Imperial College London

FAIHealth camo centre for antimicrobial optimisation

Artificial Intelligence-Driven Clinical Decision Support for Antibiotic Optimisation

William Bolton

Viva

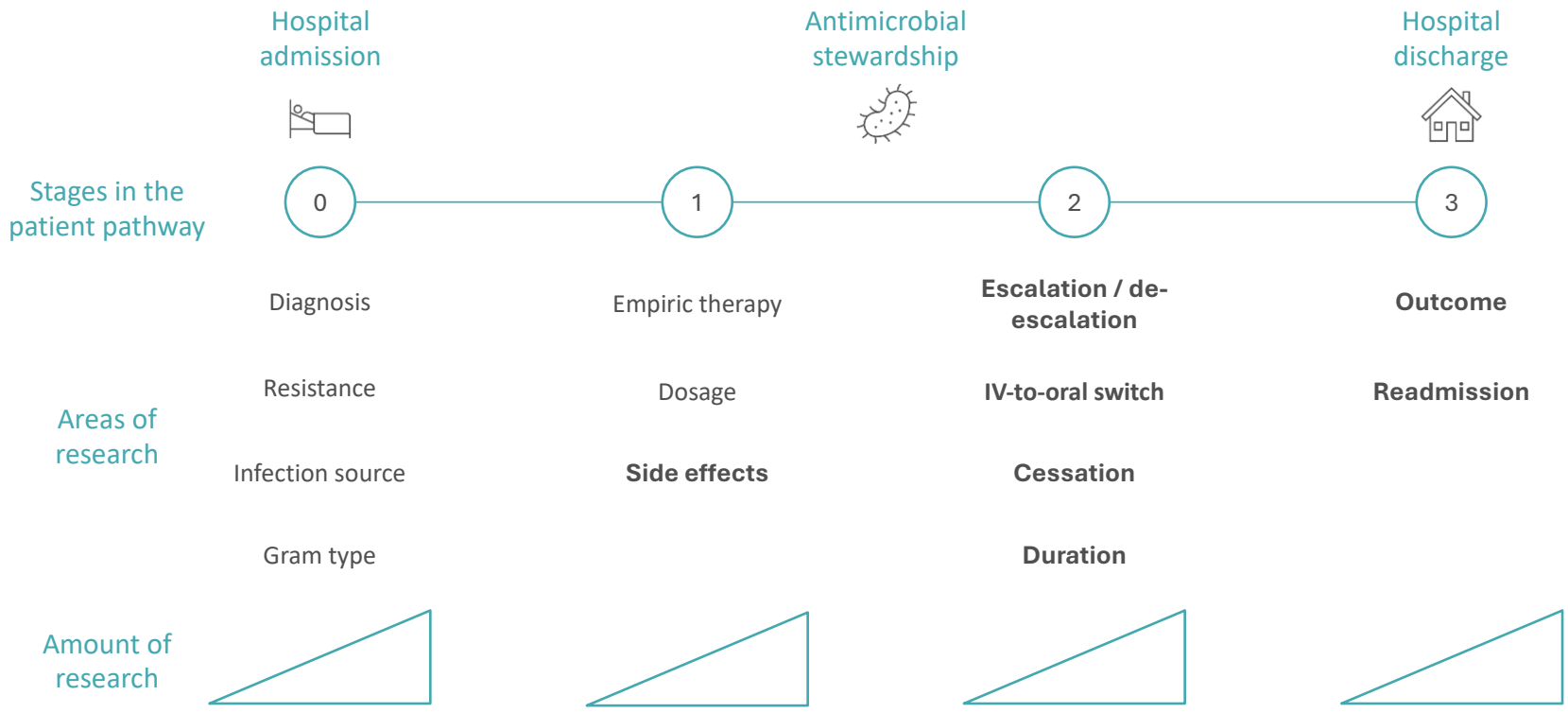
10th February 2025

Imperial College London

Contents.

Antimicrobial resistance is a global threat with data driven approaches towards stewardship neglected.

- Antimicrobial resistance is a **growing global threat**. One key strategy to tackle this is to **optimise antimicrobial use**
- **AI clinical decision support systems** have been developed to support **infection management**
- However antimicrobial stewardship has been **neglected**
- The uptake and utilisation of such systems have been limited to date, in part due to **integration and behavioural issues**.



I addressed my PhD hypothesis through three overarching themes.

Hypothesis:

Artificial intelligence-based clinical decision support systems utilising routinely collected electronic health record data can support antibiotic stewardship

Chapter 3

Co-morbidity Representation in AI

Chapter 4

Antibiotic Cessation Decision Support

Create methods to appropriately represent routinely collected patient data

Chapter 4

Antibiotic Cessation Decision Support

Chapter 5

Intravenous to Oral Switch Decision Support

Develop novel decision support systems

2.4 Moral AI for antimicrobial resistance

Chapter 6

Clinician Evaluation

Understand the ethical, and behavioral components of decision making

Using AI to optimise antimicrobial prescribing raises important ethical questions.

How can a **moral balance** be obtained between the needs of an **individual** patient and those of **wider and future society**?

Variables	Description	Exemplar of starting antimicrobial treatment	Corresponding ad-hoc utility value
Intensity	How strong is the pleasure?	Treating a relevant infection with antimicrobials has the potential to save that person's life	Highly positive utility
Duration	How long will the pleasure last?	Any extension of life is immeasurable while it is reasonable AMR will continue in the near-term future	Positive utility
Certainty or uncertainty	How likely or unlikely is it that the pleasure will occur?	Limited information often means treatment may or may not be helpful and there is always an inherent risk of developing AMR	Neutral utility, without more information
Propinquity	How soon will the pleasure occur?	Treatment can be effective immediately however the same is true for the evolution of AMR	Neutral utility, without more information
Fecundity	The likelihood of further sensations of the same kind	-	Unable to assign
Purity	The likelihood of not being followed by opposite sensations	-	Unable to assign
Extent	How many people will be affected?	Prescribing antimicrobials effects the patient and those close to them, while the development of AMR is a certainty and may affect everyone, causing significant suffering and mortality	Immense negative utility

Ethical frameworks can help ensure AI systems are fair and moral.

FINDINGS

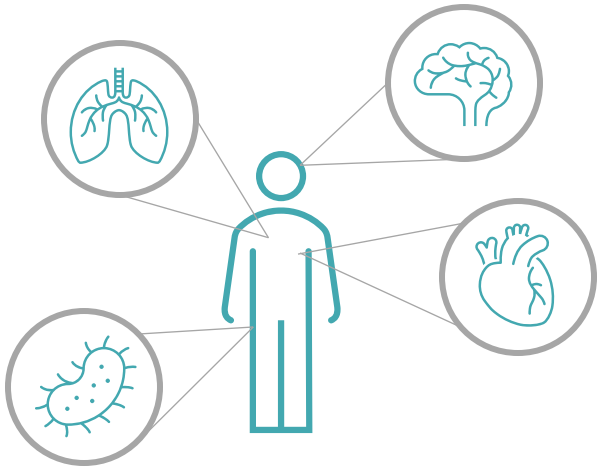
- A **utilitarian approach** may be most suitable for developing AI-based CDSSs for AMR, given the potential **number of people affected** and **aligned aims** of extending life and providing equality
- **Spatial and temporal considerations** and heterogeneity are critical for infections diseases
- Aspects of AI such as **accountability, interpretability, and biases**, require further research
- We have a **responsibility** towards the health of future generations

nature machine intelligence

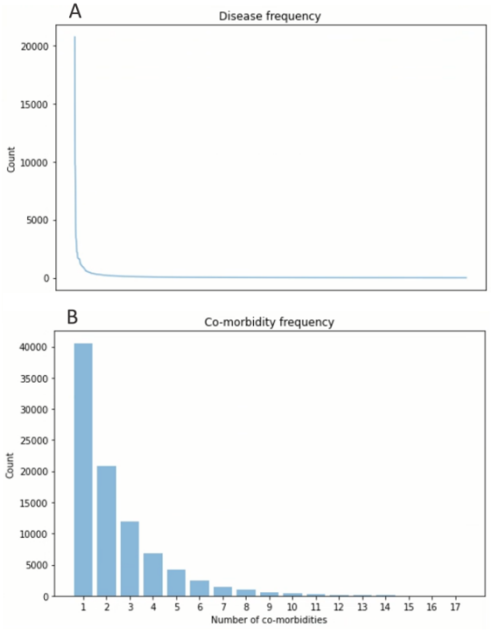
Contributions

Applied ethical theories to the nuanced problem of optimised antimicrobial prescribing through **AI to tackle AMR**

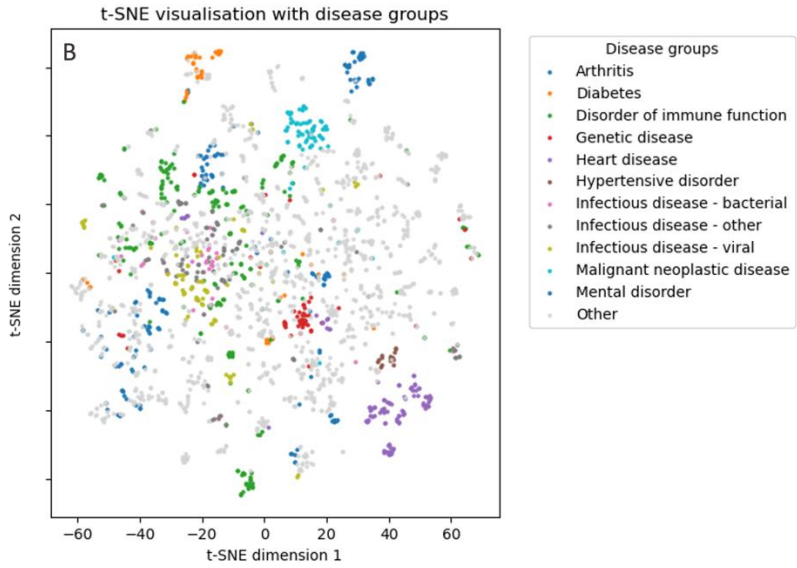
Co-morbidities are a problem for healthcare and AI systems and increase an individual's risk of infection.



Co-morbidities or chronic long-term medical conditions increase infection risk and are a **major challenge in healthcare**



Challenges such as **combinatorial complexity**, **heterogeneity**, and a lack of data make using disease data in AI systems difficult

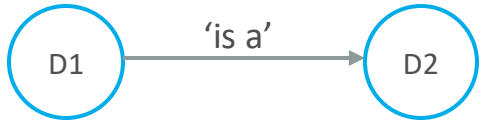


We created meaningful **embeddings from external medical knowledge**, to help overcome these challenges

Medical knowledge can help create informative disease and patent embeddings.

NHS Imperial College Healthcare NHS Trust

SNOMED CT The global language of healthcare



95,157 patients

2,133 chronic conditions

Table 2: Mean results for the similar patient retrieval task.

Method	SNOMED similarity score	Charlson Jaccard index
One hot encodings	4.40 (SD 2.32)	**0.88 (SD 0.30)**
Rocheteau's method	3.52 (SD 3.26)	0.69 (SD 0.20)
Co-morbid patient embeddings	**1.78 (SD 1.90)**	0.84 (SD 0.34)

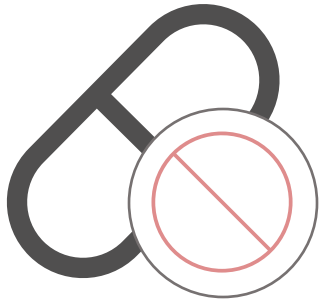
	Co-morbidities			
Question 8 patient	Gestational diabetes mellitus	Hypertensive disorder	Pre-eclampsia	Varicella
Co-morbid patient embeddings	Gestational diabetes mellitus	Pregnancy-induced hypertension	Pre-eclampsia	Varicella
Rocheteau score	Gestational diabetes mellitus	Hypertensive disorder	-	Varicella
One hot encodings	Gestational diabetes mellitus	-	Pre-eclampsia	Varicella

Contributions

Developed a **novel generalisable pipeline** to extract and utilize medical knowledge to **represent diseases and co-morbid patients** and demonstrated its **utility in multiple tasks**

aaai SpringerLink

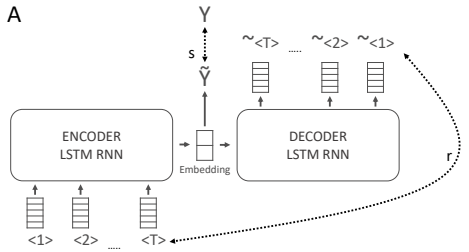
Understanding antibiotic cessation with artificial intelligence.



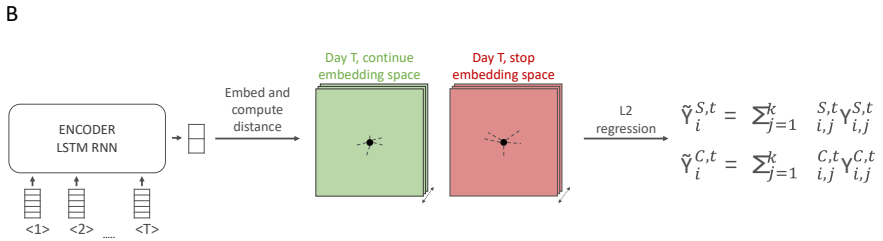
Understanding when it is appropriate to **stop antibiotic treatment** is complex

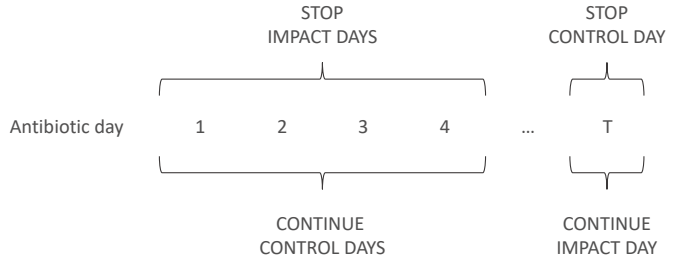
Currently treatment durations are often decided based on habit or limited historical population evidence, with **extended durations common**

AUTOENCODER TRAINING



SYNTHETIC OUTCOME ESTIMATION





Use artificial intelligence to estimate the **impact of stopping antimicrobial treatment** on a patient's treatment response or outcome. Important factors influence cessation decision making

Machine learning and synthetic outcome estimation for individualised antimicrobial cessation.

AUTOENCODER PREDICTIONS

	Metric	Result
Mortality Classification	AUROC	0.77
	Accuracy	0.73
LOS Regression	RMSE	3.88

SYNTHETIC OUTCOME ESTIMATION

SCENARIO	DAY(S)	LOS				Mortality		
		Mean delta (days, p-value)	MAPE	MAE	RMSE	Mean delta	MAE	AUROC
STOP	IMPACT	2.71*, <0.01	0.36	3.30	4.80	0.06	0.25	0.66
	CONTROL	0.24, 0.60	0.26	1.32	1.93	0.05	0.15	0.72
CONTINUE	IMPACT	-2.09*, <0.01	0.77	2.85	3.16	0.05	0.18	0.67
	CONTROL	0.42*, 0.01	0.48	2.72	3.76	0.07	0.24	0.64

Contributions

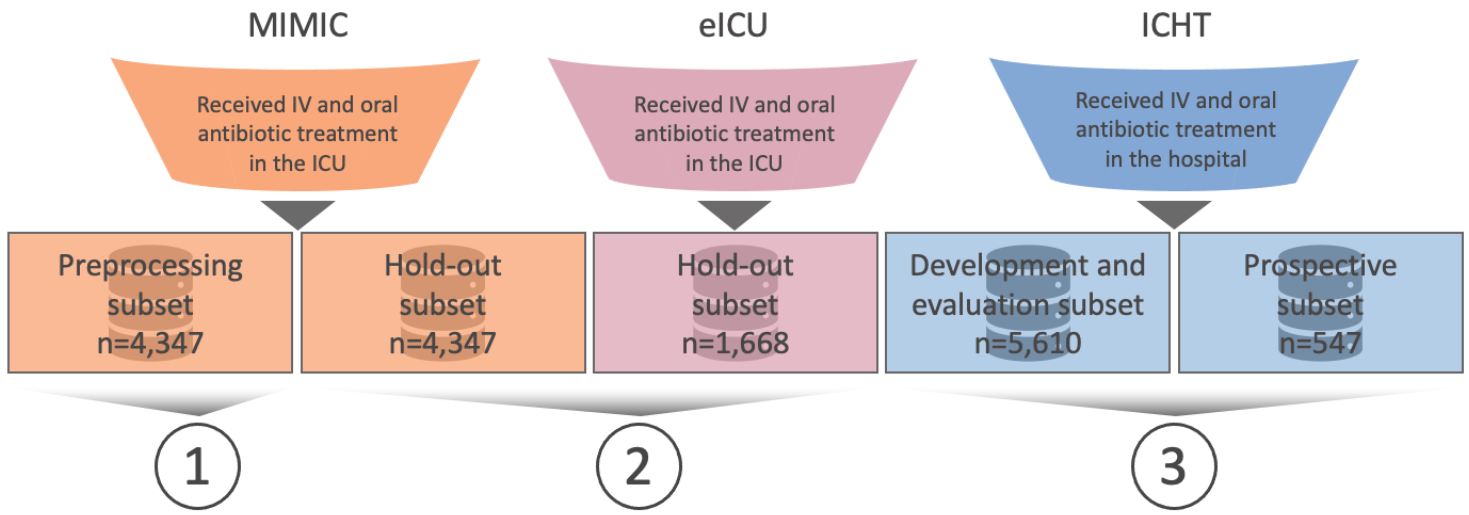
Created a **novel bi-directional LSTM autoencoder** model to estimate patient outcomes and treatment response under the **contrasting scenarios** of stopping or continuing antibiotic treatment

frontiers in Digital Health

Switching from IV-to-oral antibiotic treatment is complex and under-researched.

One key challenge of stewardship is **determining when to switch** antibiotics from **IV-to-oral administration**

Oral therapy can be **non-inferior** to IV but there is a **poor understanding** of the factors that facilitate or inhibit an individual from receiving oral therapy



UKHSA IVOS Criteria

Two feature sets

Optimised models

Evaluation versus baseline

Subgroup analysis

IV treatment length

Fairness

Interpretability

Developed models that also incorporated demographics and co-morbidities

Evaluation including subgroups and prospective dataset

Point prevalence survey

Models achieve generalisable performance across a range of datasets and patient populations.

MIMIC

Metric	1ST threshold results	2nd threshold results	IVOS criteria baseline
AUROC	**0.78 (SD 0.02)**	0.69 (SD 0.03)	0.66
FPR	0.25 (SD 0.02)	**0.10 (SD 0.02)**	0.43

eICU

Metric	1st threshold results	2nd threshold results	IVOS criteria baseline
AUROC	**0.72 (SD 0.02)**	0.65 (SD 0.05)	0.55
FPR	0.24 (SD 0.04)	**0.05 (SD 0.02)**	0.28

NHS Imperial College Healthcare NHS Trust

Metric	Initial Results	Prospective dataset	Prospective PPS
AUROC	0.79 (SD 0.01)	0.77	0.68
FPR	0.21 (SD 0.03)	0.20	0.28

Contributions

Researched **interpretable, fair and generalisable models** to predict when a patient could **switch from IV-to-oral antibiotic treatment**, with **rigorous and prospective evaluation**

nature communications

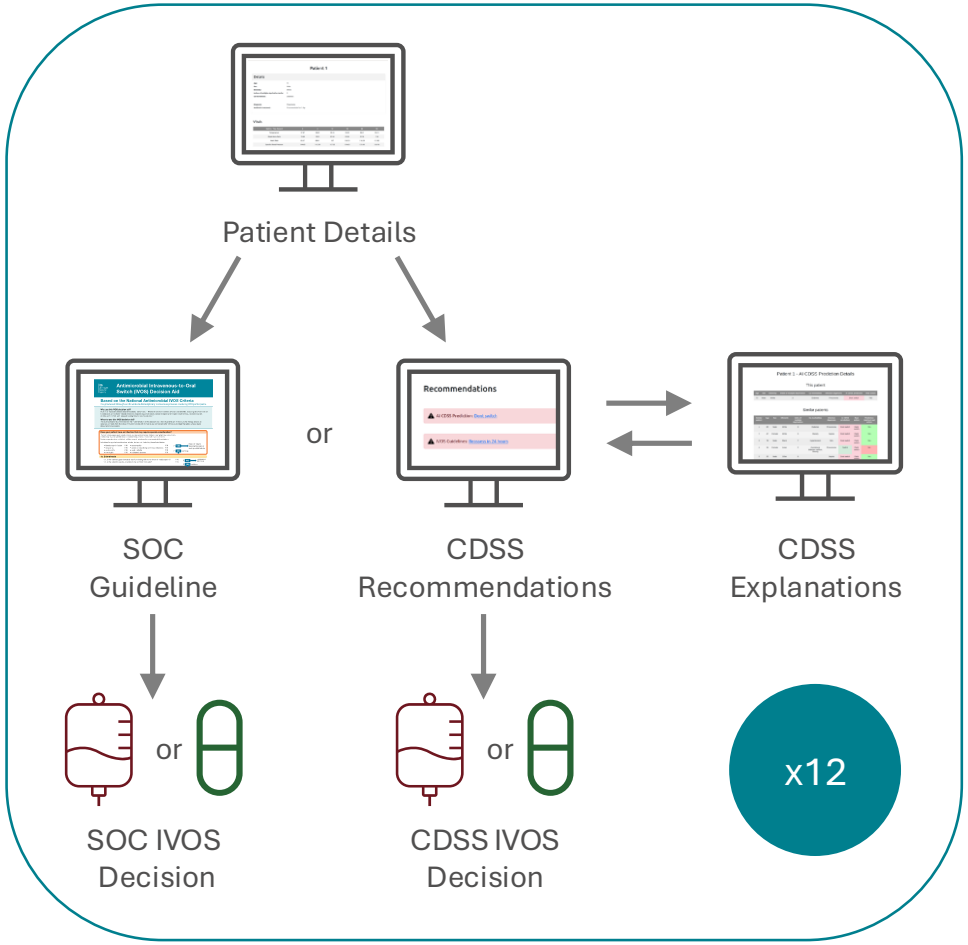
Clinician evaluation was conducted through case vignettes, interviews and questionnaires.

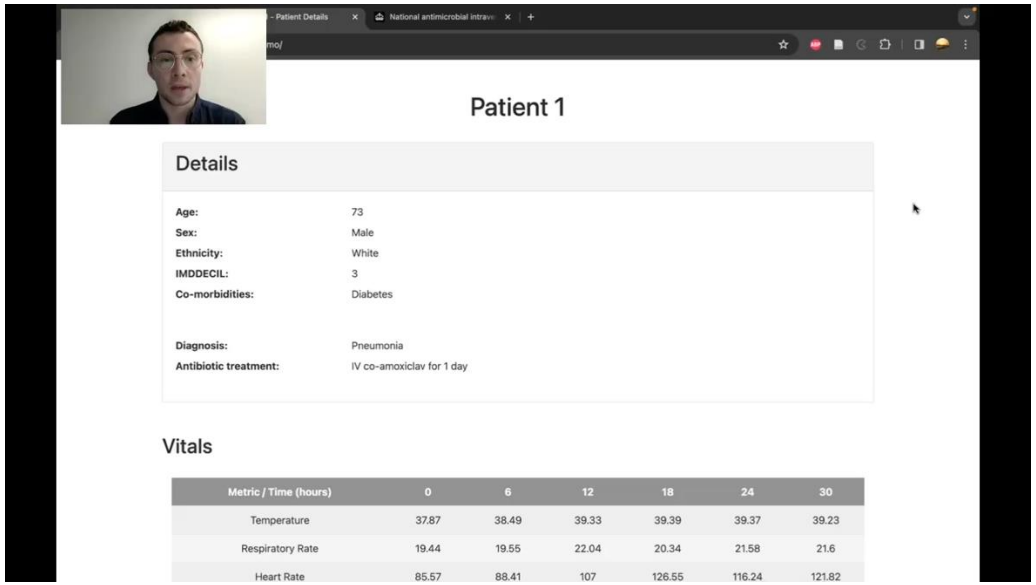
Clinicians

- Information sheet
- Informed consent
- Video demonstration

Semi-structured Interview

SUS & TAM Questionnaires





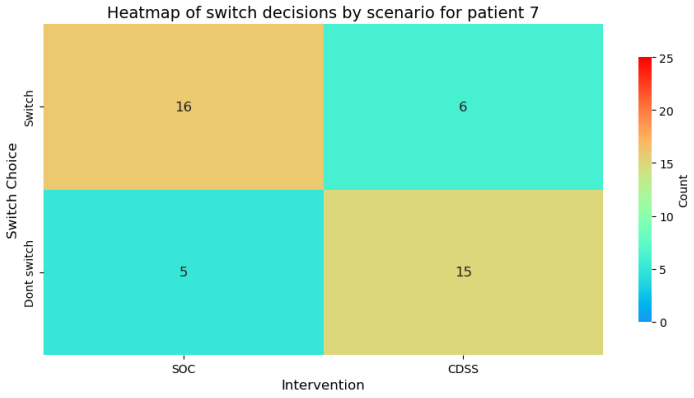
A greater impact was observed when the AI-IVOS CDSS recommended don't switch.

NO DIFFERENCES

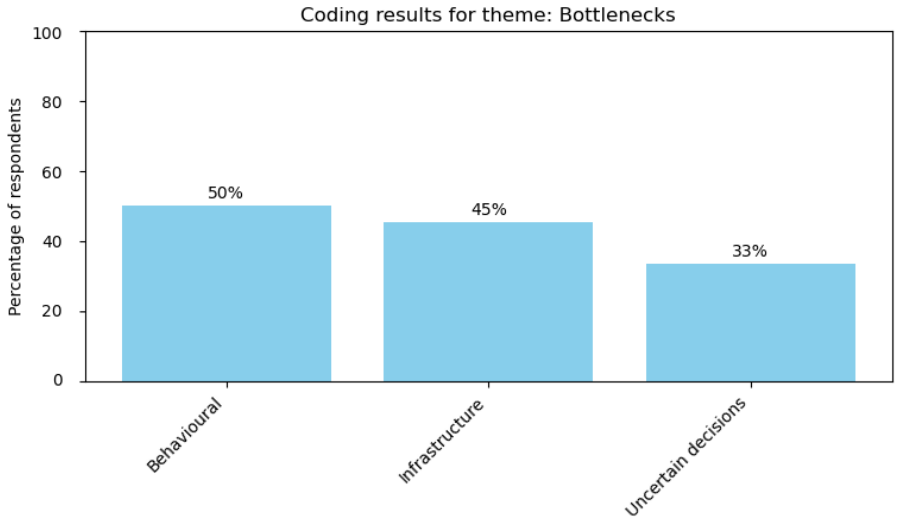
11/12 cases

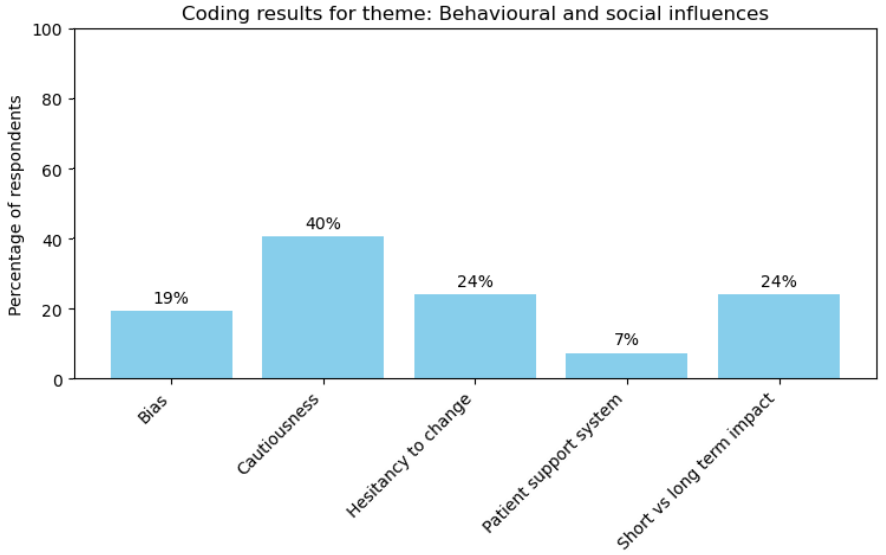
SIGNIFICANT DIFFERENCES

Patient 7 and subgroup analysis for 3 patients



System Usability score:	72.32 / 100
Perceived usefulness:	3.59 / 5
Perceived ease of use:	3.83 / 5
Self efficacy:	4.05 / 5





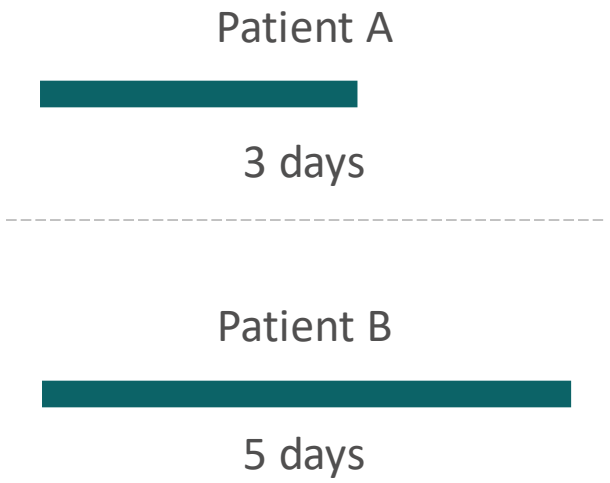
Contributions

Study providing an understanding of the **potential use case and benefit** of the AI-IVOS CDSS with **wider learnings** for the AI CDSSs ecosystem

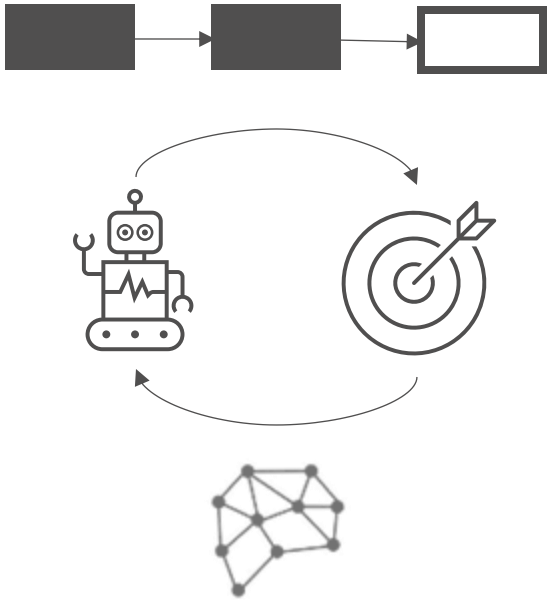
TBC

Future work includes clinical translation, other parts of prescribing and alternative AI methodologies.

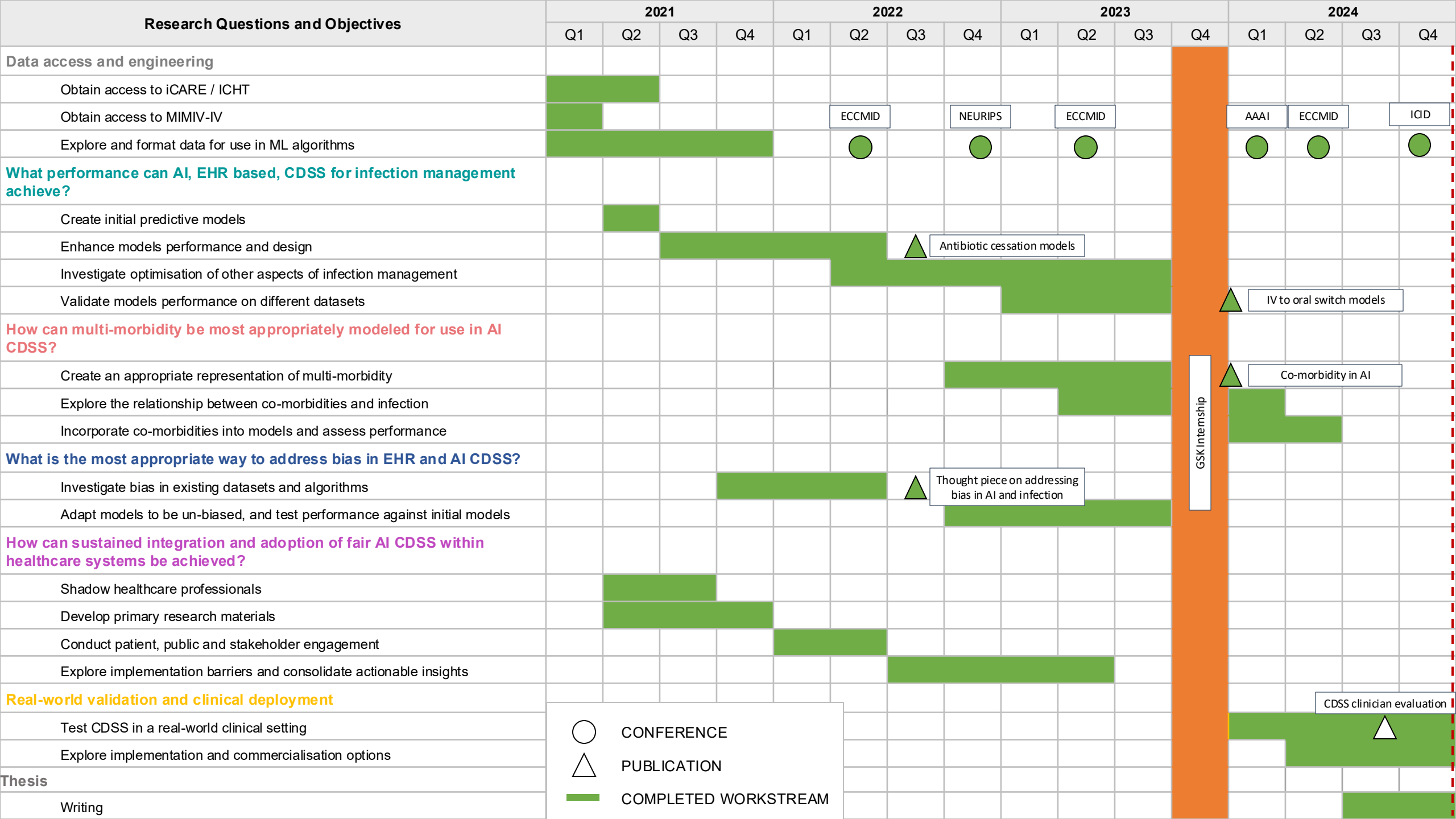
Clinical translation including developing infrastructure, conducting pilots and gathering evidence



Other aspects of optimised antimicrobial prescribing including treatment duration



Alternative AI methodologies: self-supervised learning, reinforcement learning, foundational models


Research Questions and Objectives
2021
2022
2023
2024
Q1
Q2
Q3
Q4
Data access and engineering
Obtain access to iCARE / ICHT
Obtain access to MIMIV-IV
Explore and format data for use in ML algorithms
What performance can AI, EHR based, CDSS for infection management achieve?
Create initial predictive models
Enhance models performance and design
Investigate optimisation of other aspects of infection management
Validate models performance on different datasets
How can multi-morbidity be most appropriately modeled for use in AI CDSS?
Create an appropriate representation of multi-morbidity
Explore the relationship between co-morbidities and infection
Incorporate co-morbidities into models and assess performance
What is the most appropriate way to address bias in EHR and AI CDSS?
Investigate bias in existing datasets and algorithms
Adapt models to be un-biased, and test performance against initial models
How can sustained integration and adoption of fair AI CDSS within healthcare systems be achieved?
Shadow healthcare professionals
Develop primary research materials
Conduct patient, public and stakeholder engagement
Explore implementation barriers and consolidate actionable insights
Real-world validation and clinical deployment
Test CDSS in a real-world clinical setting
Explore implementation and commercialisation options
Thesis
Writing
ECCMID
NEURIPS
ECCMID
AAAI
ECCMID
ICID
Antibiotic cessation models
IV to oral switch models
Co-morbidity in AI
Thought piece on addressing bias in AI and infection
GSK Internship
CDSS clinician evaluation
CONFERENCE
PUBLICATION
COMPLETED WORKSTREAM

GSK internship exploring the intricacies of ensuring safety and understanding hallucinations in large language models.

Published at ICLR 2024 Workshop on Reliable and Responsible Foundation Models

RAmBLA: A FRAMEWORK FOR EVALUATING THE RELIABILITY OF LLMs AS ASSISTANTS IN THE BIOMEDICAL DOMAIN

William James Bolton*
Imperial College London†
william.bolton@imperial.ac.uk

Rafael Poyiadzi*
GSK.ai
rafael.x.poyiadzi@gsk.com

Edward Morrell*
GSK.ai
ed.r.morrell@gsk.com

Gabriela van Bergen Gonzalez Bueno*
GSK.ai
gabriela.v.vanbergengonzalez-bueno@gsk.com

Lea Goetz
GSK.ai
lea.x.goetz@gsk.com

ABSTRACT

Large Language Models (LLMs) increasingly support applications in a wide range of domains, some with potential high societal impact such as biomedicine, yet their reliability in realistic use cases is under-researched. In this work we introduce the Reliability AssesMent for Biomedical LLM Assistants (RAmBLA[1]) framework and evaluate whether four state-of-the-art foundation LLMs can serve as reliable assistants in the biomedical domain. We identify prompt robustness, high recall, and a lack of hallucinations as necessary criteria for this use case. We design shortform tasks and tasks requiring LLM freeform responses mimicking real-world user interactions. We evaluate LLM performance using semantic similarity with a ground truth response, through an evaluator LLM.


William Bolton
Imperial College London


Rafael Poyiadzi
GSK AI/ML


Edward Morrell
GSK AI/ML


Gabriela van Bergen Gonzalez-Bueno
GSK AI/ML


Lea Goetz
GSK AI/ML

ICLR International Conference On Learning Representations

Artificial Intelligence-Driven Clinical Decision Support for Antibiotic Optimisation.

Conclusion

- Created novel data pipelines and **artificial intelligence methodologies**
- **Research covers a wide area** spanning from fundamental data analysis and model development, to deploying artificial intelligence-based decision support applications with clinicians
- Tackled **key aspects of antibiotic prescribing** under-represented in the literature but impactful regarding a clinical team, their patient, and antimicrobial resistance
- Antimicrobial prescribing and stewardship decisions are are **technically, ethically, and behaviourally complex and numerous barriers exist** to implementing AI technology
- Artificial intelligence has potential to tackle antimicrobial resistance through **optimised and individualised antibiotic prescribing decisions**

Imperial College London

Thanks everyone for the support.

Dr Tim Rawson

Professor Pantelis Georgiou

Professor Alison Holmes

Professor Mark Gilchrist

Richard Wilson

Dr David Antcliffe

Dr Bernard Hernandez Perez

Cosmin Badea

Britta Ross

Imperial College London

FIHealth

camo centre for antimicrobial optimisation

Thank you!

William Bolton

Viva

10th February 2025

Imperial College London

FHi Health

camo centre for antimicrobial optimisation

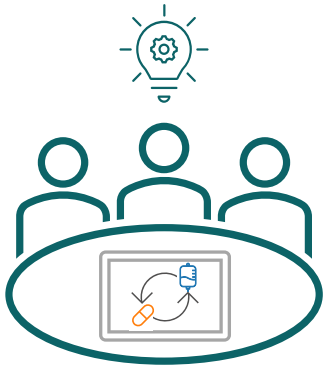
Appendix

William Bolton

Viva

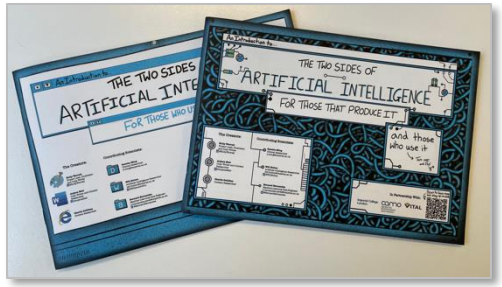
10th February 2025

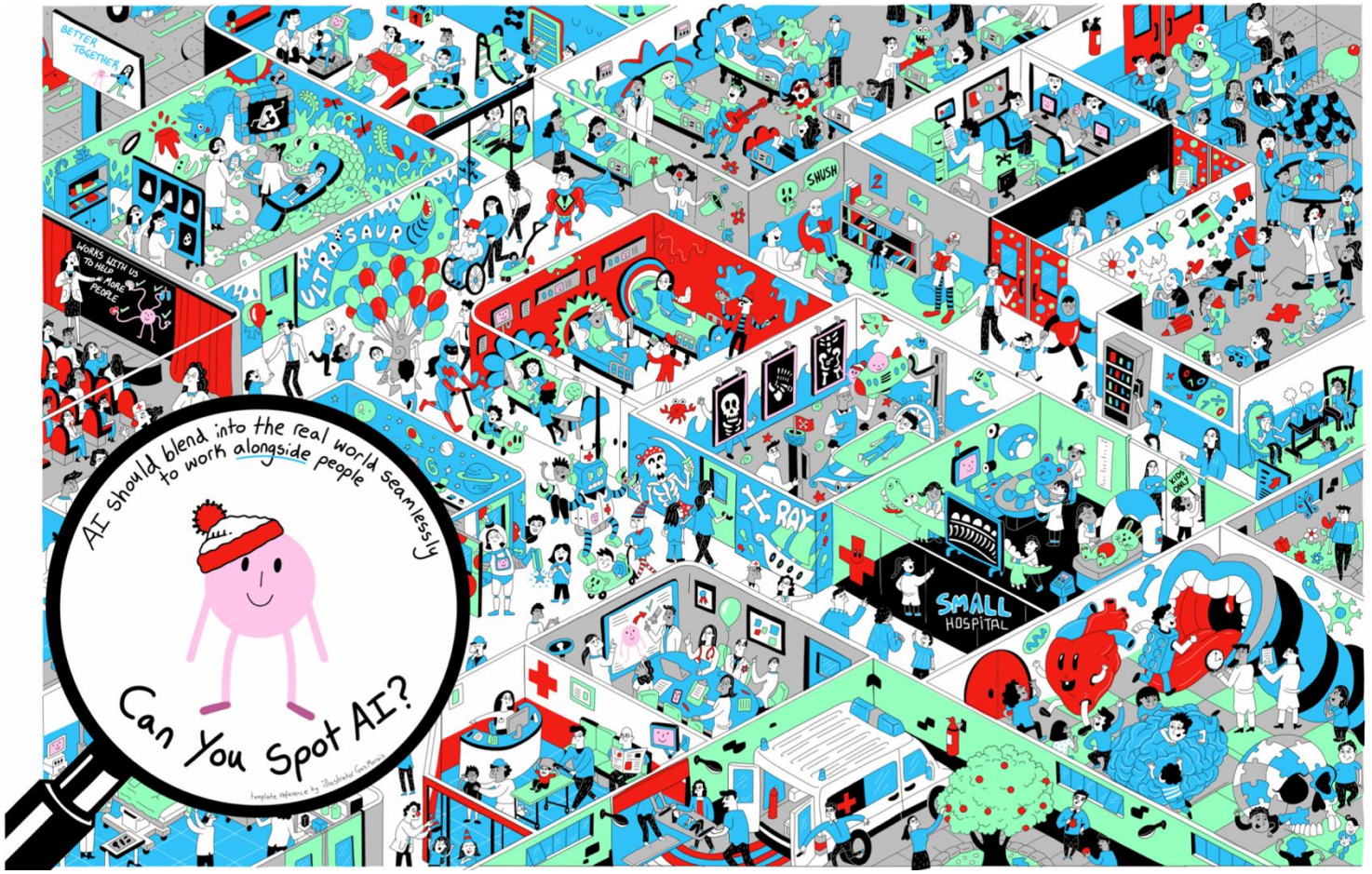
Prospective evaluation and education are essential for technological adoption, implementation and impact.



We have conducted **end user assessment** and **prospective testing** with clinicians in **simulated settings**

PRIMARY RESEARCH AND EDUCATION

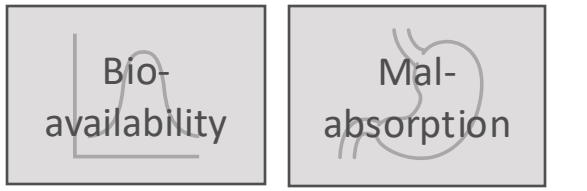


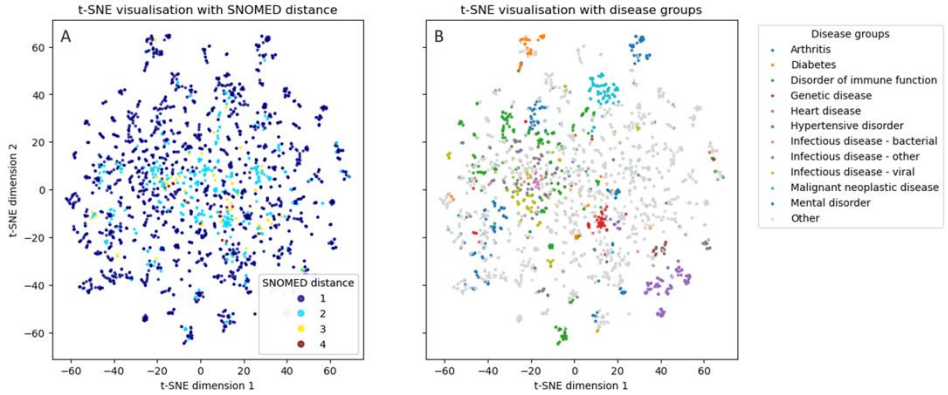


Data often poses a challenge for AI systems in healthcare, particularly those focusing on AMR.

DATA QUALITY AND MISSINGNESS

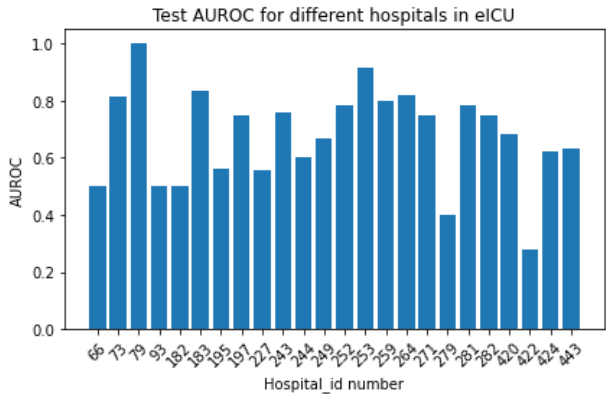
- **Lack of reliable data** on important factors such as absorption
- Appling some important parameters such as **co-morbidities** to AI systems is **combinatorially complex**





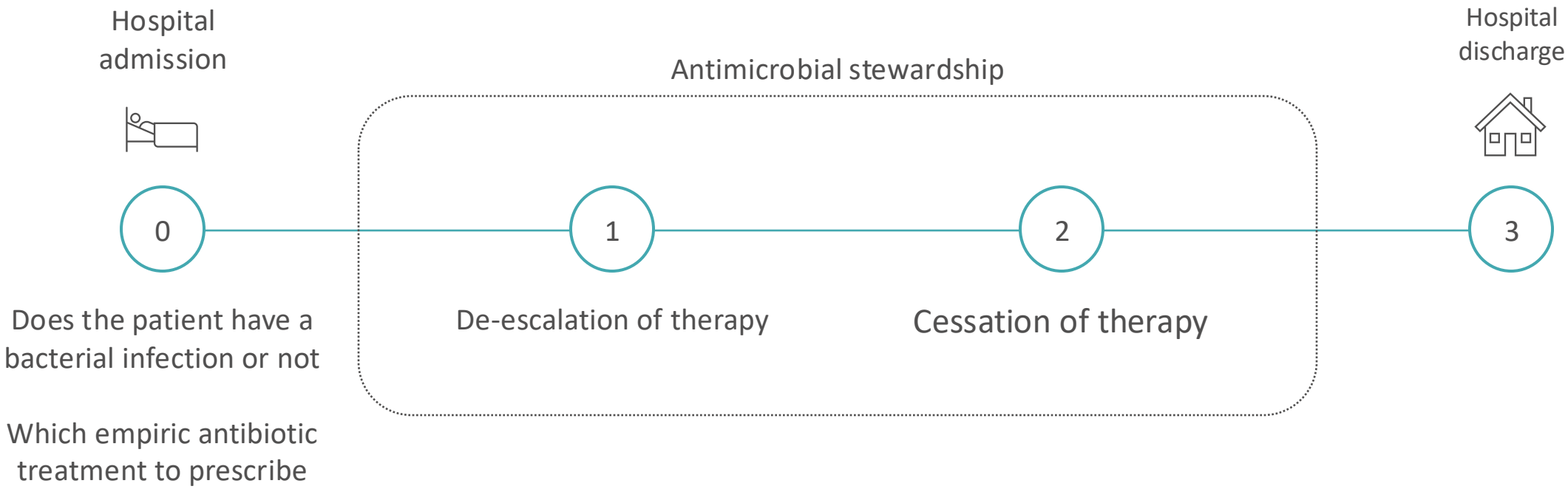
HUMAN BEHAVIOUR IS HETEROGENEOUS

- Antimicrobial stewardship is driven by **human actions** which can be **difficult to model and predict**



Antimicrobial stewardship aims to optimise antibiotic decision making.

STAGES OF ANTIBIOTIC DECISION MAKING



Antimicrobial stewardship

A coordinated effort and set of practices aimed at **optimising antimicrobial use** and **prolonging their therapeutic life**, to improve infection patient **outcomes** while minimizing the development of **antimicrobial resistance**

Artificial intelligence can support optimised antibiotic decision making.

STAGES OF ANTIBIOTIC DECISION MAKING

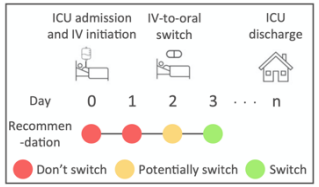
Hospital admission

Antimicrobial stewardship

Hospital discharge

0 — 1 — 2 — 3

IV-to-oral switch



Antibiotic readmission

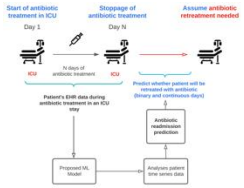


Figure 1.2: Proposed ML-based decision support model

Side effects

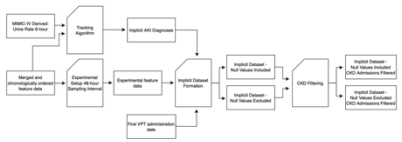
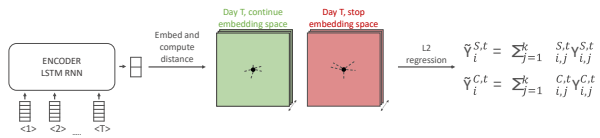


Figure 13: Implicit dataset formation workflow.

Antibiotic cessation



Imperial College London

Can we create a roadmap for responsibly designing, evaluating and integrating AI in healthcare

nature medicine

Explore content ▾ About the journal ▾ Publish with us ▾

nature › nature medicine › perspectives › article

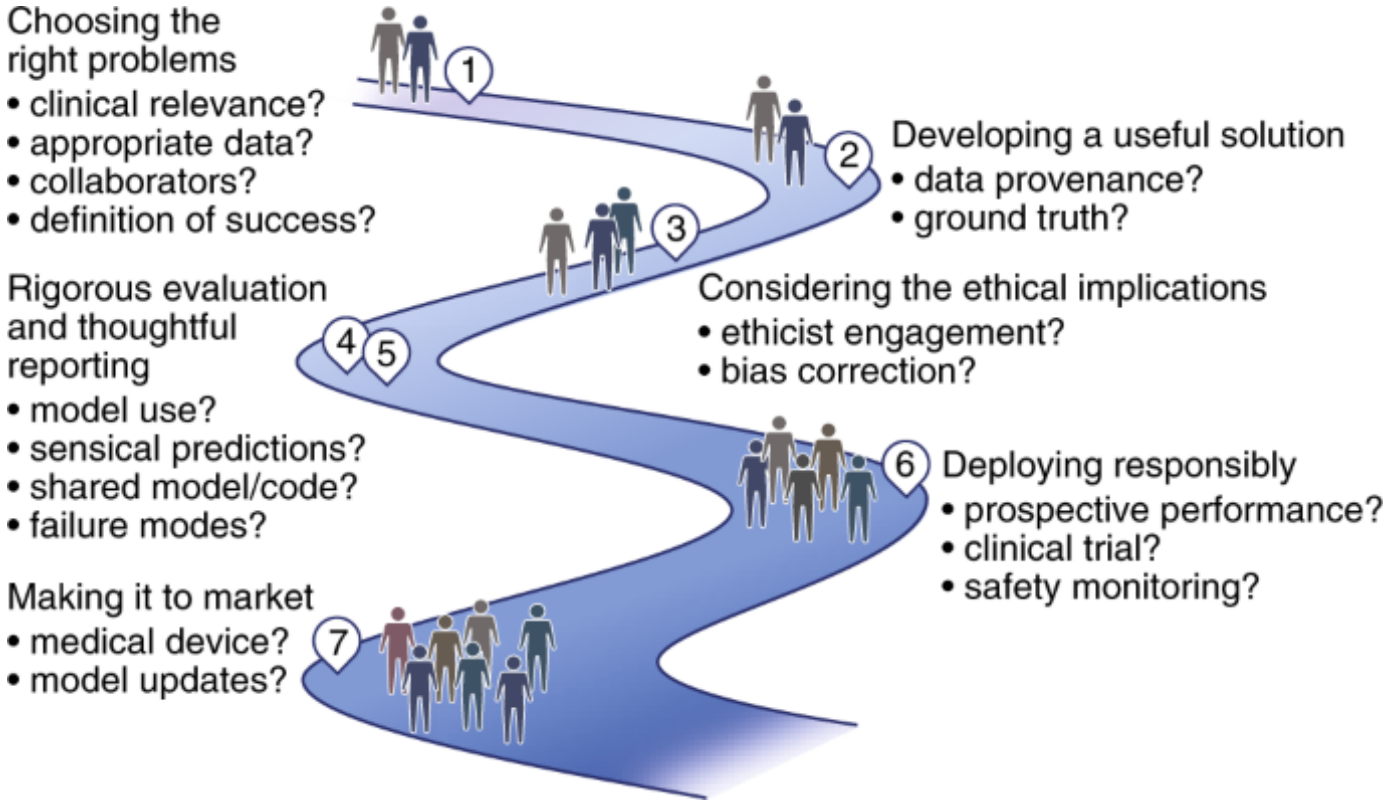
Perspective | Published: 19 August 2019

Do no harm: a roadmap for responsible machine learning for health care

Jenna Wiens ✉, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X. Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, Pilar N. Ossorio, Sonoo Thadaney-Israni & Anna Goldenberg ✉

Nature Medicine **25**, 1337–1340 (2019) | Cite this article

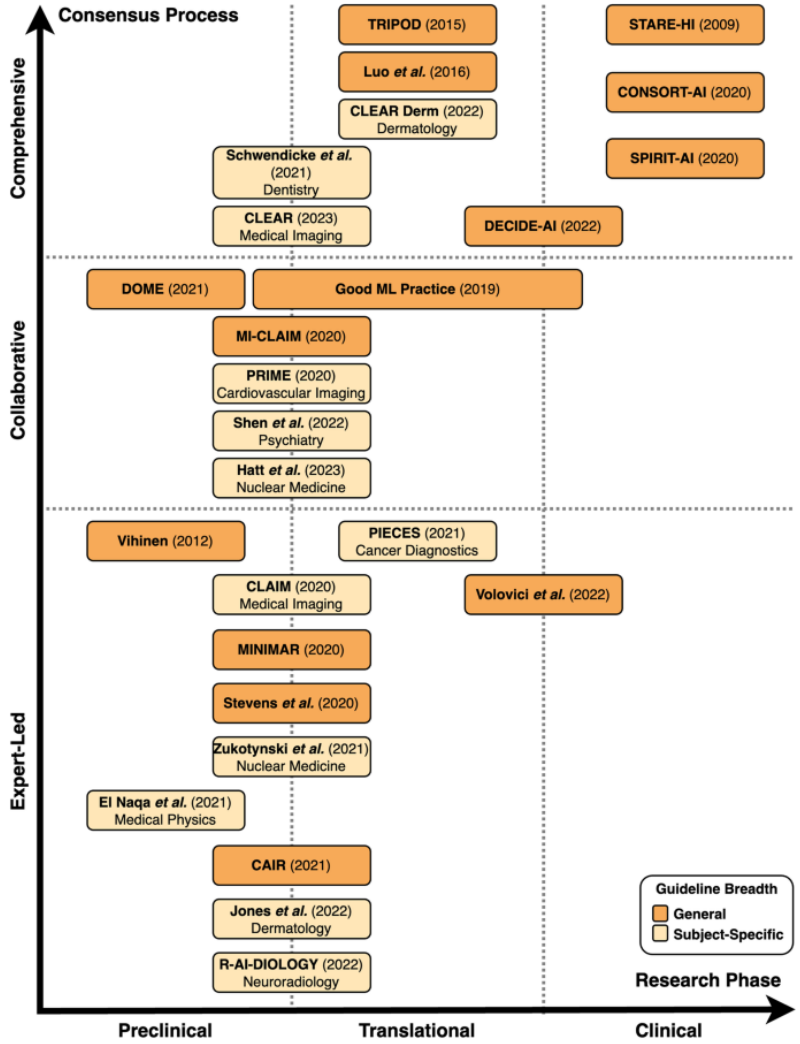
33k Accesses | **409** Citations | **704** Altmetric | Metrics

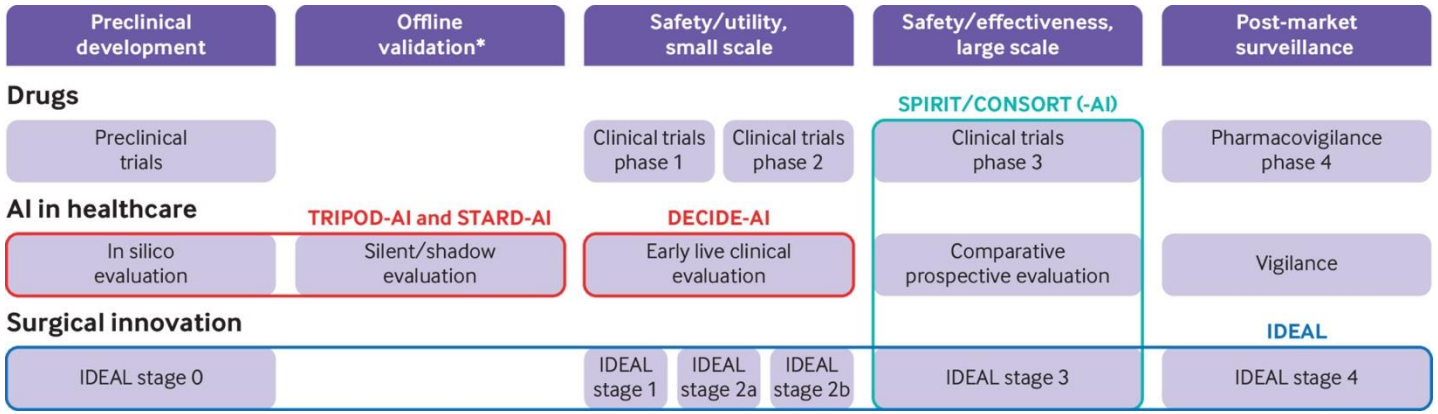


Wiens, J., Saria, S., Sendak, M. *et al.* Do no harm: a roadmap for responsible machine learning for health care. *Nat Med* **25**, 1337–1340 (2019). https://doi.org/10.1038/s41591-019-0548-6

Imperial College London

Many guidelines exist for reporting AI in medicine





Box 2: Noteworthy changes and additions to TRIPOD 2015

- New checklist of reporting recommendations to cover prediction model studies using any regression or machine learning method (eg, random forests, deep learning), and harmonise nomenclature between regression and machine learning communities
- New TRIPOD+AI checklist supersedes the TRIPOD 2015 checklist, which should no longer be used
- Particular emphasis on fairness (box 1) to raise awareness and ensure that reports mention whether specific methods were used to deal with fairness. Aspects of fairness are embedded throughout the checklist
- Inclusion of TRIPOD+AI for Abstracts for guidance on reporting abstracts
- Modification of the model performance item recommending that authors evaluate model performance in key subgroups (eg, sociodemographic)
- Inclusion of a new item on patient and public involvement to raise awareness and prompt authors to provide details on any patient and public involvement during the design, conduct, reporting (and interpretation), and dissemination of the study
- Inclusion of an open science section with subitems on study protocols, registration, data sharing and code sharing

TRIPOD=Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis; AI=artificial intelligence.

Medicines & Healthcare products Regulatory Agency

Guidance

Good Machine Learning Practice for Medical Device Development: Guiding Principles

Published 27 October 2021

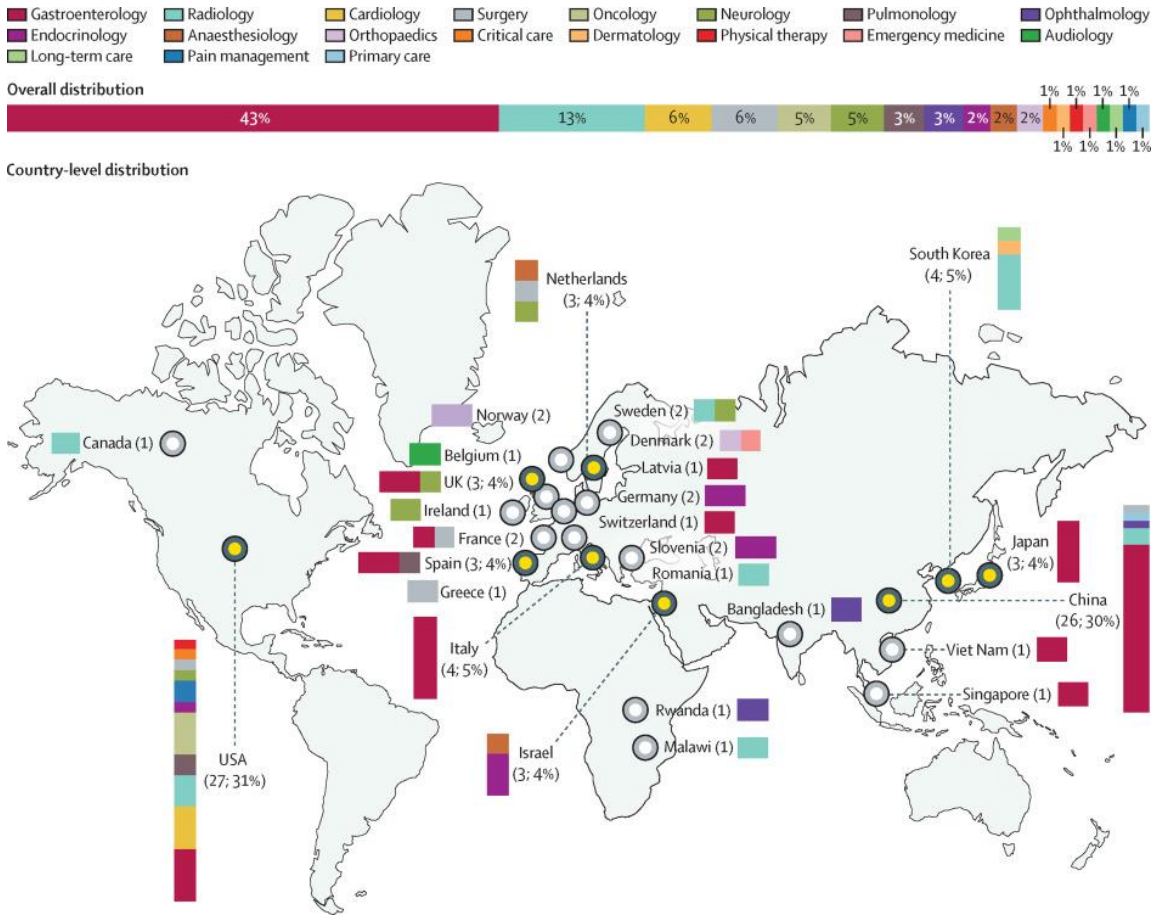
Kolbinger, F.R., Veldhuizen, G.P., Zhu, J. *et al.* Reporting guidelines in medical artificial intelligence: a systematic review and meta-analysis. *Commun Med* **4**, 71 (2024). https://doi.org/10.1038/s43856-024-00492-0

Vasey B, Nagendran M, Campbell B, Clifton D A, Collins G S, Denaxas S et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI BMJ 2022; 377 :e070904 doi:10.1136/bmj-2022-070904

Collins G S, Moons K G M, Dhiman P, Riley R D, Beam A L, Van Calster B et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods BMJ 2024; 385 :e078378 doi:10.1136/bmj-2023-078378

Imperial College London

Few clinical trials of AI in real clinical practice exist - especially in infectious diseases.



nature medicine

Explore content | About the journal | Publish with us

nature > nature medicine > articles > article

Article | Published: 21 July 2022

Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis

Roy Adams, Katharine E. Henry, Anirudh Sridharan, Hossein Soleimani, Andong Zhan, Nishi Rawat, Lauren Johnson, David N. Hager, Sara E. Cosgrove, Andrew Markowski, Eili Y. Klein, Edward S. Chen, Mustapha O. Saheed, Maureen Henley, Sheila Miranda, Katrina Houston, Robert C. Linton, Anushree R. Ahluwalia, Albert W. Wu ✉ & Suchi Saria ✉

Nature Medicine **28**, 1455–1460 (2022) | Cite this article

32k Accesses | **73** Citations | **496** Altmetric | Metrics

Han R, Acosta JN, Shakeri Z, Ioannidis JP, Topol EJ, Rajpurkar P. Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. The Lancet Digital Health. 2024 May 1;6(5):e367-73.

Adams, R., Henry, K.E., Sridharan, A. et al. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. Nat Med 28, 1455–1460 (2022). https://doi.org/10.1038/s41591-022-01894-0

Imperial College London

AI clinical decision support systems are regulated at a minimum as Class II software as a medical device in the UK.

The Journey

Pre-Market

1. **Intended Use**
   The critical first step in the development of AI and health tech products. A clear intended use prioritises safety, effectiveness and gives clarity on how to position your SaMD for success.
   READ MORE →

2. **Risk Classification**
   The risk level of a medical device or AI product determines the required clinical evidence and regulatory oversight. Read on to learn more!
   READ MORE →

3. **Notified / Approved Body Engagement**
   To launch a product in the UK or EU, an independent notified body or approved body must review and comply with European legislation, granting UKCA marks.
   READ MORE →

4. **Quality Management Systems (ISO 13485)**
   Elevate compliance in medical device manufacturing with our QMS. We cover design, supply, risk management, and CAPAs for a solid regulatory strategy. Read more!
   READ MORE →

5. **Medical Device File Design**
   Your MDF provides evidence to demonstrate compliance of the device to all the applicable regulations. Its structure and design is key.
   READ MORE →

6. **MDSAP**
   The Medical Device Single Audit Programme (MDSAP) streamlines quality management systems by enabling compliance proof through a single audit for five markets.
   READ MORE →

7. **Clinical Evaluation Plan**
   The Clinical Evaluation Plan (CEP) is a vital tool in product development, guiding device clinical evaluation through the Valid Clinical Association. Read more!
   READ MORE →

8. **QMS Deployment & Training**
   Hardian is an AI-based SaMD company that offers training and guidance to help clients meet international standards for quality management system implementation.
   READ MORE →

9. **Software Verification & Validation**
   The process of the software development lifecycle is essential to ensure all requirements are met before testing the product in the real world. Read more!
   READ MORE →

10. **Clinical Evaluation Report**
    The Clinical Evaluation Report (CER) is a crucial document in clinical evaluation, containing development activities and clinical evidence for device marketing.
    READ MORE →

11. **Responsible Person (PRRC and UKRP)**
    Depending on your jurisdiction of deployment, you may need to appoint responsible persons across these jurisdictions.
    READ MORE →

12. **Product Registration**
    Now the hard work is done, registering your product on the relevant regulatory databases is the final step to legally place your product on the market.
    READ MORE →

Medicines & Healthcare products Regulatory Agency

Guidance

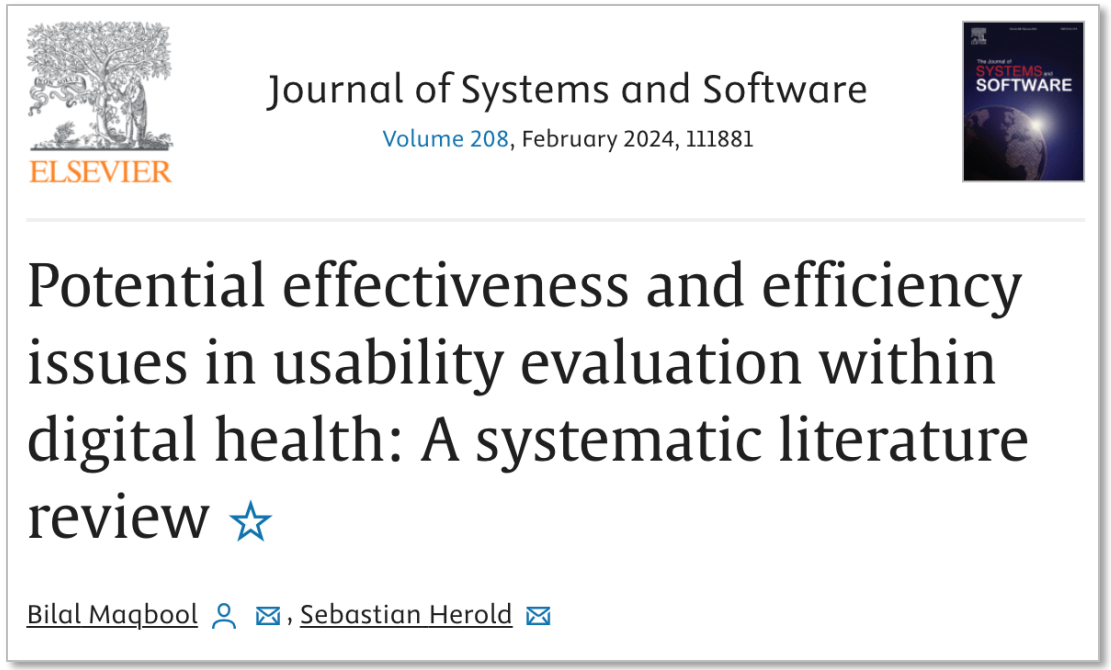
Software and artificial intelligence (AI) as a medical device

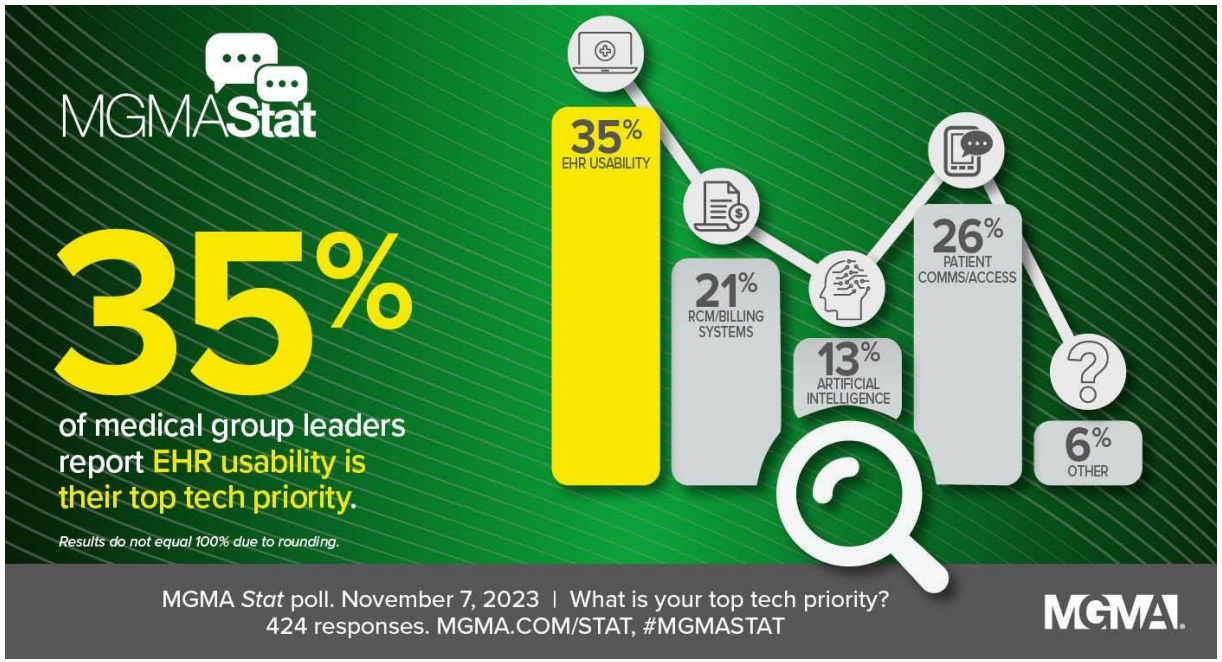
Updated 13 June 2024

https://www.england.nhs.uk/long-read/artificial-intelligence-ai-and-machine-learning/
https://www.hardianhealth.com/regulatory
https://www.gov.uk/government/publications/software-and-artificial-intelligence-ai-as-a-medical-device/software-and-artificial-intelligence-ai-as-a-medical-device

Imperial College London

Usability is also essential for trust and adoption

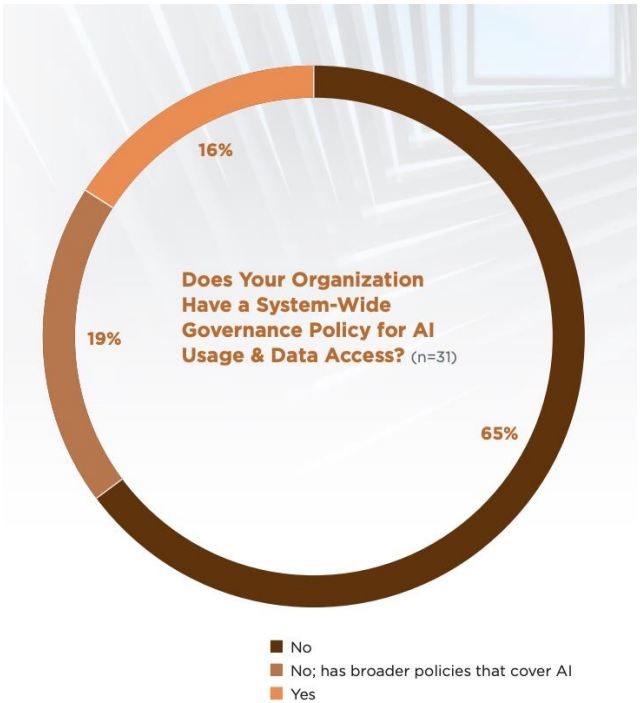


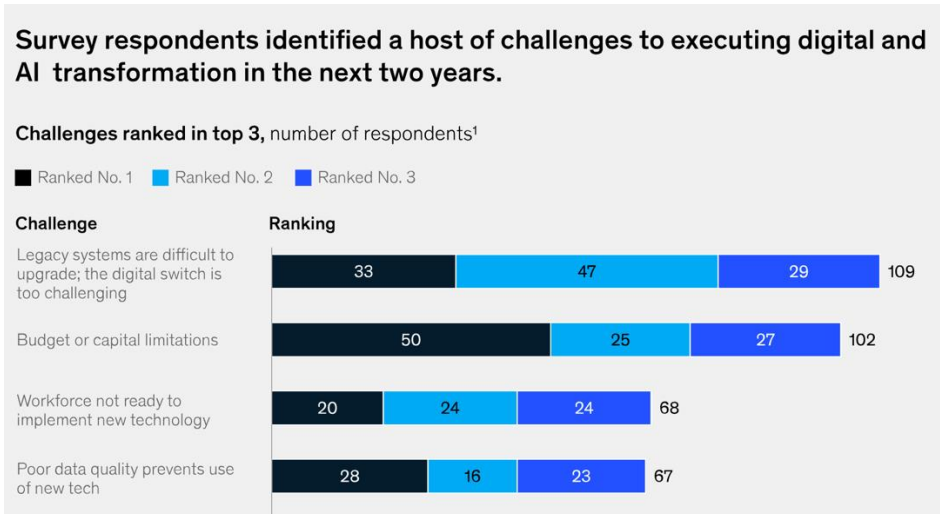


https://www.mgma.com/mgma-stat/ehr-usability-patient-communications-billing-outrank-ai-as-top-tech-priorities

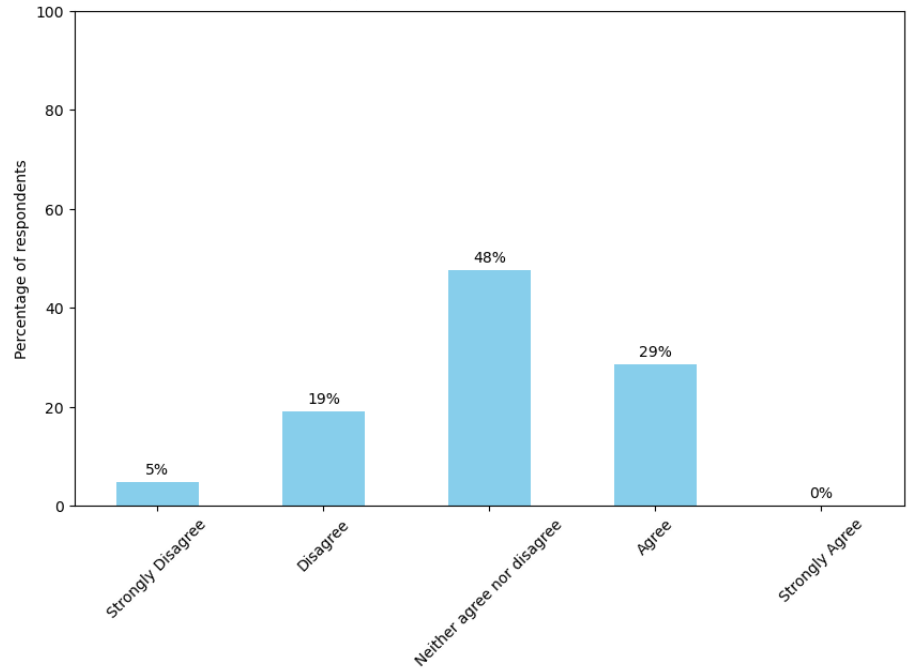
Maqbool, B. and Herold, S., 2023. Potential effectiveness and efficiency issues in usability evaluation within digital health: A systematic literature review. Journal of Systems and Software, p.111881.

Are hospitals ready for AI?





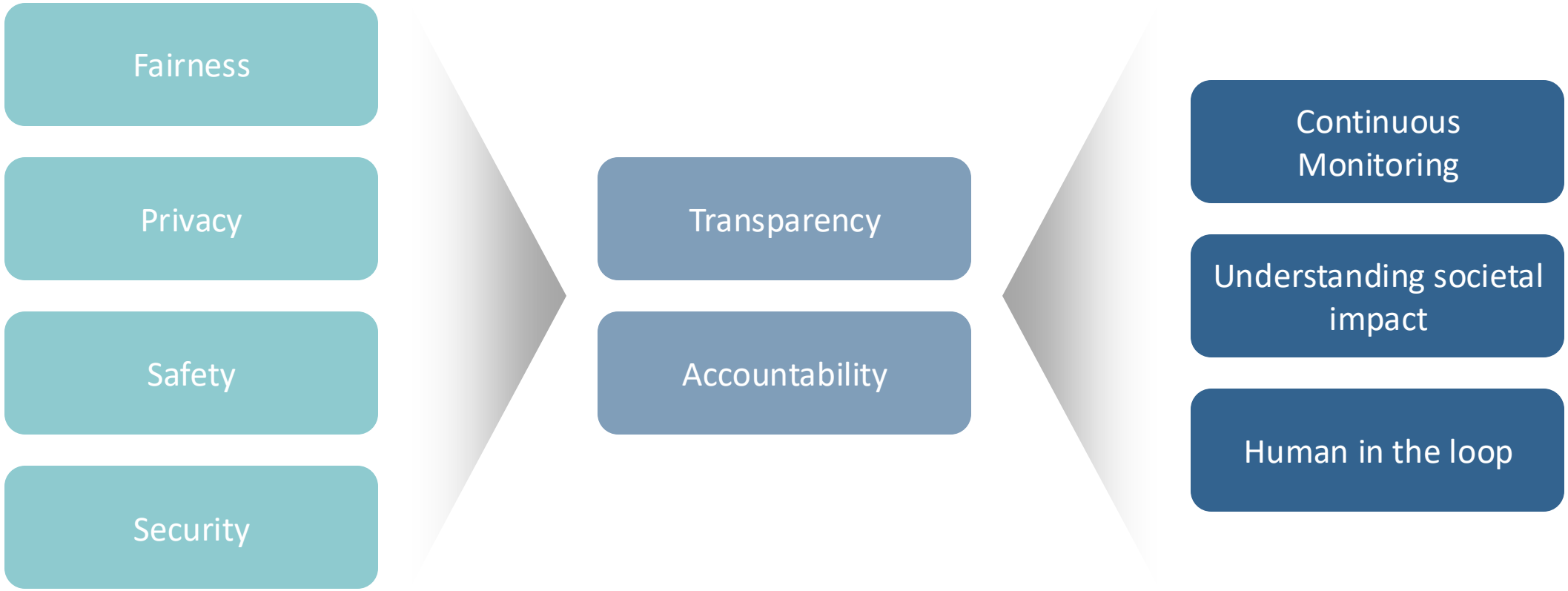
I think my healthcare institution has the necessary infrastructure to support this AI CDSS*



https://info.connectedmed.com/l/689353/2024-02-09/2lvknc/689353/1707510824kuJAqb0f/How_Health_Systems_Are_Navigating_The_Complexities_Of_AI_CCM_Reports.pdf?
https://www.mckinsey.com/industries/healthcare/our-insights/digital-transformation-health-systems-investment-priorities?utm_source=substack&utm_medium=email
*Unpublished research – n = 21

Imperial College London

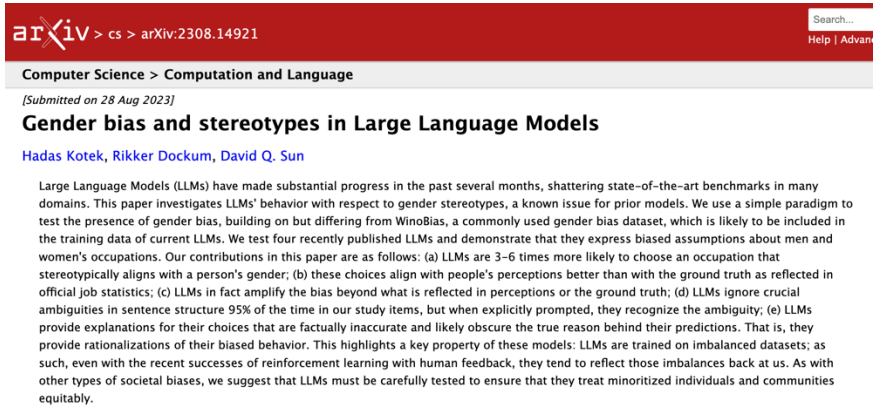
What is ethical and responsible AI?



Imperial College London

Ensuring models are responsible and ethical becomes more complex as AI advances.

BIASES



arXiv > cs > arXiv:2308.14921

Search... Help | Advanc

Computer Science > Computation and Language

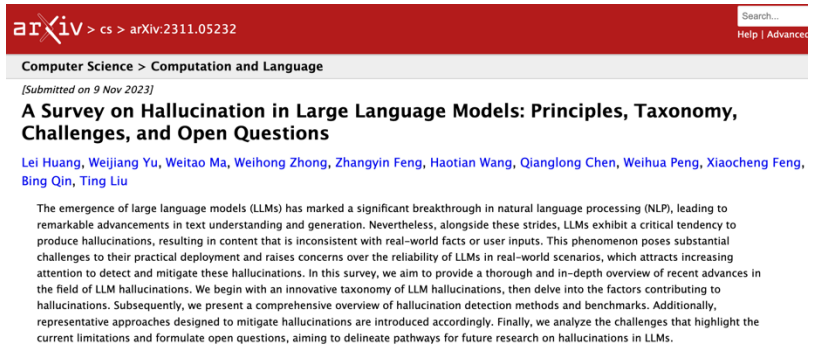
[Submitted on 28 Aug 2023]

Gender bias and stereotypes in Large Language Models

Hadas Kotek, Rikker Dockum, David Q. Sun

Large Language Models (LLMs) have made substantial progress in the past several months, shattering state-of-the-art benchmarks in many domains. This paper investigates LLMs' behavior with respect to gender stereotypes, a known issue for prior models. We use a simple paradigm to test the presence of gender bias, building on but differing from WinoBias, a commonly used gender bias dataset, which is likely to be included in the training data of current LLMs. We test four recently published LLMs and demonstrate that they express biased assumptions about men and women's occupations. Our contributions in this paper are as follows: (a) LLMs are 3-6 times more likely to choose an occupation that stereotypically aligns with a person's gender; (b) these choices align with people's perceptions better than with the ground truth as reflected in official job statistics; (c) LLMs in fact amplify the bias beyond what is reflected in perceptions or the ground truth; (d) LLMs ignore crucial ambiguities in sentence structure 95% of the time in our study items, but when explicitly prompted, they recognize the ambiguity; (e) LLMs provide explanations for their choices that are factually inaccurate and likely obscure the true reason behind their predictions. That is, they provide rationalizations of their biased behavior. This highlights a key property of these models: LLMs are trained on imbalanced datasets; as such, even with the recent successes of reinforcement learning with human feedback, they tend to reflect those imbalances back at us. As with other types of societal biases, we suggest that LLMs must be carefully tested to ensure that they treat minoritized individuals and communities equitably.

HALLUCINATIONS



arXiv > cs > arXiv:2311.05232

Search... Help | Advanced

Computer Science > Computation and Language

[Submitted on 9 Nov 2023]

A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, Ting Liu

The emergence of large language models (LLMs) has marked a significant breakthrough in natural language processing (NLP), leading to remarkable advancements in text understanding and generation. Nevertheless, alongside these strides, LLMs exhibit a critical tendency to produce hallucinations, resulting in content that is inconsistent with real-world facts or user inputs. This phenomenon poses substantial challenges to their practical deployment and raises concerns over the reliability of LLMs in real-world scenarios, which attracts increasing attention to detect and mitigate these hallucinations. In this survey, we aim to provide a thorough and in-depth overview of recent advances in the field of LLM hallucinations. We begin with an innovative taxonomy of LLM hallucinations, then delve into the factors contributing to hallucinations. Subsequently, we present a comprehensive overview of hallucination detection methods and benchmarks. Additionally, representative approaches designed to mitigate hallucinations are introduced accordingly. Finally, we analyze the challenges that highlight the current limitations and formulate open questions, aiming to delineate pathways for future research on hallucinations in LLMs.

PRIVACY LEAKAGE

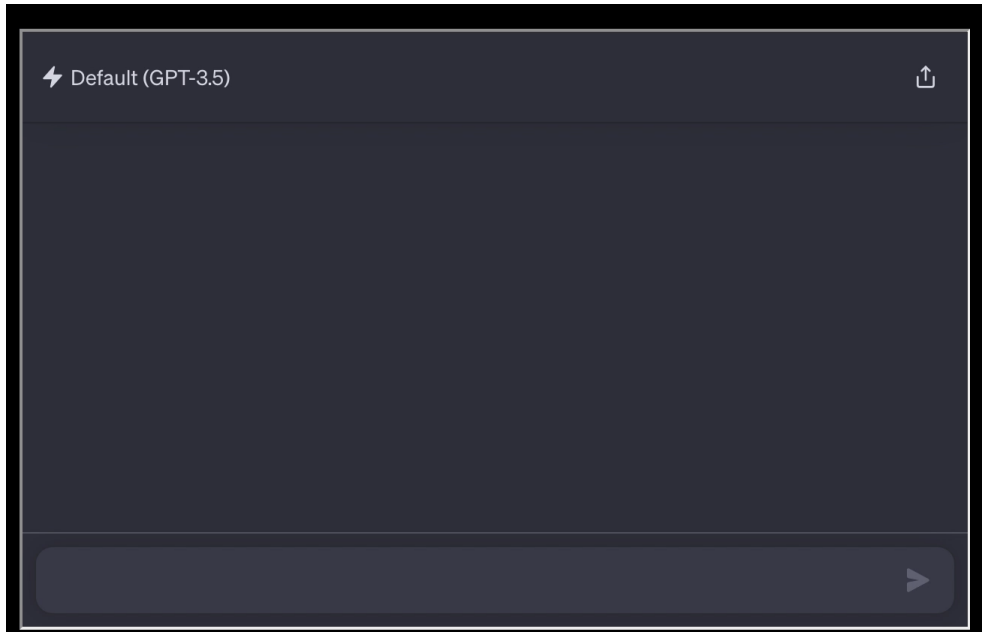
National Cyber Security Centre

ABOUT NCSC

Home Information for... Advice & guidance Education & skills Products & serv

BLOG POST

ChatGPT and large language models: what's the risk?



Default (GPT-3.5)

Regulation, frameworks, and standard operating procedures can help ensure responsible AI development.

European Parliament

EU AI Act: first regulation on artificial intelligence

Society Updated: 14-06-2023 - 14:06
Created: 08-06-2023 - 11:40

The use of artificial intelligence in the EU will be regulated by the AI Act, the world's first comprehensive AI law. Find out how it will protect you.

U.S. FOOD & DRUG ADMINISTRATION | Health Canada Santé Canada | Medicines & Healthcare products Regulatory Agency

Good Machine Learning Practice for Medical Device Development: Guiding Principles

October 2021

The U.S. Food and Drug Administration (FDA), Health Canada, and the United Kingdom's Medicines and Healthcare products Regulatory Agency (MHRA) have jointly identified 10 guiding principles that can inform the development of Good Machine Learning Practice (GMLP). These guiding principles will help promote safe, effective, and high-quality medical devices that use artificial intelligence and machine learning (AI/ML).

Artificial intelligence and machine learning technologies have the potential to transform health care by deriving new and important insights from the vast amount of data generated during the delivery of health care every day. They use software algorithms to learn from real-world use and in some situations may use this information to improve the product's performance. But they also present unique considerations due to their complexity and the iterative and data-driven nature of their development.

These 10 guiding principles are intended to lay the foundation for developing Good Machine Learning Practice that addresses the unique nature of these products. They will also help cultivate future growth in this rapidly progressing field.

The 10 guiding principles identify areas where the

Good Machine Learning Practice for Medical Device Development: Guiding Principles	
Multi-Disciplinary Expertise Is Leveraged Throughout the Total Product Life Cycle	Good Software Engineering and Security Practices Are Implemented
Clinical Study Participants and Data Sets Are Representative of the Intended Patient Population	Training Data Sets Are Independent of Test Sets
Selected Reference Datasets Are Based Upon Best Available Methods	Model Design Is Tailored to the Available Data and Reflects the Intended Use of the Device
Focus Is Placed on the Performance of the Human-AI Team	Testing Demonstrates Device Performance During Clinically Relevant Conditions
Users Are Provided Clear, Essential Information	Deployed Models Are Monitored for Performance and Re-training Risks are Managed

A balance between regulation and guidance is needed for AI

- Define problem and assess risk
- Understand data readiness and model design
- Develop and evaluate
- Deploy

Imperial College London

FIHealth

camo centre for antimicrobial optimisation

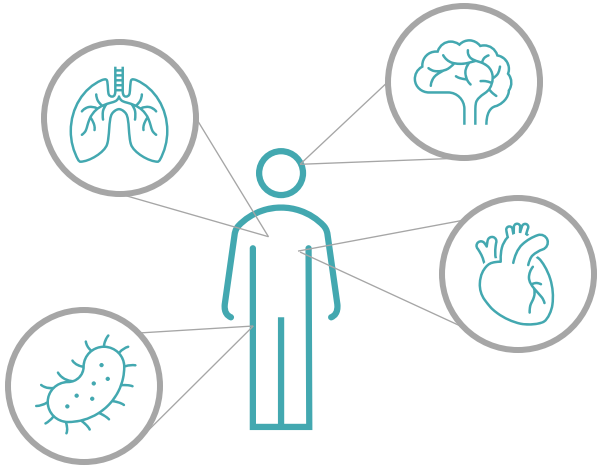
Comorbidity representation

William Bolton

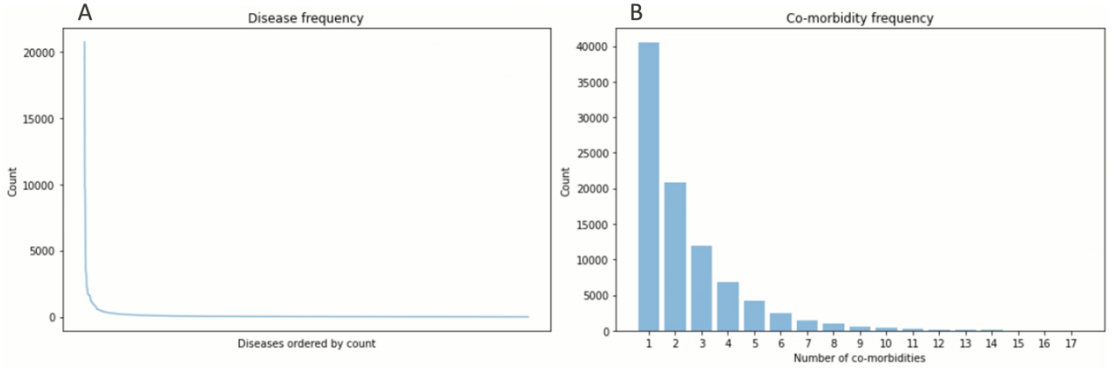
Viva

10th February 2025

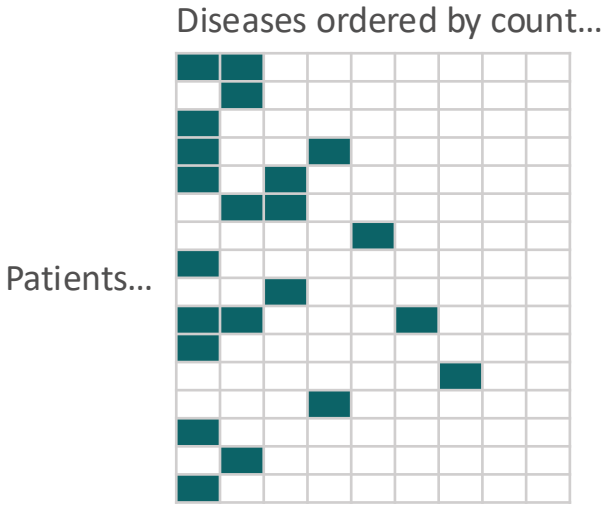
Diagnosis information can be hard to apply to AI systems.



Co-morbidities or chronic long-term medical conditions are a **major challenge in healthcare**



Challenges such as **combinatorial complexity**, **heterogeneity**, and a **lack of data** make using disease data in AI systems difficult



Existing AI research on co-morbid patients does not tackle these problems and therefore **lacks appropriate representation**

Aim

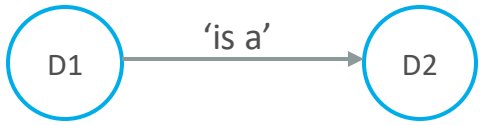
Creating **meaningful embeddings** from **external medically grounded knowledge**, to help **overcome such challenges** and **support downstream AI applications**

Our pipeline leverages a publicly available expert curated healthcare knowledge graph.

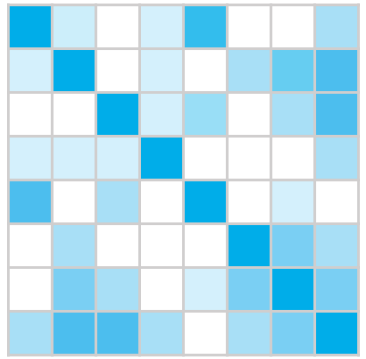
NHS Imperial College Healthcare NHS Trust

SNOMED distance

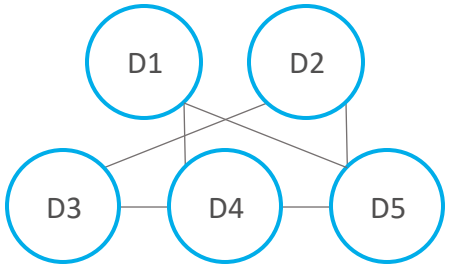
SNOMED CT
The global language of healthcare



Diseases



Graph creation

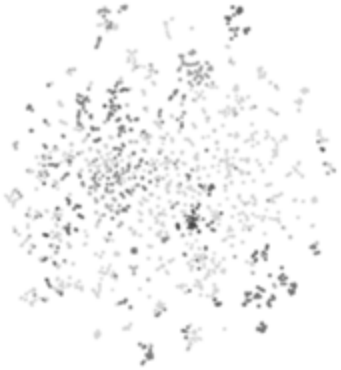


Weighting: $w = \frac{1}{s_{d_i d_j} \cdot f_s}$

where s is the SNOMED distance between two diseases and fs is the number of edges with that SNOMED distance

Edge pruning

Node2vec



Optimized for the mean SNOMED distance between each disease and their nearest neighbor

Imperial College London

We tested our methodologies against two clinically relevant AI tasks.

Supervised learning task

SNOMED disease embeddings for each patient → In-hosptial mortality within a year; Long length of stay

- Model: Set transformer
- Baseline:
  - Logistic regression with one hot encodings
  - Logistic regression with Charlson co-moribdity categories
  - Set transformer with random SNOMED disease embeddings
- Evaluation: AUROC

Similar patient retrieval task

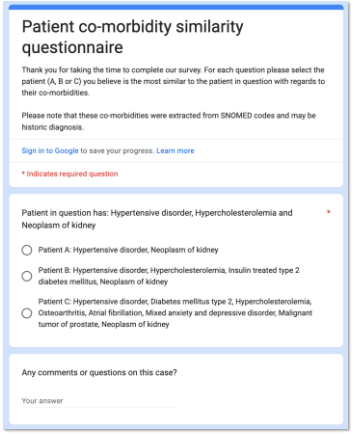
- Useful for **case-based reasoning**
- Created **co-morbid patient embeddings** as the **mean** of all the patients SNOMED disease embeddings
- Retrieved similar patients through **nearest neighbor lookup**
- Baseline:
  - One hot encodings
  - Rocheteau metric
- Evaluation:

Metrics

SNOMED similarity score

Charlson Jaccard index

Expert humans



Two novel metrics were created for the Similar patient retrieval task.

SNOMED similarity score

$$SNOMED\ sim_{p1,p2} = f(S_{p1,p2}) + f(S_{p2,p1})$$

where $S_{p1,p2}$ is a SNOMED distance matrix for the patients co-morbidities

We match each disease of $p1$ to a disease of $p2$ so that the matching minimized the following equation:

$$f(A) = \sum_{i-1}^{n} min_{j \in \{1,\ldots,m\}} \left(1 - \frac{1}{A_{ij} + 1}\right)$$

where $A \in \mathbb{R}^{n \times m}$

More similar patients return **smaller** values

Charlson Jaccard index

The Charlson co-morbidity index is a widely adopted clinical tool that classifies some specific co-morbidities to 17 different categories

$$Charlson\ Jaccard\ index_{p1,p2} = \frac{|C_{p1} \cap C_{p2}|}{|C_{p1} \cup C_{p2}|}$$

where C represents the set of Charlson co-morbidities for a particular patient

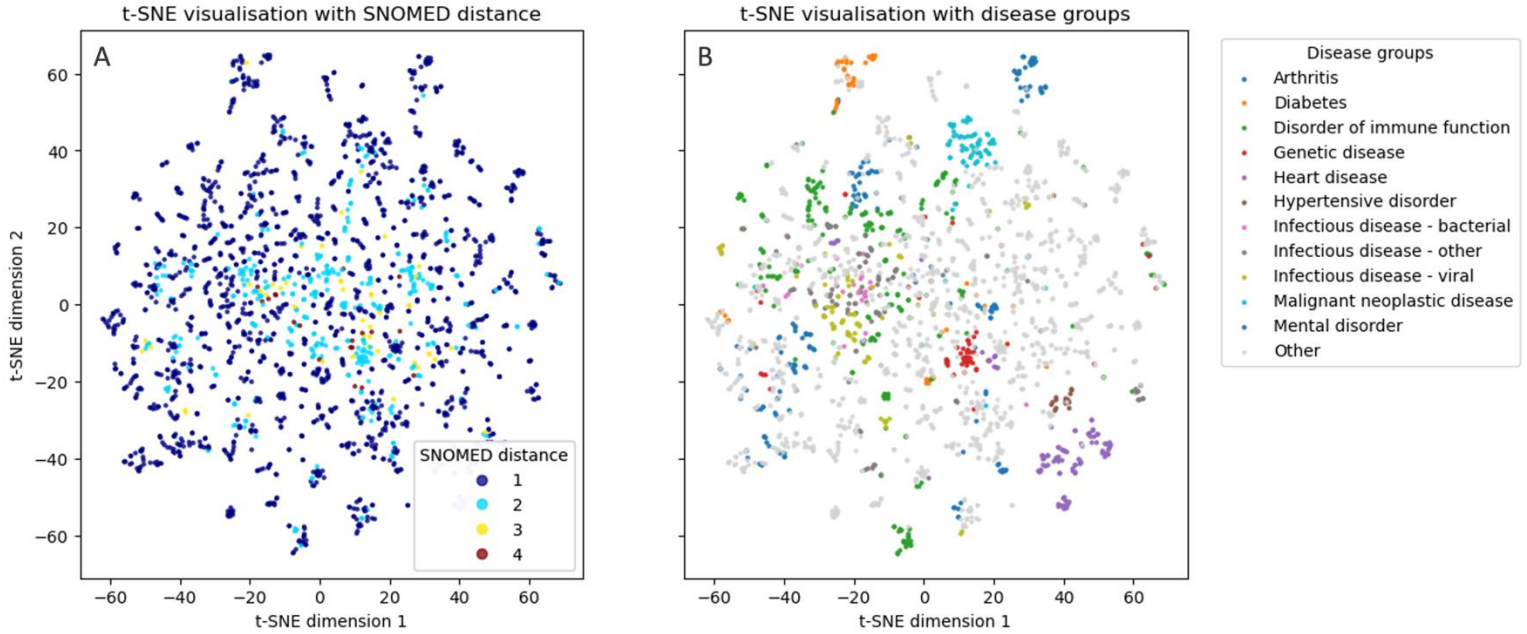
More similar patients return **larger** values

Our pipeline generated meaningful SNOMED disease embeddings.

Imperial College Healthcare NHS Trust

95,157 patients

2,133 chronic conditions



- **Optimization of pre-processing** returned a mean SNOMED distance between each node and their nearest neighbor of **1.23**
- **Related conditions are located close in the embedding space**
- For example:
  - Viral and bacterial infectious diseases
  - Heart diseases and hypertensive disorders

Imperial College London

SNOMED disease embeddings are informative features for models undertaking clinically relevant predictions.

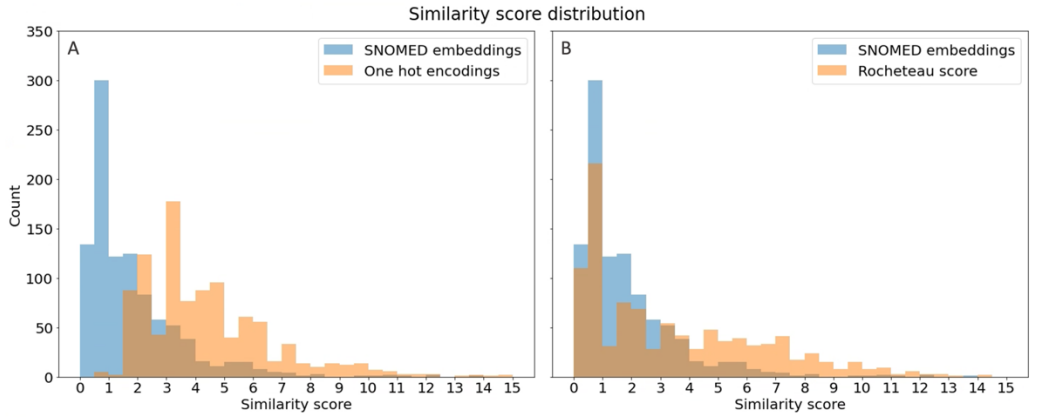
Table 1: Mean unseen test set AUROC results for supervised learning classification tasks in different populations.

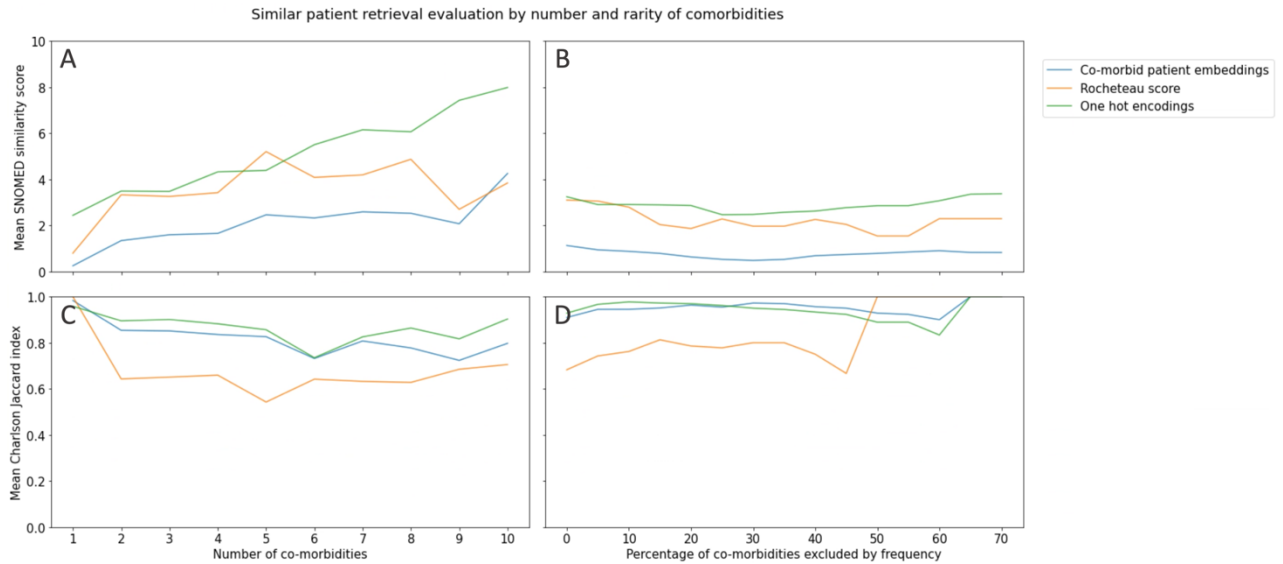
		Year Mortality		Long length of stay	
Features	Model	Overall	Rarest co-morbidities	Overall	Rarest co-morbidities
Charlson co-moribdity categories	Logistic regression	0.65 (SD 0.01)	0.50 (SD <0.01)	0.60 (SD 0.01)	0.50 (SD 0.03)
One hot encodings	Logistic regression	0.79 (SD 0.02)	0.80 (SD 0.23)	0.72 (SD 0.01)	0.55 (SD 0.11)
Random SNOMED disease embeddings	Set transformer	0.80 (SD 0.03)	0.56 (SD 0.33)	0.74(SD 0.02)	0.52 (SD 0.23)
SNOMED disease embeddings	Set transformer	**0.82 (SD 0.02)**	**0.85 (SD 0.14)**	**0.75 (SD 0.01)**	**0.61 (SD 0.20)**

Co-morbid patient embeddings finds more similar patients and is consistent across different degrees of rarity and multi-morbidity.

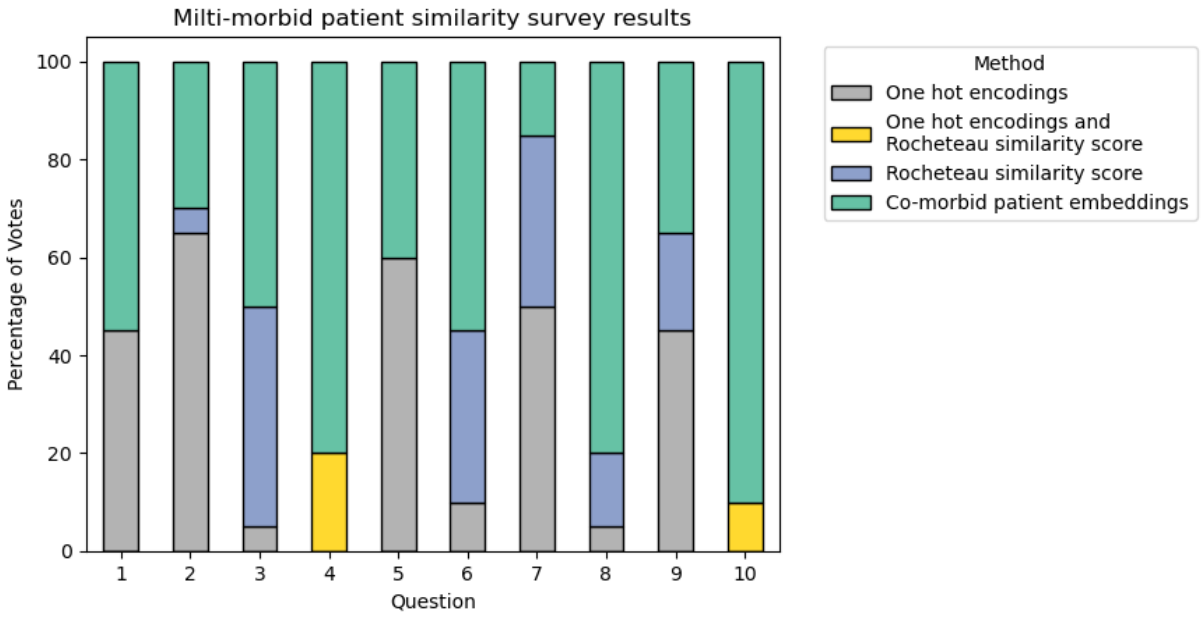
Table 2: Mean results for the similar patient retrieval task.

Method	SNOMED similarity score	Charlson Jaccard index
One hot encodings	4.40 (SD 2.32)	**0.88 (SD 0.30)**
Rocheteau's method	3.52 (SD 3.26)	0.69 (SD 0.20)
Co-morbid patient embeddings	**1.78 (SD 1.90)**	0.84 (SD 0.34)





Patients identified by our method were selected as the most similar in 60% of questions with a mean winning margin of 40%.



	Co-morbidities			
Question 8 patient	Gestational diabetes mellitus	Hypertensive disorder	Pre-eclampsia	Varicella
Co-morbid patient embeddings	Gestational diabetes mellitus	Pregnancy-induced hypertension	Pre-eclampsia	Varicella
Rocheteau score	Gestational diabetes mellitus	Hypertensive disorder	-	Varicella
One hot encodings	Gestational diabetes mellitus	-	Pre-eclampsia	Varicella
Question 10 patient	Osteo-arthritis	Alcoholism		
Co-morbid patient embeddings	Osteo-arthritis	Alcohol dependence		
Rocheteau score	Osteo-arthritis	Alcoholism	Peripheral nerve entrapment	
One hot encodings	Osteo-arthritis	Alcoholism	Peripheral nerve entrapment	

Identical
Similar
Dissimilar

Imperial College London

FHi Health

camo centre for antimicrobial optimisation

Cessation CDSS

William Bolton

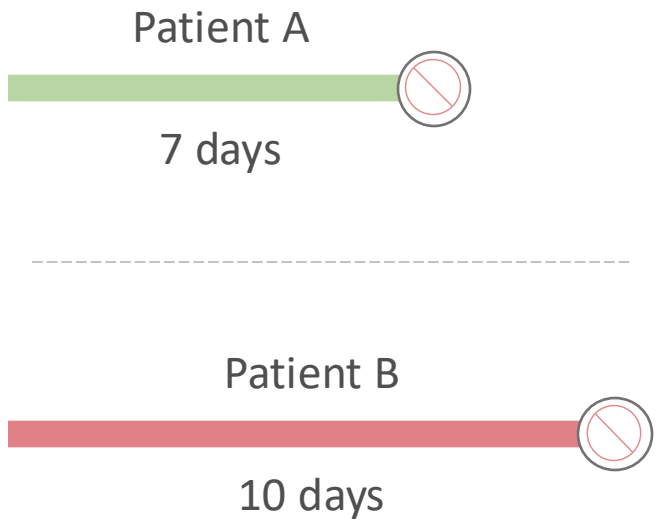
Viva

10th February 2025

Antibiotic cessation decision making is complex and under-researched.

One key challenge when treating a patient who has a bacterial infection is determining when it is appropriate to stop antibiotic treatment

Numerous studies have shown that on a population level, shorter treatment durations are often non-inferior to longer ones



There is a poor understanding of the factors that facilitate or inhibit an individual from receiving a short duration of therapy

Aim

Utilise a machine learning and synthetic control-based approach to estimate patients total white blood cell count for any given day, if they were to stop vs. continue antibiotic treatment

Machine learning and synthetic outcome estimation for individualised antimicrobial cessation.

1 DATASET

Filter relevant patient stays from MIMIC-IV, **extract and aggregate features** including lab test results, clinical parameters, ventilation settings and demographics to create a **regular temporal dataset** for estimation of their **length of stay and mortality outcomes**.

MIMIC-IV
>40,000 ICU patients

ICU intravenous antibiotic treatment (1<days<21)

OUR DATASET
18,988 ICU patients 22,845 unique stays

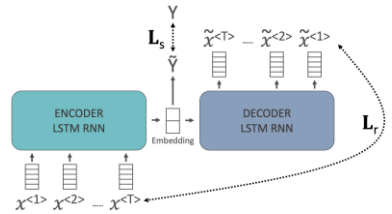
43 FEATURES

MORTALITY

LENGTH OF STAY (LOS)

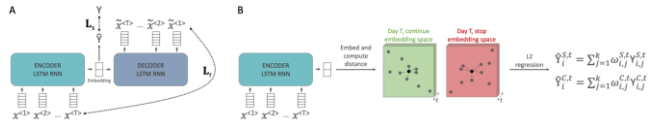
2 AUTO ENCODER

Train autoencoders to create an **embedding that is representative** of the patients' temporal features and a **linear predictor** of their outcomes.



3 SYNTHETIC OUTCOME ESTIMATION

Apply the **adapted synthetic control methodology** to **estimate patient outcomes** if they were to **stop vs continue treatment** on each antibiotic day within their ICU stay.

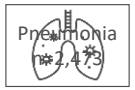


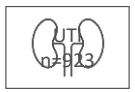
4 EVALUATION AND VALIDATION

Evaluate stop and continue estimations through **'impact' and 'control' days**. Validate the model through numerous tests and application to **pneumonia and UTI datasets**.

STOP IMPACT	STOP CONTROL
CONTINUE CONTROL	CONTINUE IMPACT

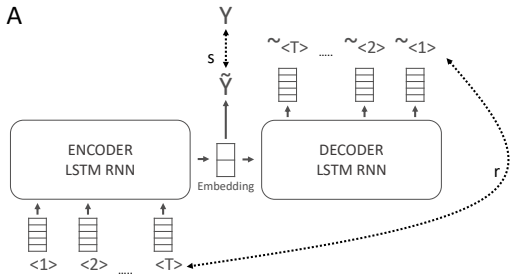
Metrics:
- Mean delta
- Wilcoxon rank-sum test
- RMSE
- AUROC





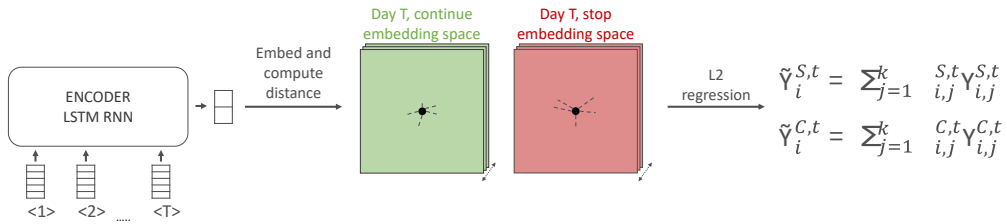
AUTOENCODER TRAINING

A



SYNTHETIC OUTCOME ESTIMATION

B

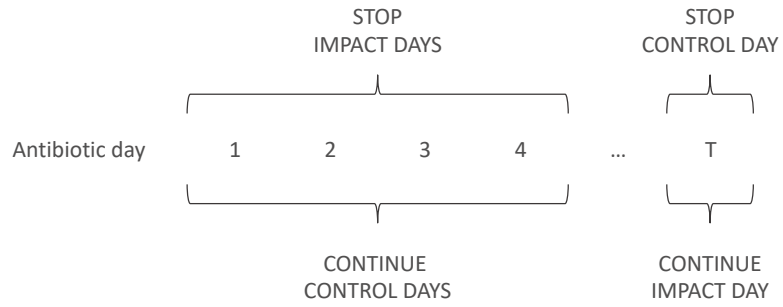


Machine learning and synthetic outcome estimation for individualised antimicrobial cessation.

AUTOENCODER PREDICTIONS

	Metric	Result
Mortality Classification	AUROC	0.77 (95% CI 0.73–0.80)
	Accuracy	0.73 (95% CI 0.71–0.75)
	Precision	0.44 (95% CI 0.36–0.46)
	Recall	0.67 (95% CI 0.61–0.72)
	F1 Score	0.75 (95% CI 0.72–0.78)
	AUPRC	0.55 (95% CI 0.42–0.56)
LOS Regression	RMSE	3.88 (95% CI 3.84–3.92)

SYNTHETIC OUTCOME ESTIMATION



SCENARIO	DAY(S)	LOS				Mortality		
		Mean delta (days, p-value)	MAPE	MAE	RMSE	Mean delta	MAE	AUROC
STOP	IMPACT	2.71*, <0.01	0.36	3.30	4.80	0.06	0.25	0.66
	CONTROL	0.24, 0.60	0.26	1.32	1.93	0.05	0.15	0.72
CONTINUE	IMPACT	-2.09*, <0.01	0.77	2.85	3.16	0.05	0.18	0.67
	CONTROL	0.42*, 0.01	0.48	2.72	3.76	0.07	0.24	0.64

frontiers in Digital Health

Routinely collected electronic health record data and an autoencoder were used for white blood cell count prediction.

DATASET

MIMIC-IV
>40,000 ICU patients

ICU intravenous antibiotic treatment (days<8)

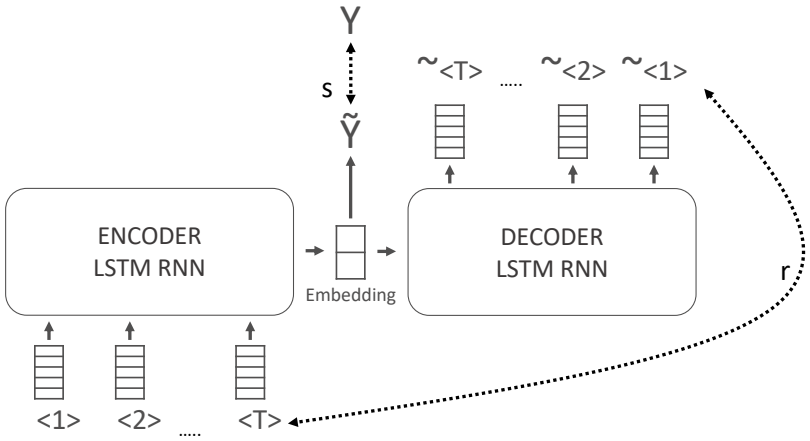
OUR DATASET
7,867 unique ICU stays

77
features

Total white blood cell count (WBC)

MODEL

Autoencoder



The encoder is trained using both a supervised loss (L_s) and reconstruction loss (L_r)

The autoencoder achieves reasonable WBC prediction performance and can be used for synthetic scenario estimation.

AUTOENCODER PREDICTIONS

77 features

Evaluation metric	Value
RMSE	3.28
MAE	2.68
MAPE	0.31

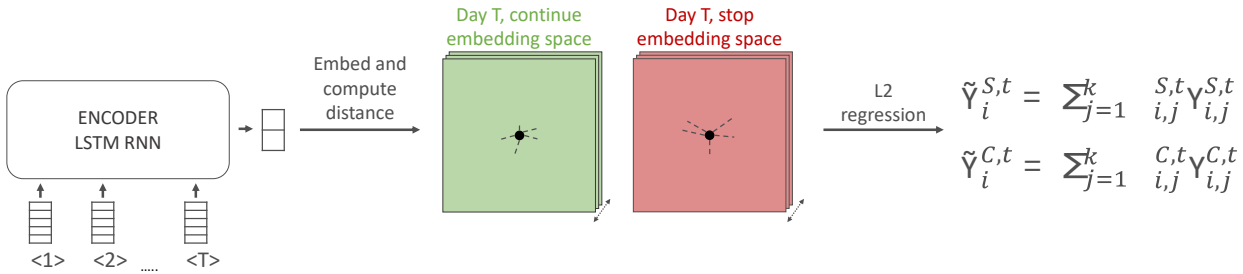
Feature selection

22 features

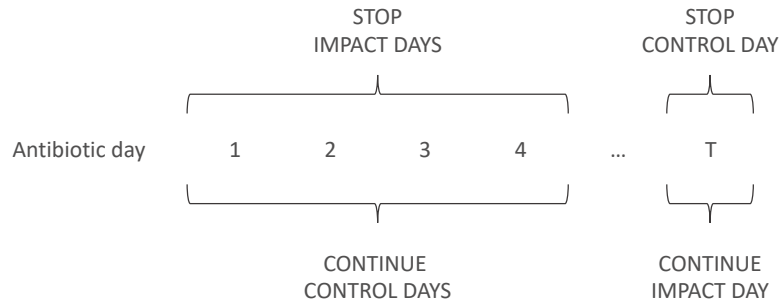
Evaluation metric	Value
RMSE	3.33
MAE	2.66
MAPE	0.36

SYNTHETIC ESTIMATION

Methodology



Evaluation



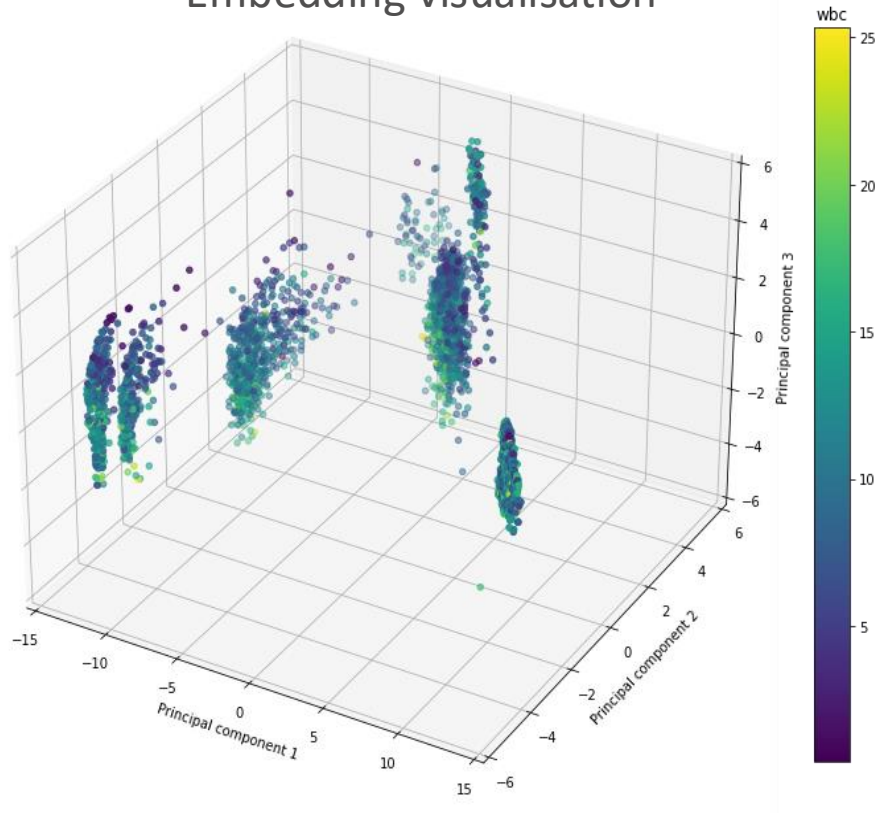
Our model can estimate patients white blood cell count under alternative antibiotic treatment.

SYNTHETIC WBC ESTIMATION RESULTS

Evaluation table

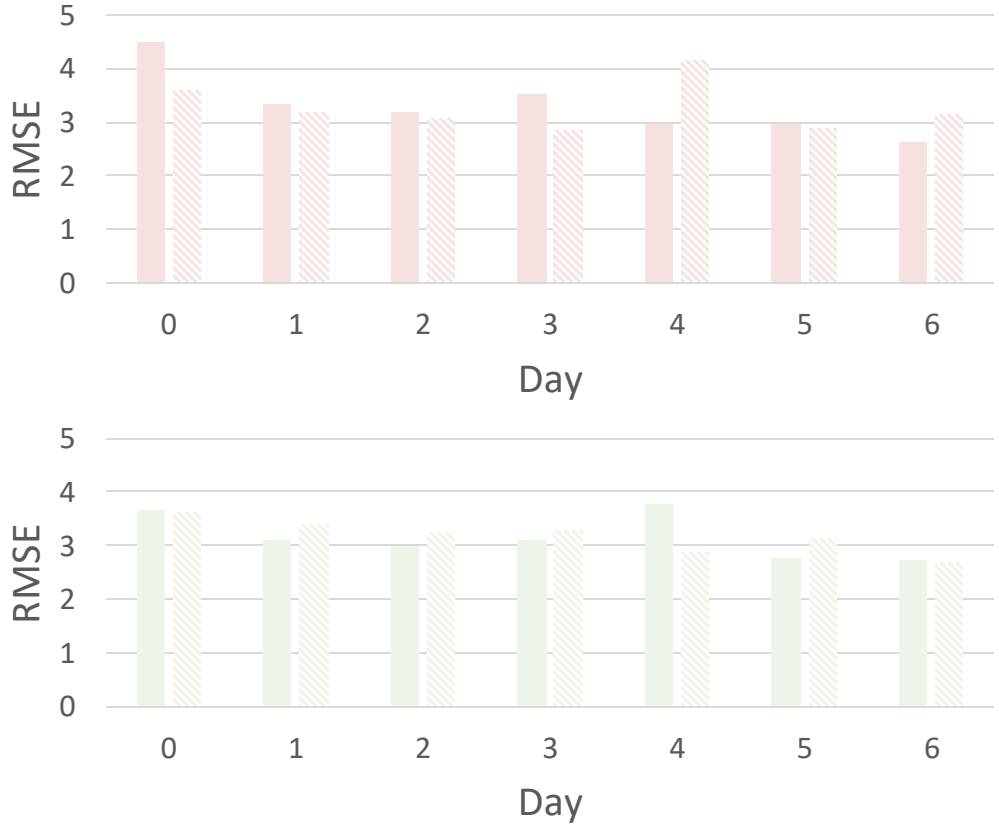
Scenario	Day(s)	WBC			
		Mean delta (days, p-value)	MAPE	MAE	RMSE
Stop	Impact	0.71*, <0.01	0.34	2.99	3.89
	Control	0.00, 0.06	0.32	2.50	3.24
Continue	Impact	-0.31*, <0.01	0.33	2.53	3.18
	Control	0.02*, 0.01	0.32	2.70	3.44

Embedding visualisation



Machine learning and synthetic outcome estimation for individualised antimicrobial cessation.

CONSISTENT ESTIMATION RESULTS



SYNTHETIC OUTCOME ESTIMATION

43 features

Mortality

Length of stay

frontiers in Digital Health

SCENARIO	DAY(S)	LOS				Mortality		
		Mean delta (days, p-value)	MAPE	MAE	RMSE	Mean delta	MAE	AUROC
STOP	IMPACT	2.71*, <0.01	0.36	3.30	4.80	0.06	0.25	0.66
	CONTROL	0.24, 0.60	0.26	1.32	1.93	0.05	0.15	0.72
CONTINUE	IMPACT	-2.09*, <0.01	0.77	2.85	3.16	0.05	0.18	0.67
	CONTROL	0.42*, 0.01	0.48	2.72	3.76	0.07	0.24	0.64

Imperial College London

FIHealth

camo centre for antimicrobial optimisation

Switch CDSS

William Bolton

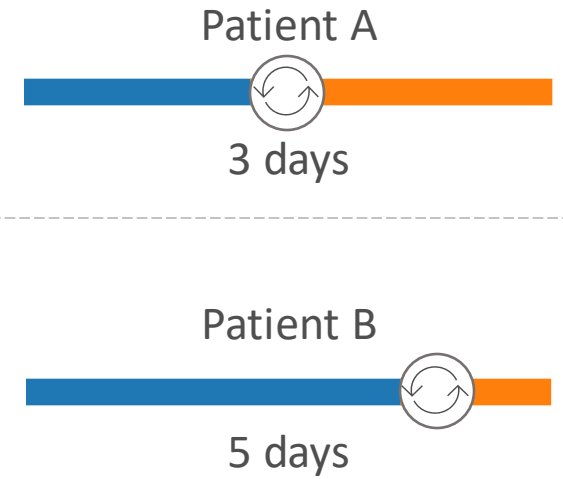
Viva

10th February 2025

Switching from IV-to-oral antibiotic treatment is complex and under-researched.

One key challenge of stewardship is **determining when to switch** antibiotics from **IV-to-oral administration**

Numerous studies have shown that **oral therapy can be non-inferior to IV**



There is a **poor understanding** of the factors that facilitate or inhibit an individual from receiving oral therapy

Aim

Utilise a **machine learning** and **routinely collected clinical parameters** to predict whether a patient could be **suitable for switching** from IV-to-oral antibiotics on **any given day**

Imperial College London

Routinely collected electronic health record data were used, with clinical guided features.

DATASET

MIMIC dataset
n=8,694

eICU dataset
n=1,668

FEATURES

UK Health Security Agency

Antimicrobial Intravenous-to-Oral Switch (IVOS) Decision Aid

Based on the National Antimicrobial IVOS Criteria

Co-produced through a UK-wide multidisciplinary consensus process involving 279 participants

Open Access | Published: 09 August 2019

catch22: CAnonical Time-series CHaracteristics

Selected through highly comparative time-series analysis

Carl H. Lubba, Sarab S. Sethi, Philip Knaute, Simon R. Schultz, Ben D. Fulcher & Nick S. Jones

Data Mining and Knowledge Discovery **33**, 1821–1852 (2019) | Cite this article

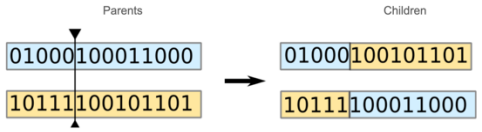
17k Accesses | **97** Citations | **34** Altmetric | Metrics

FEATURE SELECTION

1 SHAP Values

SHAP

2 Genetic algorithm

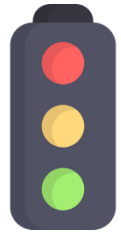


MODEL SELECTION

1 Hyperparameter optimization

OPTUNA

2 Cutoff point



The model achieves generalisable performance across a range of datasets and patient populations.

MIMIC

Metric	1ST threshold results	2nd threshold results	IVOS criteria baseline
AUROC	**0.78 (SD 0.02)**	0.69 (SD 0.03)	0.66
FPR	0.25 (SD 0.02)	**0.10 (SD 0.02)**	0.43

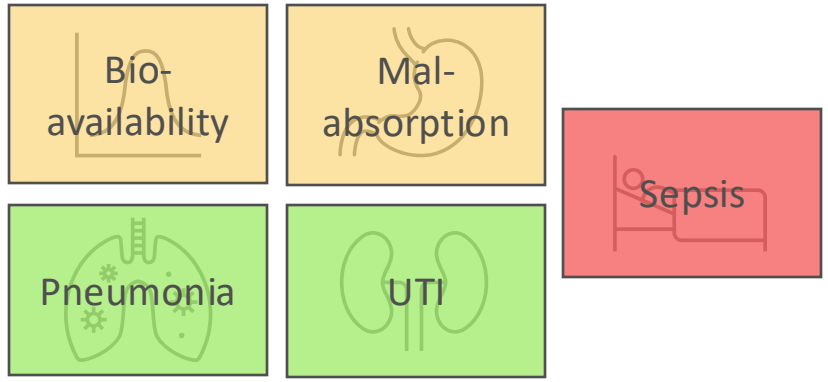
eICU

Metric	1st threshold results	2nd threshold results	IVOS criteria baseline
AUROC	**0.72 (SD 0.02)**	0.65 (SD 0.05)	0.55
FPR	0.24 (SD 0.04)	**0.05** (SD 0.02)	0.28

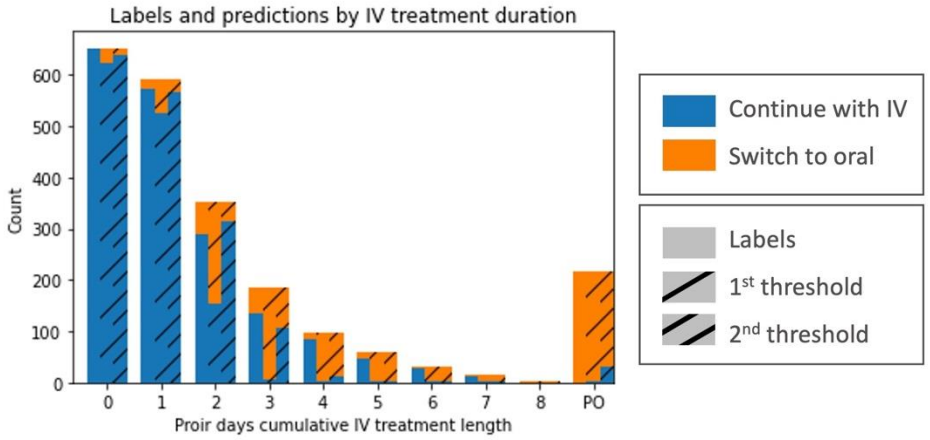
Imperial College Healthcare NHS Trust

Metric	Results	Prospective data
AUROC	0.78 (SD 0.01)	0.77
FPR	0.23 (SD 0.02)	0.46

SUBGROUPS



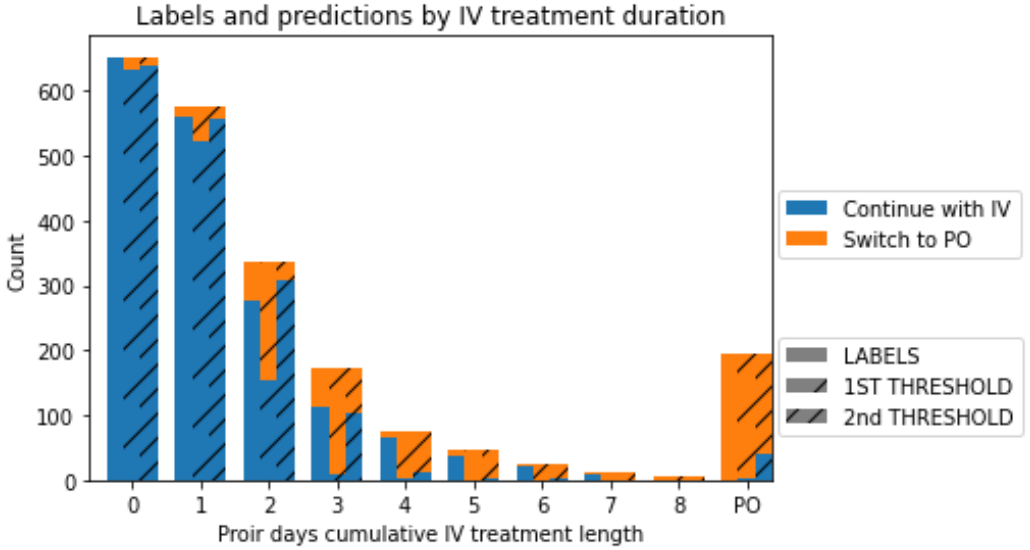
ANALYSIS

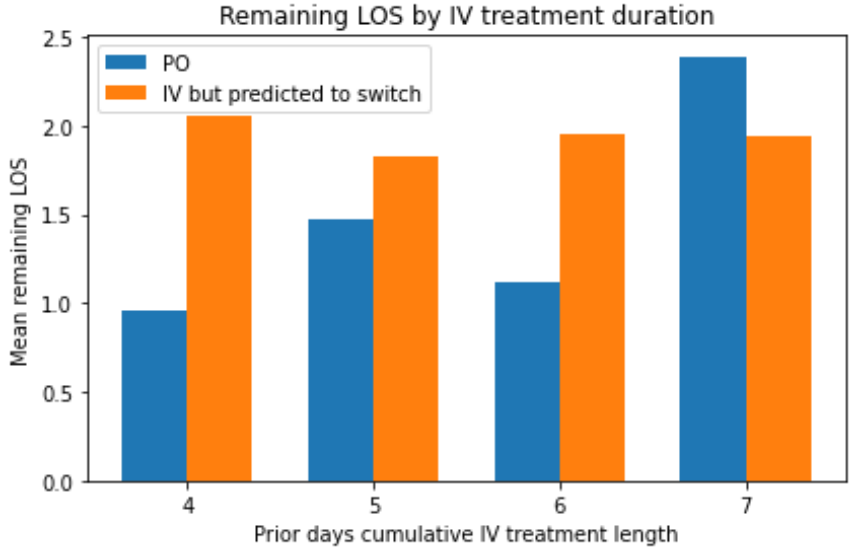


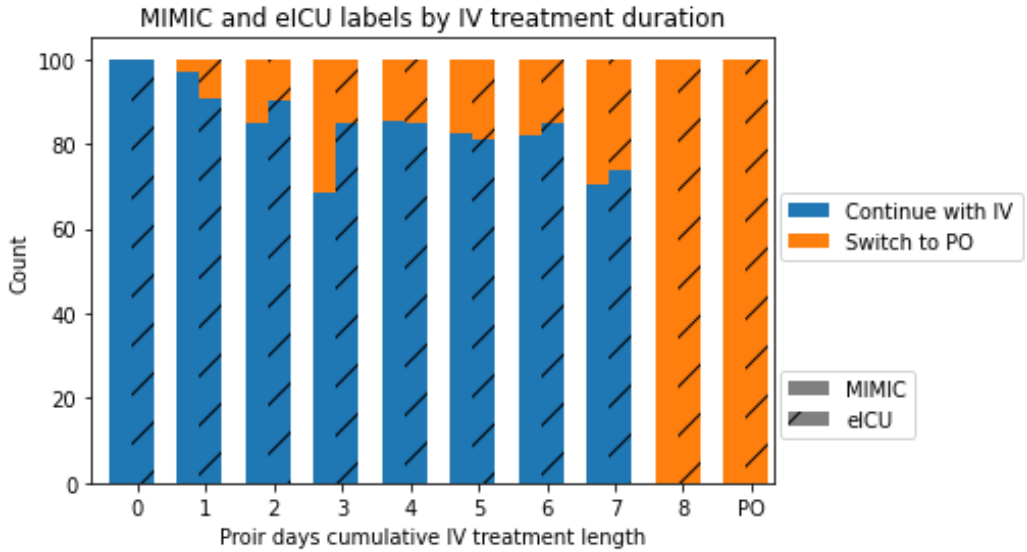
Models predict some patients could be **suitable for switching to oral administration earlier**

Imperial College London

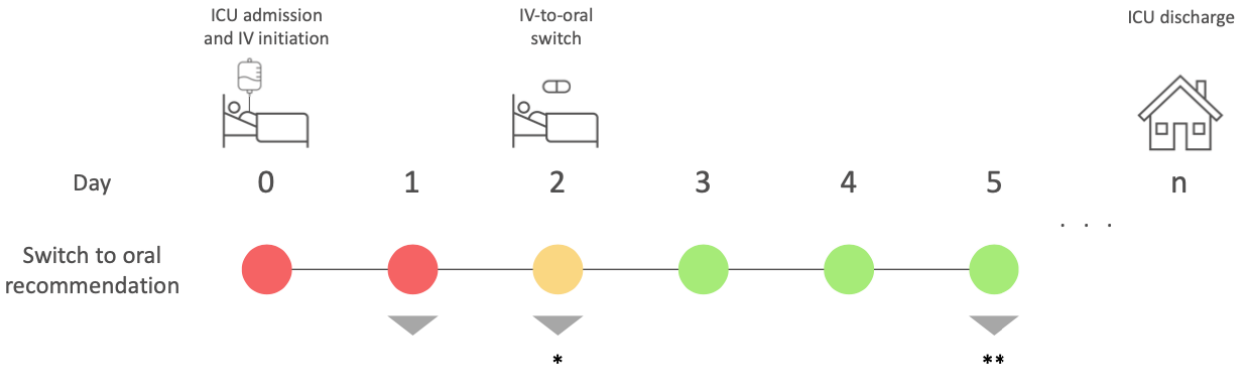
Fair interpretable machine learning for individualised IV to oral switch decision making.







Traffic light recommendations and informative visual representations improve model interpretability.



Day 1

Highlights
- Both thresholds predict switching is likely **not appropriate** at this time
- Predictions were correct for **100%** of similar examples
- O2 saturation pulseoximetry (feature 4) was of particular interest for these predictions

			Feature					Switch to oral label	Switch to oral prediction	
		Importance	1	2	3	4	5		1st threshold	2nd threshold
Patient		-	0.32	0.51	0.37	0.50	0.41	0	0	0
Example	1	0.28	0.38	0.54	0.29	0.48	0.46	0	0	0
	2	0.25	0.31	0.55	0.28	0.51	0.50	0	0	0
	3	0.21	0.29	0.52	0.45	0.52	0.46	0	0	0
	4	0.13	0.32	0.55	0.36	0.51	0.00	0	0	0

* Day 2

Highlights
- **Clinical guidance should be sought**, model thresholds disagree on whether switching could be appropriate or not at this time
- Predictions were correct for **50%** of similar examples (0% for the 1st threshold and 100% for the 2nd threshold)
- O2 saturation pulseoximetry (feature 4) was of particular interest for these predictions

			Feature					Switch to oral label	Switch to oral prediction	
		Importance	1	2	3	4	5		1st threshold	2nd threshold
Patient		-	0.24	0.25	0.28	0.43	0.77	1	1	0
Example	1	0.38	0.25	0.20	0.25	0.42	0.73	0	1	0
	2	0.12	0.21	0.12	0.20	0.43	0.85	0	1	0

** Day 5

Highlights
- Both thresholds predict switching could be **appropriate** at this time
- Predictions were correct for **75%** of similar examples (75% for the 1st threshold and 75% for the 2nd threshold)
- Systolic blood pressure (feature 1) and O2 saturation pulseoximetry (feature 4) were of particular interest for these predictions

			Feature					Switch to oral label	Switch to oral prediction	
		Importance	1	2	3	4	5		1st threshold	2nd threshold
Patient		-	0.16	0.49	0.45	0.37	0.59	1	1	1
Example	1	0.21	0.20	0.58	0.39	0.37	0.45	1	1	1
	2	0.20	0.15	0.47	0.43	0.36	0.70	1	1	1
	3	0.16	0.16	0.43	0.48	0.36	0.76	1	1	1
	4	0.15	0.18	0.49	0.42	0.38	0.59	0	1	1

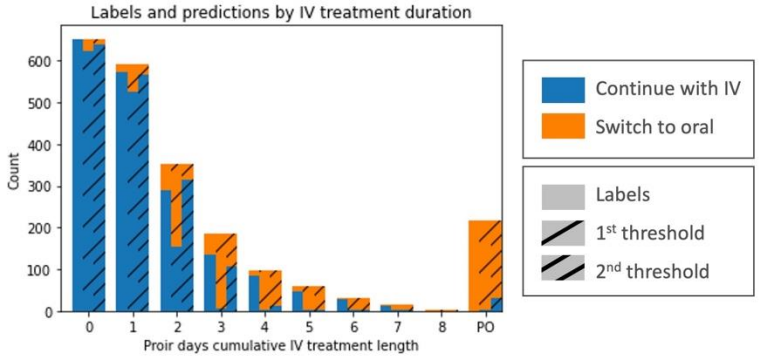
Note this system does not cover all aspects of the switch decision making process and should only be used as decision support to highlight when a patient may be suitable for switch assessment

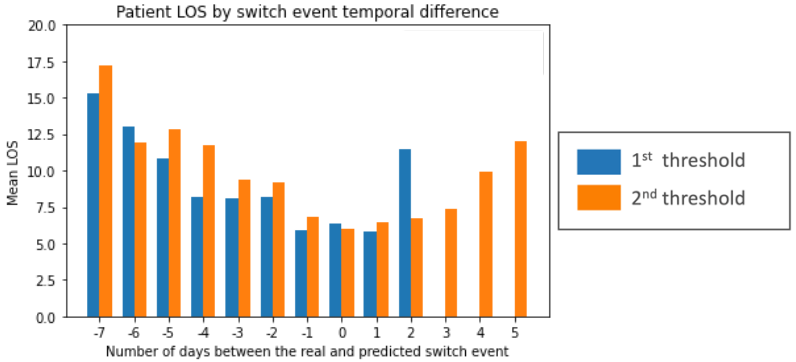
Imperial College London

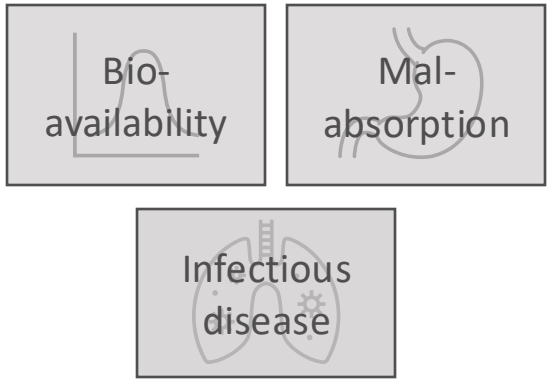
Models demonstrate reasonably fair performance and threshold optimisation can improve results.

Sensitive attribute	Group	Equalised odds demonstrated	
		Initially	With threshold optimisation
Sex	Female	✓	-
	Male	✓	-
Age	20	✓	✗
	30	✓	✓
	40	✓	✓
	50	✓	✓
	60	✓	✓
	70	✓	✓
	80	✓	✓
	90	✗	✓
Race	Asian	✓	✓
	Black	✓	✓
	Hispanic	✓	✓
	Native	✗	✗
	Other	✓	✓
	Unknown	✓	✓
	White	✓	✓
Insurance	Medicaid	✗	✓
	Medicare	✓	✓
	Other	✓	✓

Such technology could provide appropriate decision support and promote switching when appropriate.







Models predict some patients could be **suitable for switching to oral administration earlier** from a clinical parameter, health status perspective

When the difference between the real and predicted switch event was minimal, mean patient **LOS outcomes were lower**

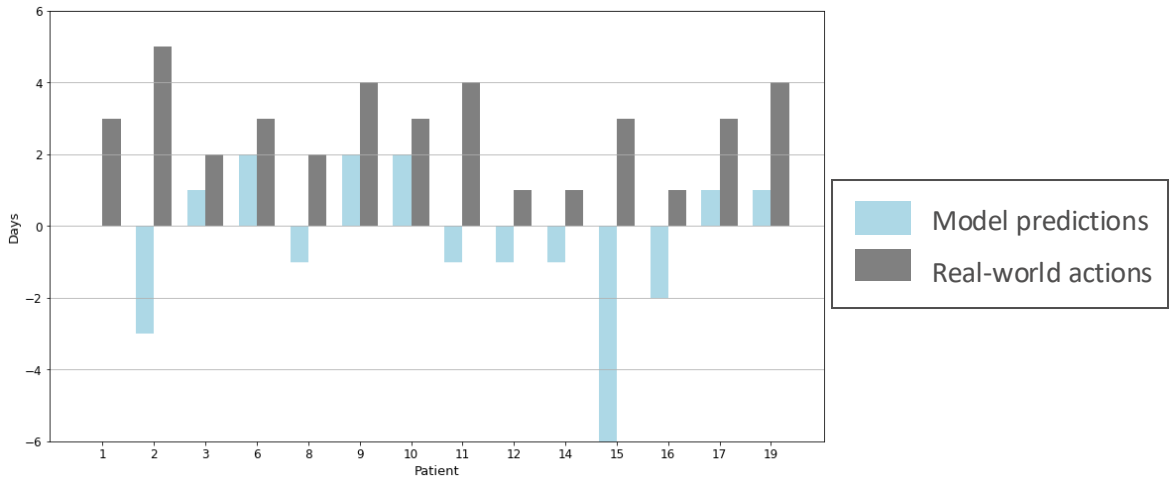
Models only analyse a **snapshot of the patient** and **not all factors** that are clinically used to assess a patient's suitability for switching

Prospective evaluation was conducted on 40 patients at Imperial NHS Trust against gold-standard pharmacists' recommendations.

Metric	Result
AUROC	0.68
Accuracy	0.70
FPR	0.28

Temporal difference	Percentage
Early	37.50
Same day	32.50
Late	30

Overall absolute temporal difference of 1.23 days (SD 1.42)



TAKAWAYS

- **Reasonable performance** on small patient sample with a slight preference for **early predictions**
- Such prospective evaluation is essential for highlighting the AI systems **successes and pitfalls**

Imperial College London

FIHealth

camo centre for antimicrobial optimisation

Clinician evaluation

William Bolton

Viva

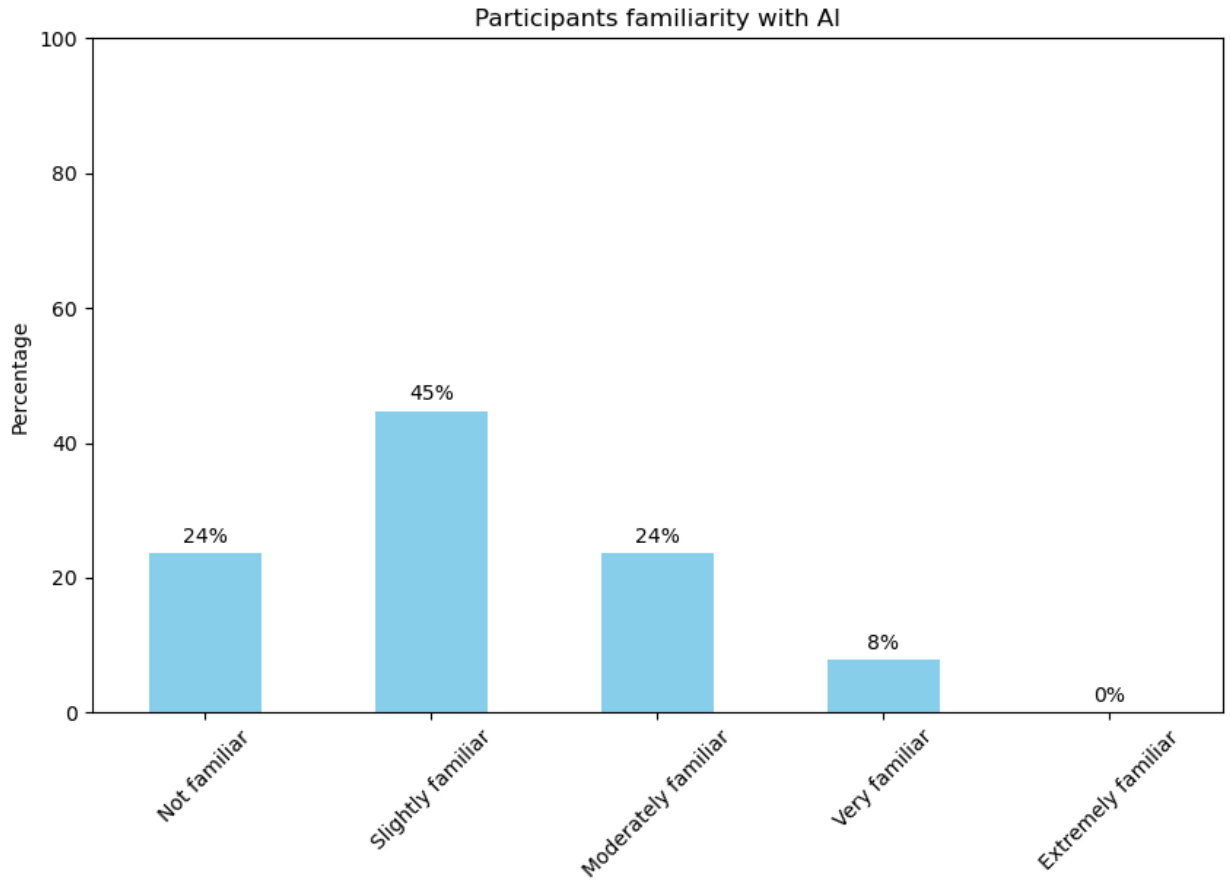
10th February 2025

Imperial College London

Forty-two clinicians completed the study with most UK consultant level and specializing in infectious disease.

42 PARTICIPANTS

Demographic	No. (%) of participants
Grade	
Consultant	26 (61.90%)
Other	16 (38.10%)
Medical Speciality	
Infectious Diseases	14 (33.33%)
Microbiology	12 (28.57%)
Pharmacist	6 (14.29%)
Other	10 (23.81%)
Sex	
Male	23 (54.76%)
Female	19 (45.24%)
Age	
20s	4 (9.52%)
30s	17 (40.48%)
40s	12 (28.57%)
50s	6 (14.29%)
60s	2 (4.76%)
Prefer not to say	1 (2.38%)
Age Statistics	
Mean	42
Median	37
Standard Deviation	8.84



The AI-IVOS CDSS has a greater impact on clinician's decision making when it recommends don't switch.

NO DIFFERENCES

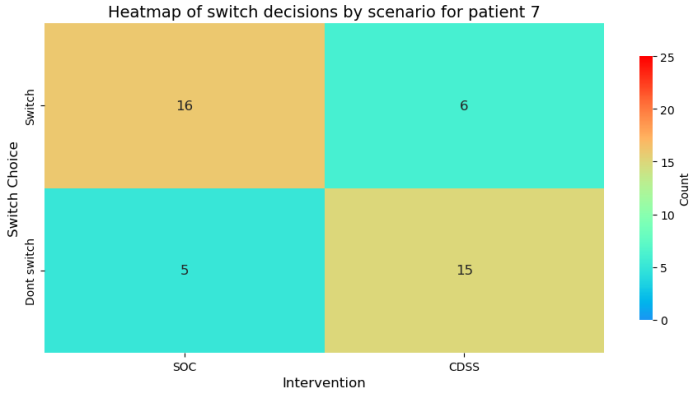
11/12 cases

TAKAWAYS

- Prescribing decisions are **nuanced** and **complex**
- Clinicians could correctly identify incorrect AI recommendations and **ignore the support**

STATISTICALLY SIGNIFICANT DIFFERENCES

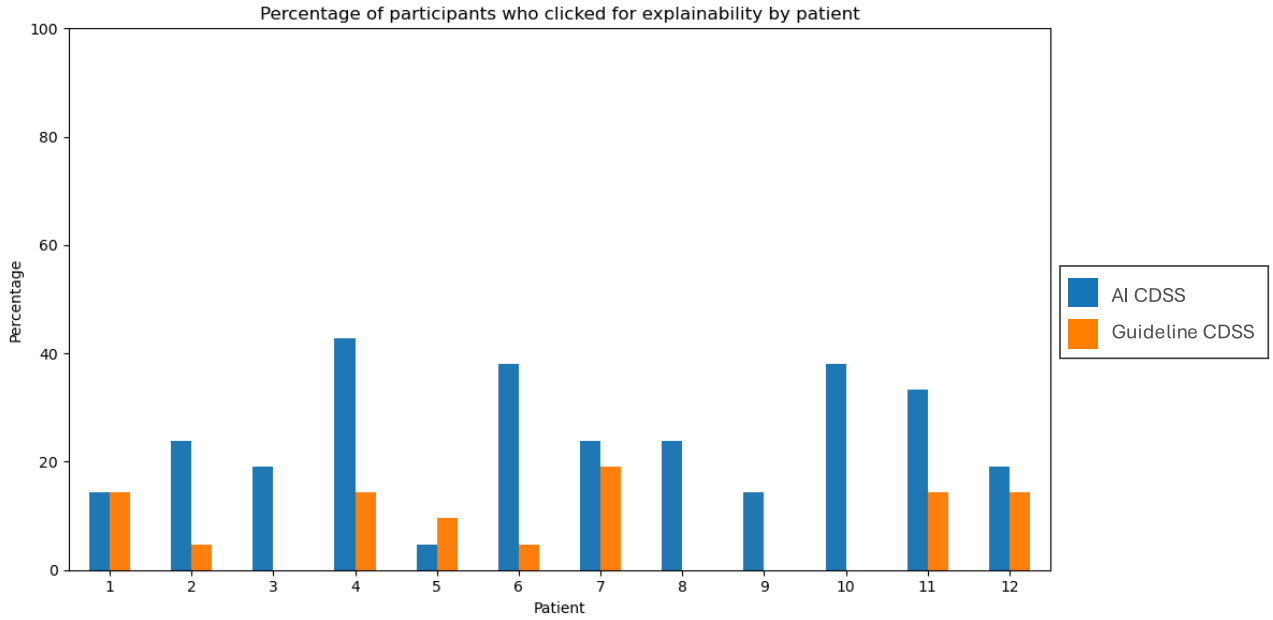
Patient 7 and subgroup analysis for 3 patients



- The system has a **greater influence** on clinicians when it recommends **don't switch**
- Participants were **persuaded to take the option perceived as safer** which matches the culture of cautiousness (40%) and hesitancy to change (24%) reported in interviews.

Decision support explanations had little impact on clinicians' decisions.

	Percentage of participants that clicked	
	At any time	When available
AI CDSS explanations	55%	9%
Guideline CDSS explanations	29%	5%



TAKAWAYS

- CDSS explanations were **not frequently used when available**
- Implying that it was the **presence of an AI recommendation** itself driving changes in decision making
- Explainability methods including displaying **similar** patients (including their outcomes) and summarizing information in **free text** warrants further research

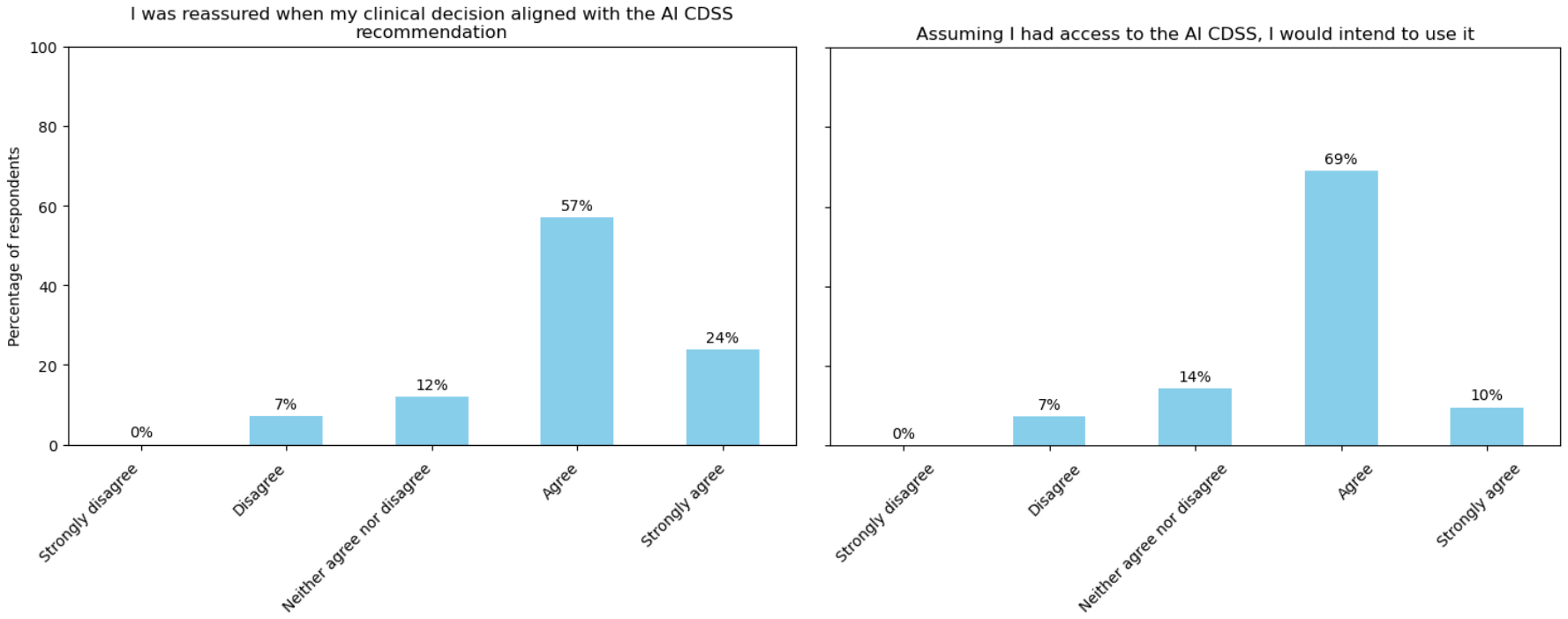
The technology was useful, usable and perceived well by participants.

System Usability score: 72.32 / 100

Perceived usefulness: 3.59 / 5

Perceived ease of use: 3.83 / 5

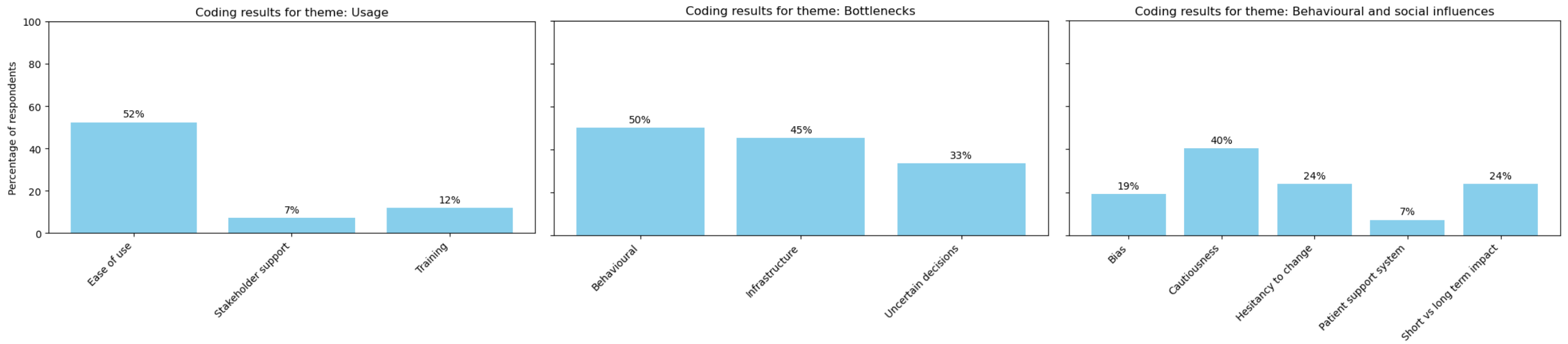
Self efficacy: 4.05 / 5



TAKAWAYS

- Clinicians are **reassured** and more confident when given AI recommendations
- The AI CDSS was well perceived being **useful, usable and appropriate**

Numerous barriers exist to implementing AI technology with evidence and the system being easy to use essential for adoption.



TAKAWAYS

- There was disagreement on whether healthcare institutions have the necessary **infrastructure to support AI technology**
- **Behavioural factors** such as cautiousness play an important role in antimicrobial prescribing and can act as a **barrier** to technology adoption
- Trust in CDSSs is built through **clinical evidence**, with **ease of use** being essential for adoption

Imperial College London

Case vignette details.

Patient	Intervention		AI CDSS recommendation	Guideline CDSS recommendation	Ground Truth	Data	Group
	Arm A	Arm B					
1	CDSS	SOC	Don't switch	Don't switch	Don't switch	ICHT	AI correct
2	SOC	CDSS	Switch	Switch	Switch	ICHT	AI correct
3	SOC	CDSS	Switch	Don't switch	Don't switch	ICHT	AI incorrect
4	CDSS	SOC	Don't switch	Switch	Switch	ICHT	AI incorrect
5	SOC	CDSS	Switch	Don't switch	-	ICHT	Uncertain
6	CDSS	SOC	Don't switch	Switch	-	ICHT	Uncertain
7	CDSS	SOC	Don't switch	Don't switch	Didn't switch	MIMIC	AI correct
8	SOC	CDSS	Switch	Switch	Switched	MIMIC	AI correct
9	SOC	CDSS	Potentially switch	Switch	Didn't switch	MIMIC	Uncertain
10	CDSS	SOC	Don't switch	Switch	Switched	MIMIC	AI incorrect
11	CDSS	SOC	Switch	Don't switch	Didn't switch	MIMIC	AI incorrect
12	SOC	CDSS	Potentially switch	Don't switch	Switched	MIMIC	Uncertain

Table 1: Details of the experimental setup and case vignettes.

Imperial College London

Shannon entropy results from the case vignette experiment.

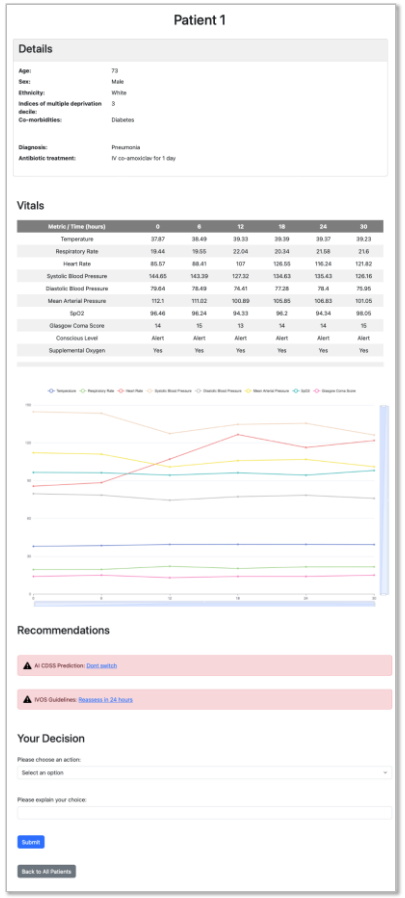
Patient	Shannon entropy
1	0.00
2	0.28
3	0.78
4	0.28
5	0.28
6	0.45
7	1.00
8	0.97
9	0.53
10	1.00
11	1.00
12	0.94

Table 9: Shannon entropy values for the switch decisions made by participants for each patient. Note due to two options (i.e., switch or dont switch) the minimum and maximum possible Shannon entropy values are 0 and 1 respectively.

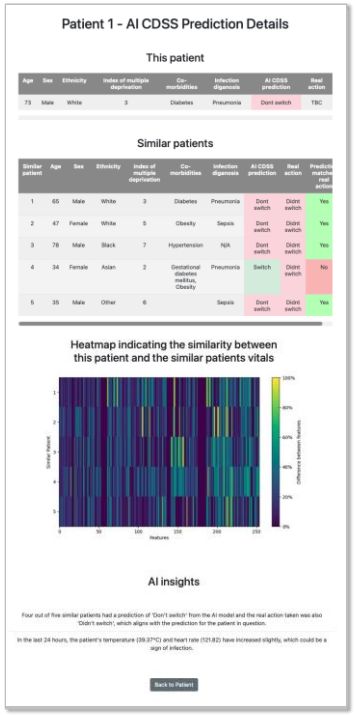
Screenshots of the web app used for the case vignette experiment.

A

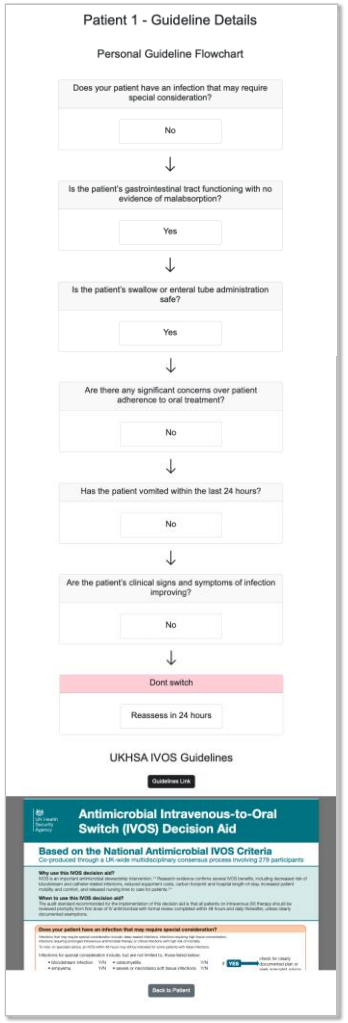
Case Vignette



AI CDSS Explanations

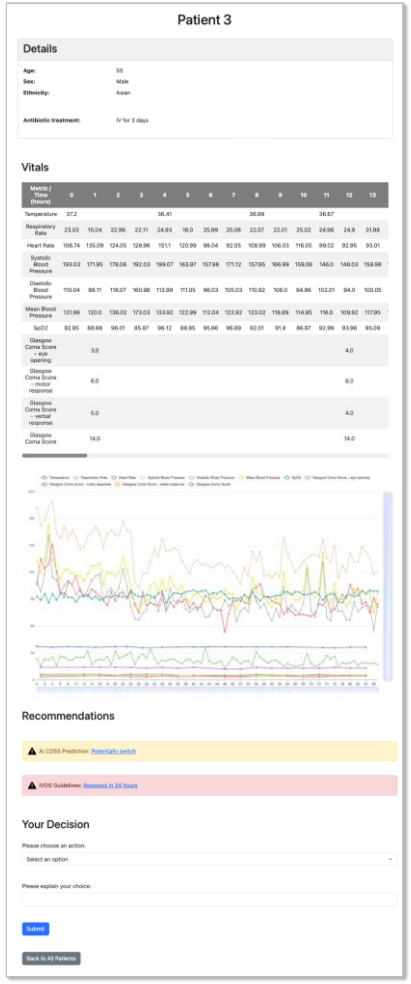


Guideline CDSS Explanations

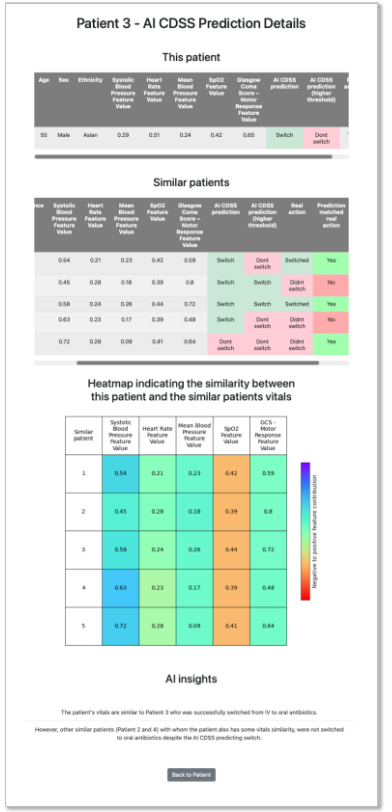


B

Case Vignette



AI CDSS Explanations



Guideline CDSS Explanations