IMPERIAL

Machine Learning for Neuroscience

ML4NS

Ethical Considerations and Responsible Machine Learning

Payam Barnaghi
Department of Brain Sciences &
School of Convergence Science in Human and Artificial Intelligence
Imperial College London

January 2025

Machine learning for healthcare and medicine

IMPERIAL

- The ultimate goal of developing machine learning models and systems in healthcare is deploying them in real-world settings and using them to improve patient and healthcare outcomes.

ML deployments in real-world settings

- The systems are often used for decision-support and have human-in-the-loop.
- However, the trustworthiness, reliability, and robustness of the systems/models must be considered and investigated prior to deployment.
- The users' perceptions of the system and appropriate training should also be considered.

Offline and online models

- Some models are designed to process offline and pre-collected datasets and can provide exploratory analysis and insights into the contributing factors to a disease, clusters and groups of features or patients based on different fractures.
- Online models are trained based on some existing data and deployed in operational settings. Sometimes, their parameters are fixed, and sometimes, they are either periodically re-trained or learned as new data emerges (i.e. online or continual learning).

Continual learning

- "Modern machine learning excels at training powerful models from fixed datasets and stationary environments, often exceeding human-level ability."
- "Yet, these models fail to emulate the process of human learning, which is efficient, robust, and able to learn incrementally, from sequential experience in a non-stationary world."

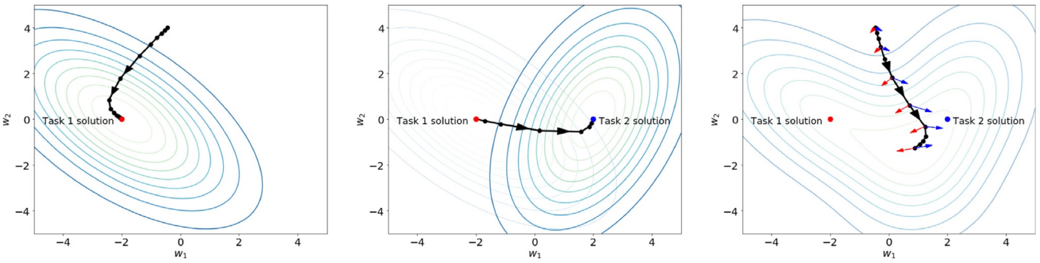
Source: R. Hadsell et al., Embracing Change: Continual Learning in Deep Neural Networks,
Trends in Cognitive Sciences, December 2020, Vol. 24, No. 12.

Why do we need continual learning?

- In continual learning, the model repeatedly receives new data, and the training data is not complete at a fixed given time.
- If we re-train the entire model whenever there are new instances, it would be inefficient, and we have to store the trained samples.
- The key challenge of continual learning scenarios in changing environments is how to incrementally and continually learn new tasks without forgetting the previous or creating highly complex models that may require accessing the entire training data.

Continual learning

IMPERIAL

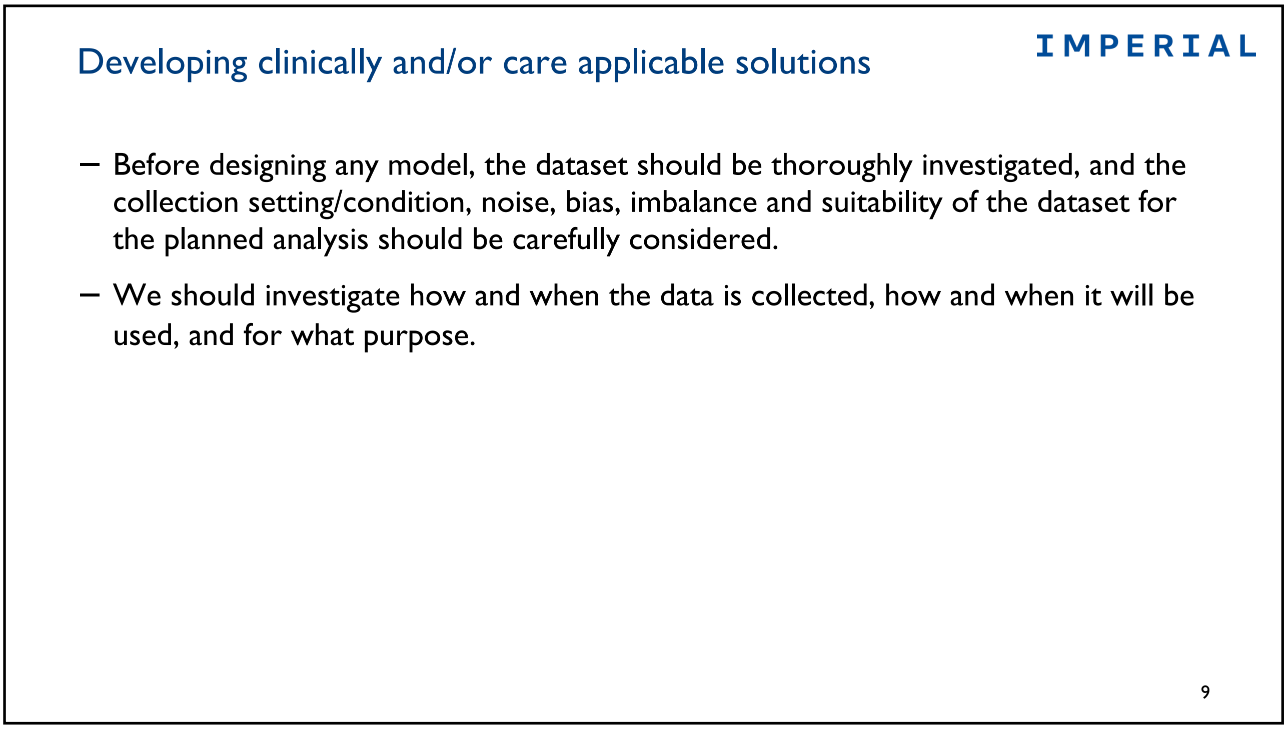


Source: R. Hadsell et al., Embracing Change: Continual Learning in Deep Neural Networks, Trends in Cognitive Sciences, December 2020, Vol. 24, No. 12.

Forgetting problem in machine learning models

IMPERIAL

- Most of the common deep learning models cannot adapt to different tasks without forgetting what they have learned in the past.
- Updating and altering tasks of an already learned model leads to the loss of previously learned knowledge as the network is not able to maintain the important weights for various distributions.
- The attempt to sequentially or continuously learn and adapt to various distributions will eventually result in a model collapse. This phenomenon is referred as catastrophic forgetting or interference (*McCloskey et al., 1989*) (*Goodfellow et al., 2013*).



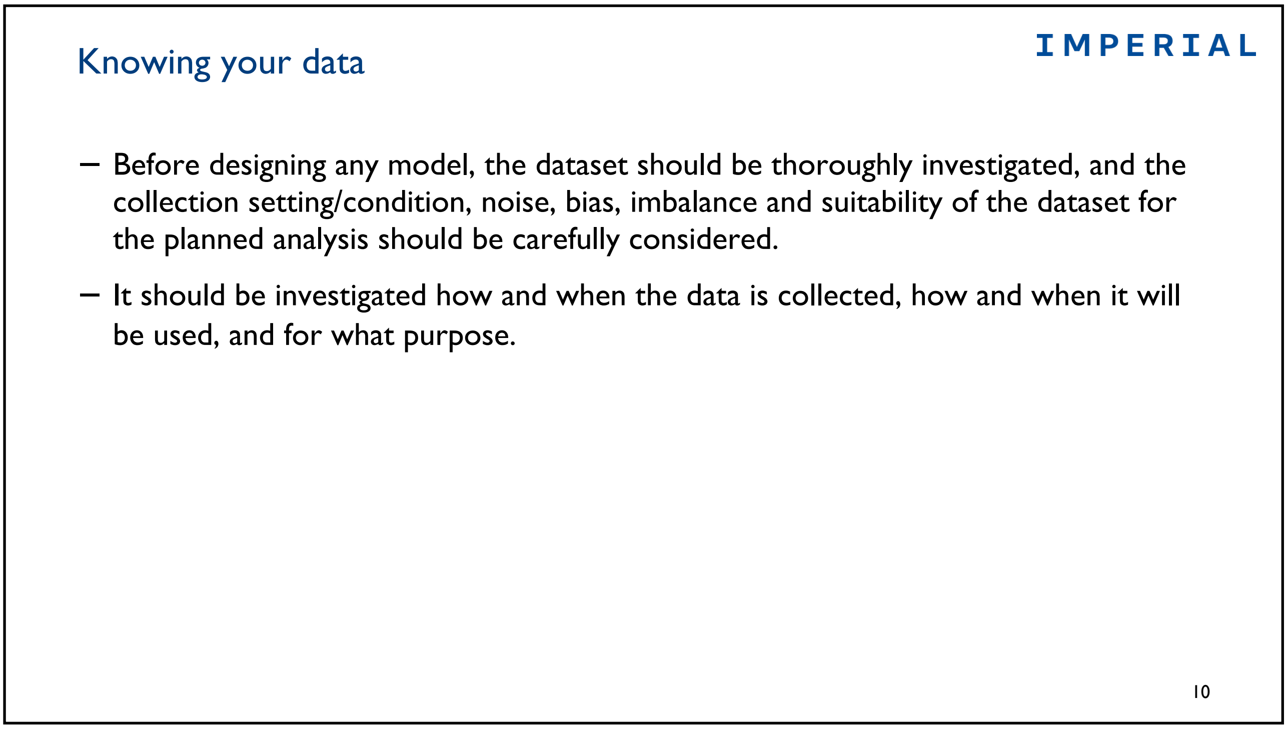
Developing clinically and/or care applicable solutions

IMPERIAL

- Before designing any model, the dataset should be thoroughly investigated, and the collection setting/condition, noise, bias, imbalance and suitability of the dataset for the planned analysis should be carefully considered.
- We should investigate how and when the data is collected, how and when it will be used, and for what purpose.

9

9



Knowing your data

IMPERIAL

- Before designing any model, the dataset should be thoroughly investigated, and the collection setting/condition, noise, bias, imbalance and suitability of the dataset for the planned analysis should be carefully considered.
- It should be investigated how and when the data is collected, how and when it will be used, and for what purpose.

10

10

Outcomes and end-points

IMPERIAL

- The clinical and care endpoints should be clearly defined.
- For example, if the model is going to be used for the prediction of an adverse health condition in a hospital setting, what timeframe would be clinically useful for the model to make the predictions? i.e., predicting a specific condition a few minutes before it happens may not be as useful in a real-world setting.
- Example:

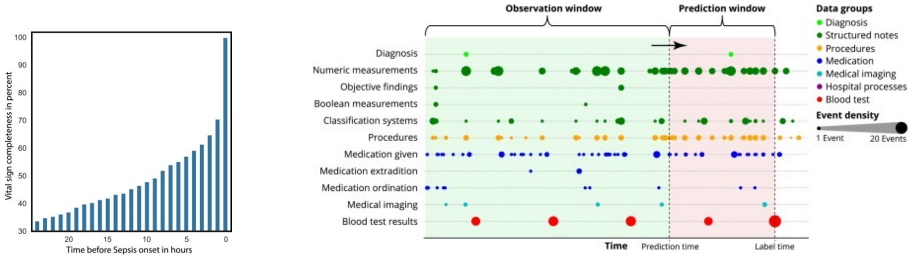
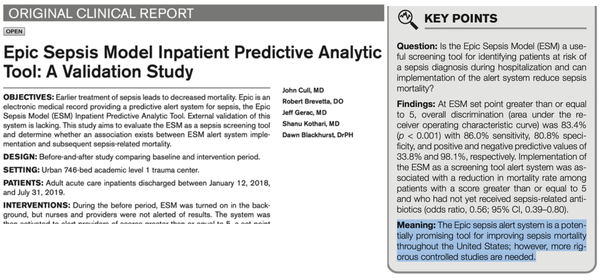


Image source: S. Meyer Lauritsen *et al.*, Early detection of sepsis utilizing deep learning on electronic health record event sequences, Artificial Intelligence in Medicine, 2020.

Robust evaluation of the models

IMPERIAL

- When a model is designed for an environment, validation within that environment requires careful thought to ensure that no unintended label leakage has occurred between the datasets used for model tuning and independent testing.
- Example:



ORIGINAL CLINICAL REPORT

OPEN

Epic Sepsis Model Inpatient Predictive Analytic Tool: A Validation Study

John Cull, MD
Robert Brevetta, DO
Jeff Gerac, MD
Shanu Kothari, MD
Dawn Blackhurst, DrPH

OBJECTIVES: Earlier treatment of sepsis leads to decreased mortality. Epic is an electronic medical record providing a predictive alert system for sepsis, the Epic Sepsis Model (ESM) Inpatient Predictive Analytic Tool. External validation of this system is lacking. This study aims to evaluate the ESM as a sepsis screening tool and determine whether an association exists between ESM alert system implementation and subsequent sepsis-related mortality.

DESIGN: Before-and-after study comparing baseline and intervention period.

SETTING: Urban 746-bed academic level 1 trauma center.

PATIENTS: Adult acute care inpatients discharged between January 12, 2018, and July 31, 2019.

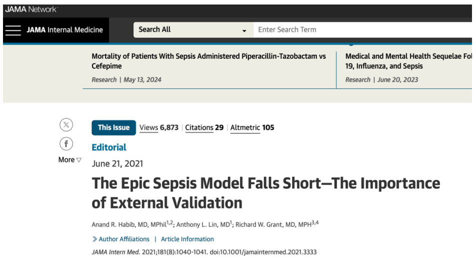
INTERVENTIONS: During the before period, ESM was turned on in the background, but nurses and providers were not alerted of results. The system was

KEY POINTS

Question: Is the Epic Sepsis Model (ESM) a useful screening tool for identifying patients at risk of a sepsis diagnosis during hospitalization and can implementation of the alert system reduce sepsis mortality?

Findings: At ESM set point greater than or equal to 5, overall discrimination (area under the receiver operating characteristic curve) was 83.4% ($p < 0.001$) with 86.0% sensitivity, 80.8% specificity, and positive and negative predictive values of 33.8% and 98.1%, respectively. Implementation of the ESM as a screening tool alert system was associated with a reduction in mortality rate among patients with a score greater than or equal to 5 and who had not yet received sepsis-related antibiotics (odds ratio, 0.56; 95% CI, 0.39–0.80).

Meaning: The Epic sepsis alert system is a potentially promising tool for improving sepsis mortality throughout the United States; however, more rigorous controlled studies are needed.



Mortality of Patients With Sepsis Administered Piperacillin-Tazobactam vs Cefepime
Research | May 13, 2024

Medical and Mental Health Sequelae Fol 19, Influenza, and Sepsis
Research | June 20, 2023

Views 6,873 | Citations 29 | Altmetric 105

Editorial

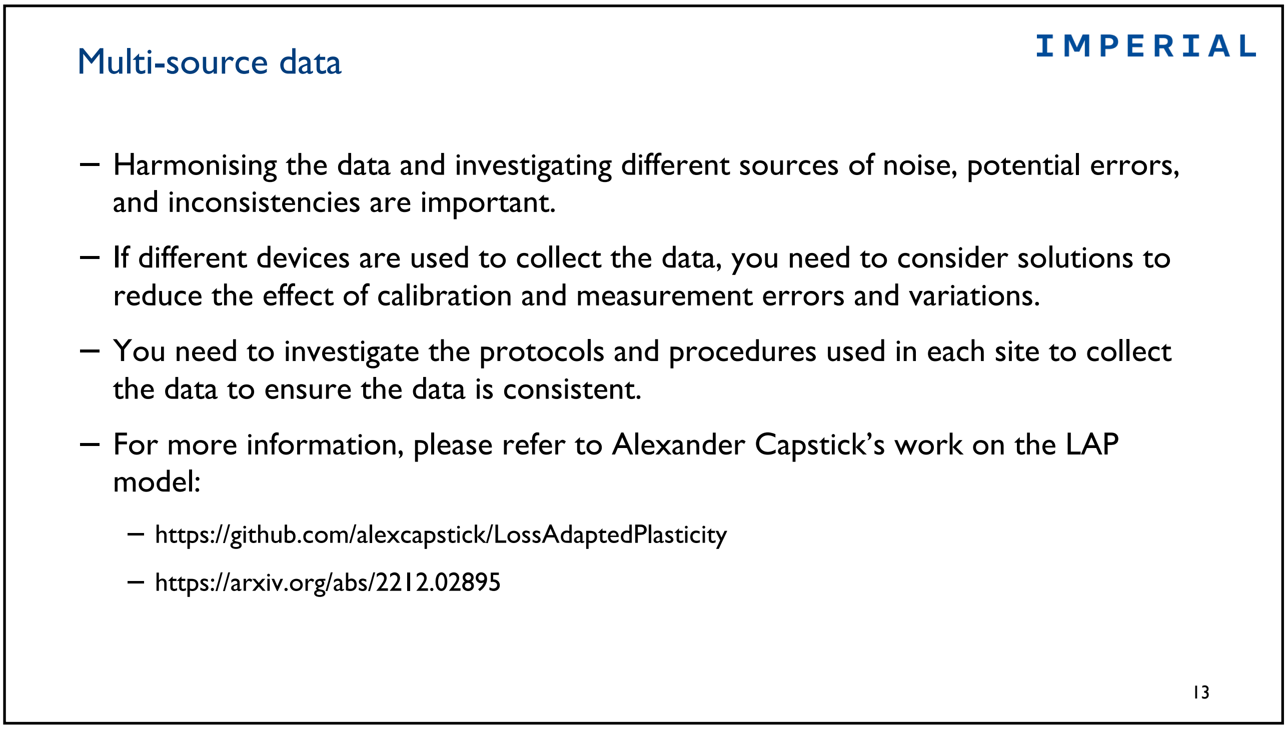
June 21, 2021

The Epic Sepsis Model Falls Short—The Importance of External Validation

Anand R. Habib, MD, MPhil[1,2]; Anthony L. Lin, MD[1]; Richard W. Grant, MD, MPH[3,4]

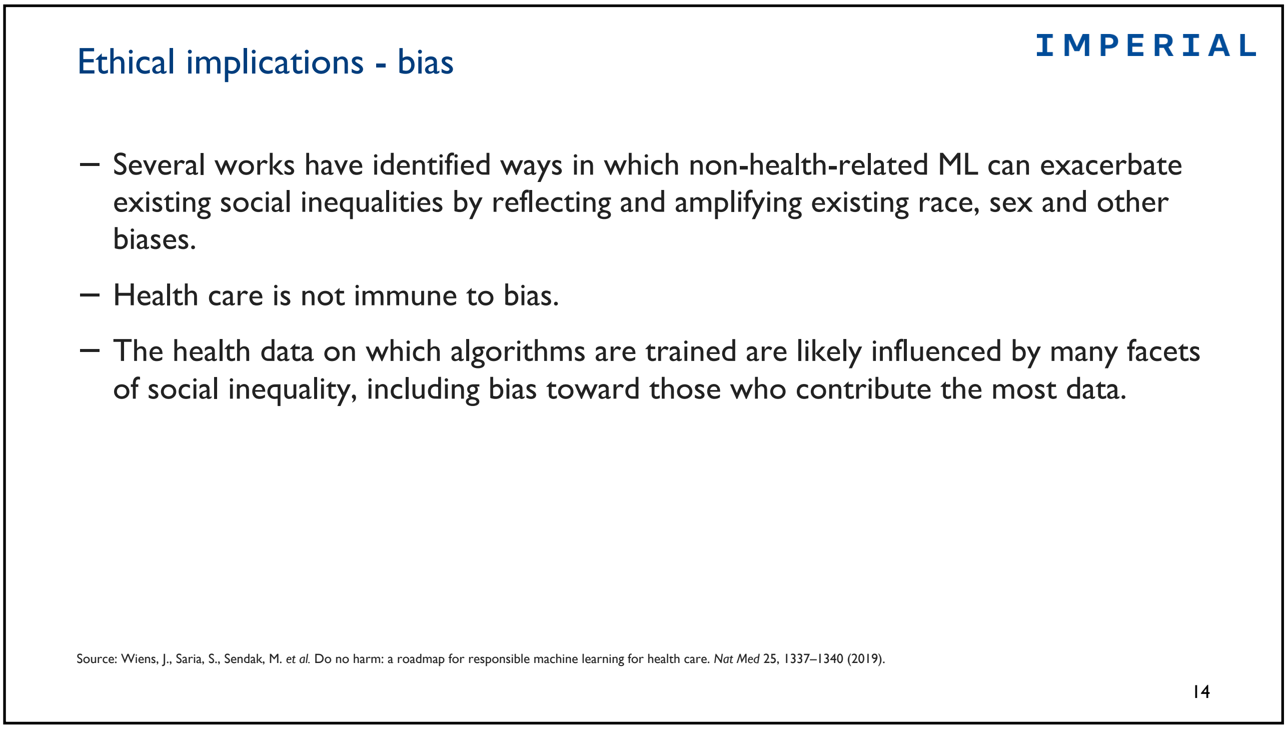
Author Affiliations | Article Information

JAMA Intern Med. 2021;181(8):1040-1041. doi:10.1001/jamainternmed.2021.3333



Multi-source data

- Harmonising the data and investigating different sources of noise, potential errors, and inconsistencies are important.
- If different devices are used to collect the data, you need to consider solutions to reduce the effect of calibration and measurement errors and variations.
- You need to investigate the protocols and procedures used in each site to collect the data to ensure the data is consistent.
- For more information, please refer to Alexander Capstick's work on the LAP model:
  - https://github.com/alexcapstick/LossAdaptedPlasticity
  - https://arxiv.org/abs/2212.02895



Ethical implications - bias

- Several works have identified ways in which non-health-related ML can exacerbate existing social inequalities by reflecting and amplifying existing race, sex and other biases.
- Health care is not immune to bias.
- The health data on which algorithms are trained are likely influenced by many facets of social inequality, including bias toward those who contribute the most data.

Source: 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).

IMPERIAL

Algorithmic bias

- Algorithms that predict an individual's risk for a condition or suitability of a specific treatment may be biased toward those who are able to access and afford the procedure.
- This could happen by feeding data from the people who have had that treatment in the past, which won't include people who couldn't afford it in the first place or didn't have access to it.
- Some of this bias can be corrected during model training when the data is divided into training, validation and test sets.
- In general, awareness is necessary to investigate when potential biases could be present in the data and what can be done to mitigate their effect.

Source: 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).

15

15

IMPERIAL

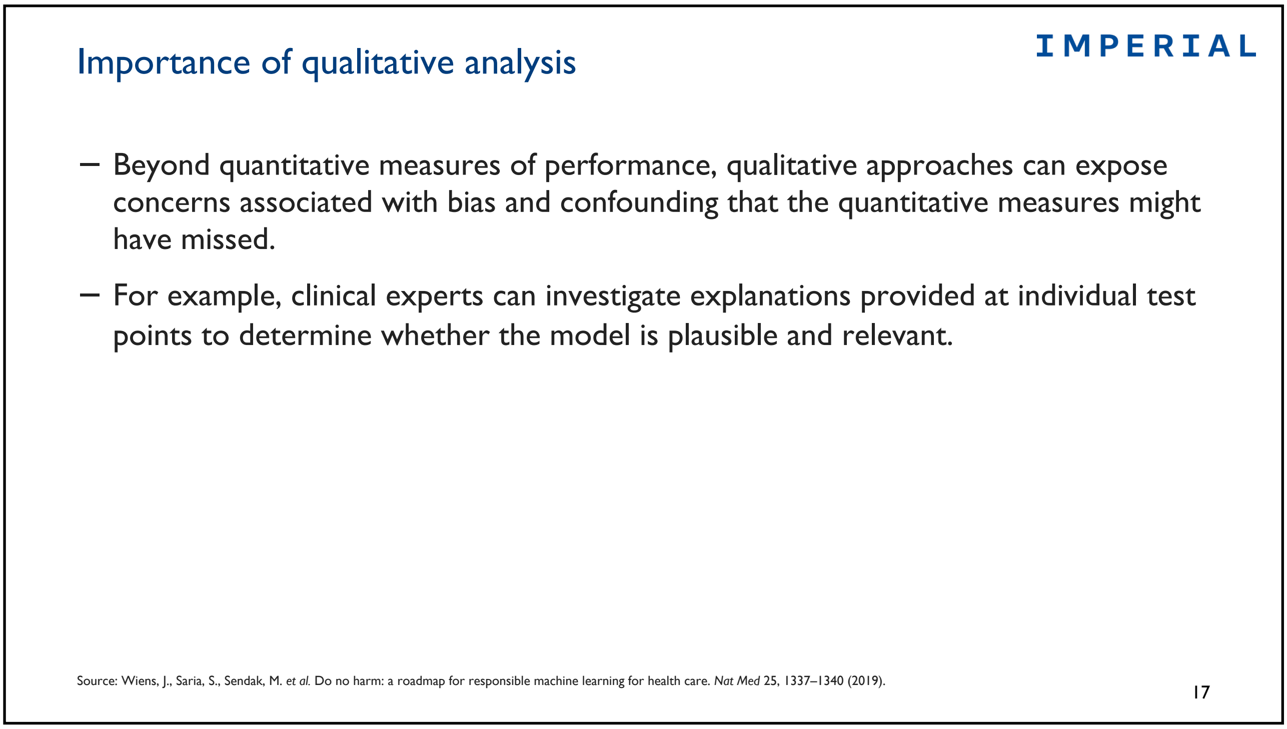
Robust evaluation of the models

- When a model is designed for an environment, validation within that environment requires careful thought to ensure that no unintended label leakage has occurred between the datasets used for model tuning and independent testing.
- For example, the 'radiologist-level' performance recently achieved across several tasks using chest X-rays. The data used in the analysis consisted of multiple X-ray images per patient. It was important to split data at the patient level instead of random splitting so that no images from the same patient appeared in both the training and testing sets.

Source: 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).

16

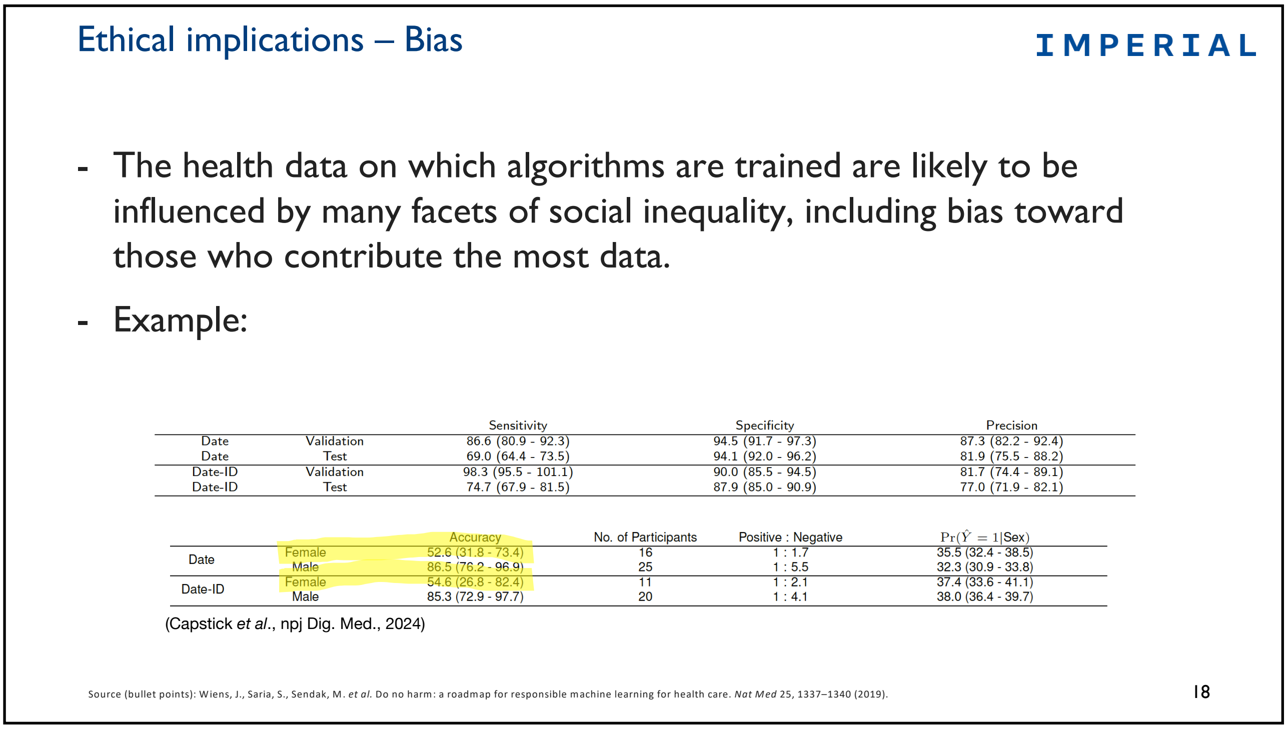
16



Importance of qualitative analysis

- Beyond quantitative measures of performance, qualitative approaches can expose concerns associated with bias and confounding that the quantitative measures might have missed.
- For example, clinical experts can investigate explanations provided at individual test points to determine whether the model is plausible and relevant.

Source: 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).



Ethical implications – Bias

- The health data on which algorithms are trained are likely to be influenced by many facets of social inequality, including bias toward those who contribute the most data.
- Example:

		Sensitivity	Specificity	Precision
Date	Validation	86.6 (80.9 - 92.3)	94.5 (91.7 - 97.3)	87.3 (82.2 - 92.4)
Date	Test	69.0 (64.4 - 73.5)	94.1 (92.0 - 96.2)	81.9 (75.5 - 88.2)
Date-ID	Validation	98.3 (95.5 - 101.1)	90.0 (85.5 - 94.5)	81.7 (74.4 - 89.1)
Date-ID	Test	74.7 (67.9 - 81.5)	87.9 (85.0 - 90.9)	77.0 (71.9 - 82.1)

		Accuracy	No. of Participants	Positive : Negative	$\Pr(\hat{Y} = 1 \mid \text{Sex})$
Date	Female	52.6 (31.8 - 73.4)	16	1 : 1.7	35.5 (32.4 - 38.5)
	Male	86.5 (76.2 - 96.9)	25	1 : 5.5	32.3 (30.9 - 33.8)
Date-ID	Female	54.6 (26.8 - 82.4)	11	1 : 2.1	37.4 (33.6 - 41.1)
	Male	85.3 (72.9 - 97.7)	20	1 : 4.1	38.0 (36.4 - 39.7)

(Capstick *et al*., npj Dig. Med., 2024)

Source (bullet points): 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).

Reporting and context of the application

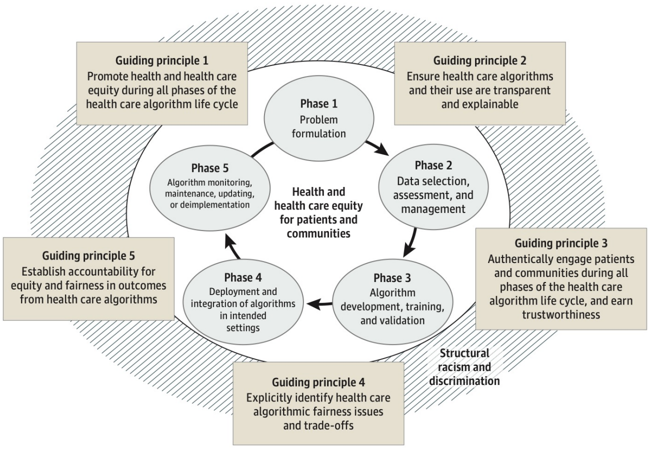
IMPERIAL

- Proposed reporting guidelines developed by the community provide good ideas to outline the importance of clear descriptions of the source of the data, participants, outcomes and predictors, and in some cases, require the model itself (e.g., regression coefficients) to be presented.
- This last requirement creates a potential for unintended consequences and even harm if the model is then applied inappropriately.
- For example, a recent study in building models to predict healthcare-associated infections found that variables associated with risk at one hospital were protective in another.
- So, it is essential to report the context(s) in which the model applies and was validated.

Source: 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).

A conceptual framework

IMPERIAL



Guiding principle 1
Promote health and health care equity during all phases of the health care algorithm life cycle

Guiding principle 2
Ensure health care algorithms and their use are transparent and explainable

Phase 1
Problem formulation

Phase 2
Data selection, assessment, and management

Phase 3
Algorithm development, training, and validation

Phase 4
Deployment and integration of algorithms in intended settings

Phase 5
Algorithm monitoring, maintenance, updating, or deimplementation

Health and health care equity for patients and communities

Guiding principle 3
Authentically engage patients and communities during all phases of the health care algorithm life cycle, and earn trustworthiness

Guiding principle 5
Establish accountability for equity and fairness in outcomes from health care algorithms

Guiding principle 4
Explicitly identify health care algorithmic fairness issues and trade-offs

Structural racism and discrimination

Guiding principles apply at each phase to mitigate and prevent bias in an algorithm.
Operationalization of the principles takes place at 3 levels: individual, institutional, and societal.

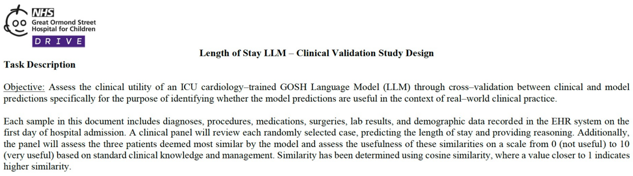
This conceptual framework builds on a US National Academy of Medicine algorithm life cycle framework adapted by Roski et al.

Source: Marshall H. Chin, *et al.*, JAMA Network.

Importance of qualitative analysis

IMPERIAL

- Beyond quantitative measures of performance, qualitative approaches can expose concerns associated with bias and confounding that the quantitative measures might have missed.
- For example, clinical experts can investigate explanations provided at individual test points to determine whether the model is plausible and relevant.
- Example:



NHS Great Ormond Street Hospital for Children DRIVE

Length of Stay LLM – Clinical Validation Study Design

Task Description

Objective: Assess the clinical utility of an ICU cardiology–trained GOSH Language Model (LLM) through cross–validation between clinical and model predictions specifically for the purpose of identifying whether the model predictions are useful in the context of real–world clinical practice.

Each sample in this document includes diagnoses, procedures, medications, surgeries, lab results, and demographic data recorded in the EHR system on the first day of hospital admission. A clinical panel will review each randomly selected case, predicting the length of stay and providing reasoning. Additionally, the panel will assess the three patients deemed most similar by the model and assess the usefulness of these similarities on a scale from 0 (not useful) to 10 (very useful) based on standard clinical knowledge and management. Similarity has been determined using cosine similarity, where a value closer to 1 indicates higher similarity.

Source (bullet points) : 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).

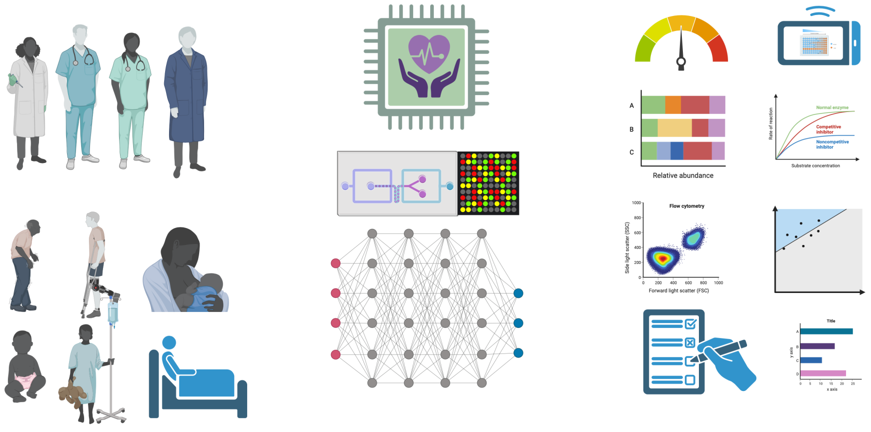
User perception and training requirements

IMPERIAL

User groups and stakeholders

Machine learning solutions

Presentation, engagement feedback, and maintenance



Created with BioRender.com

Deployment and maintenance responsibility

IMPERIAL

- Effectively applying a predictive model in an ethical, legal and morally responsible manner within a real-world healthcare setting can be substantially more difficult than developing a model in a curated experimental environment.
- Before integrating an AI/ML model into patient care, it is critical to test the system in ‘silent’ mode, in which predictions are made in real-time and exposed to a group of clinical experts but not acted upon.
- This prospective validation allows clinicians to identify and review errors in real-world settings.

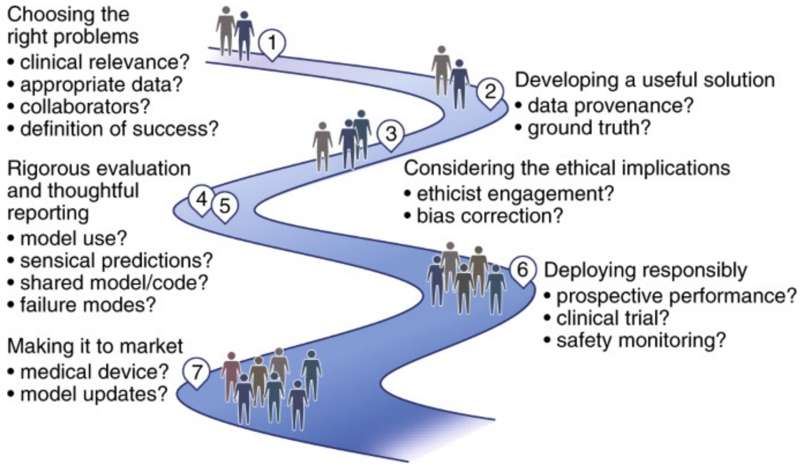
Source: Wiens, J., Saria, S., Sendak, M. *et al.* Do no harm: a roadmap for responsible machine learning for health care. *Nat Med* 25, 1337–1340 (2019).

23

23

A roadmap for deploying effective ML systems

IMPERIAL



Choosing the right problems
- clinical relevance?
- appropriate data?
- collaborators?
- definition of success?

Developing a useful solution
- data provenance?
- ground truth?

Considering the ethical implications
- ethicist engagement?
- bias correction?

Rigorous evaluation and thoughtful reporting
- model use?
- sensical predictions?
- shared model/code?
- failure modes?

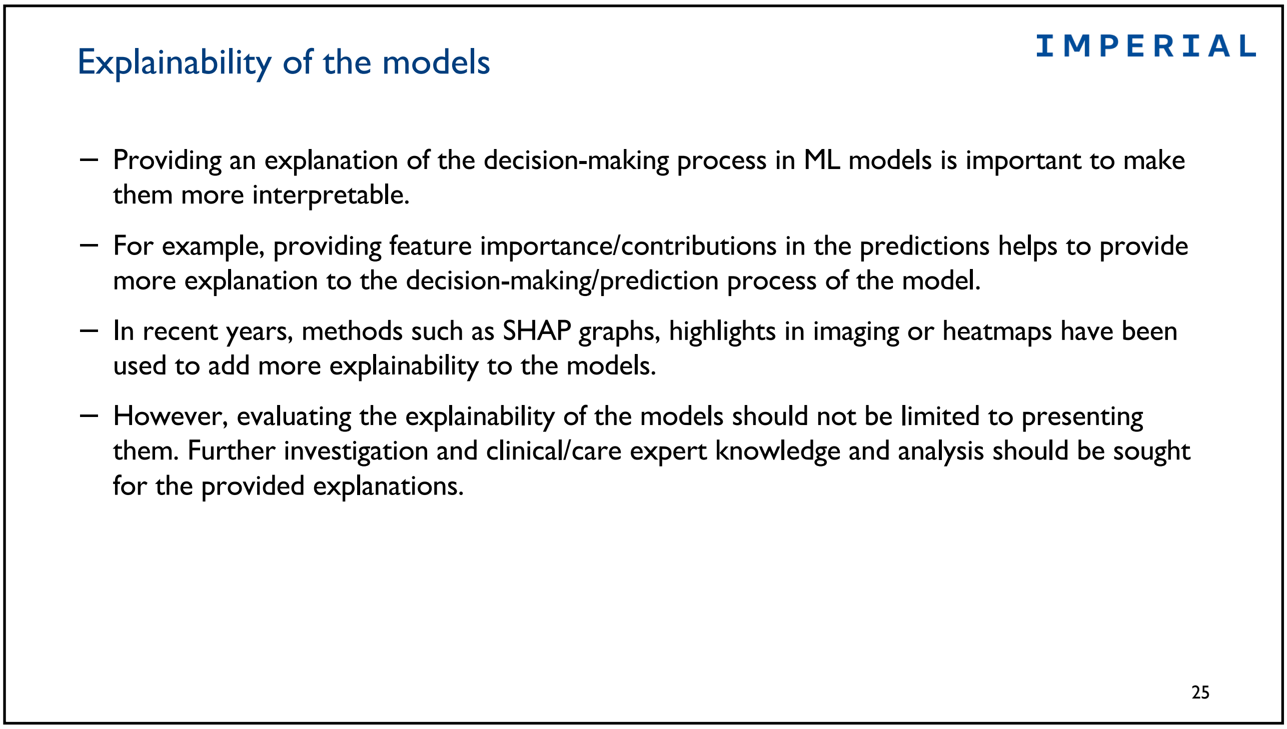
Deploying responsibly
- prospective performance?
- clinical trial?
- safety monitoring?

Making it to market
- medical device?
- model updates?

Source: 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).

24

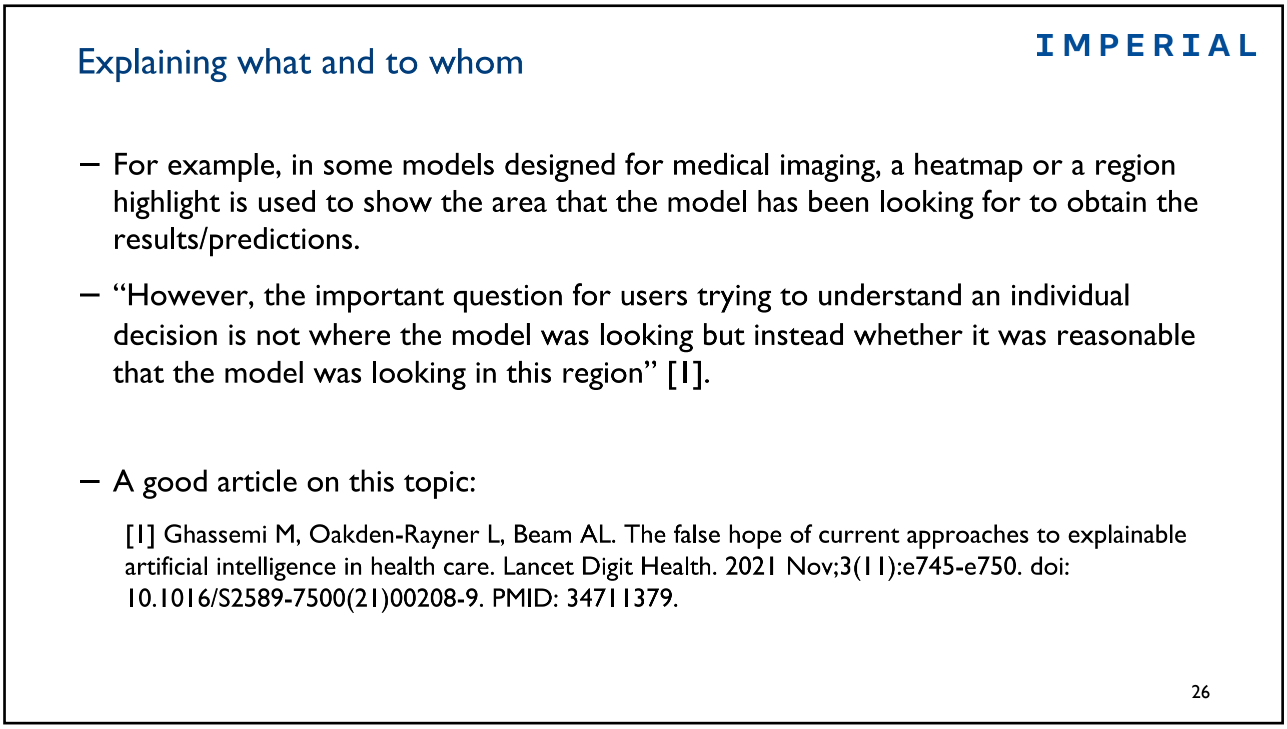
24



Explainability of the models

IMPERIAL

- Providing an explanation of the decision-making process in ML models is important to make them more interpretable.
- For example, providing feature importance/contributions in the predictions helps to provide more explanation to the decision-making/prediction process of the model.
- In recent years, methods such as SHAP graphs, highlights in imaging or heatmaps have been used to add more explainability to the models.
- However, evaluating the explainability of the models should not be limited to presenting them. Further investigation and clinical/care expert knowledge and analysis should be sought for the provided explanations.



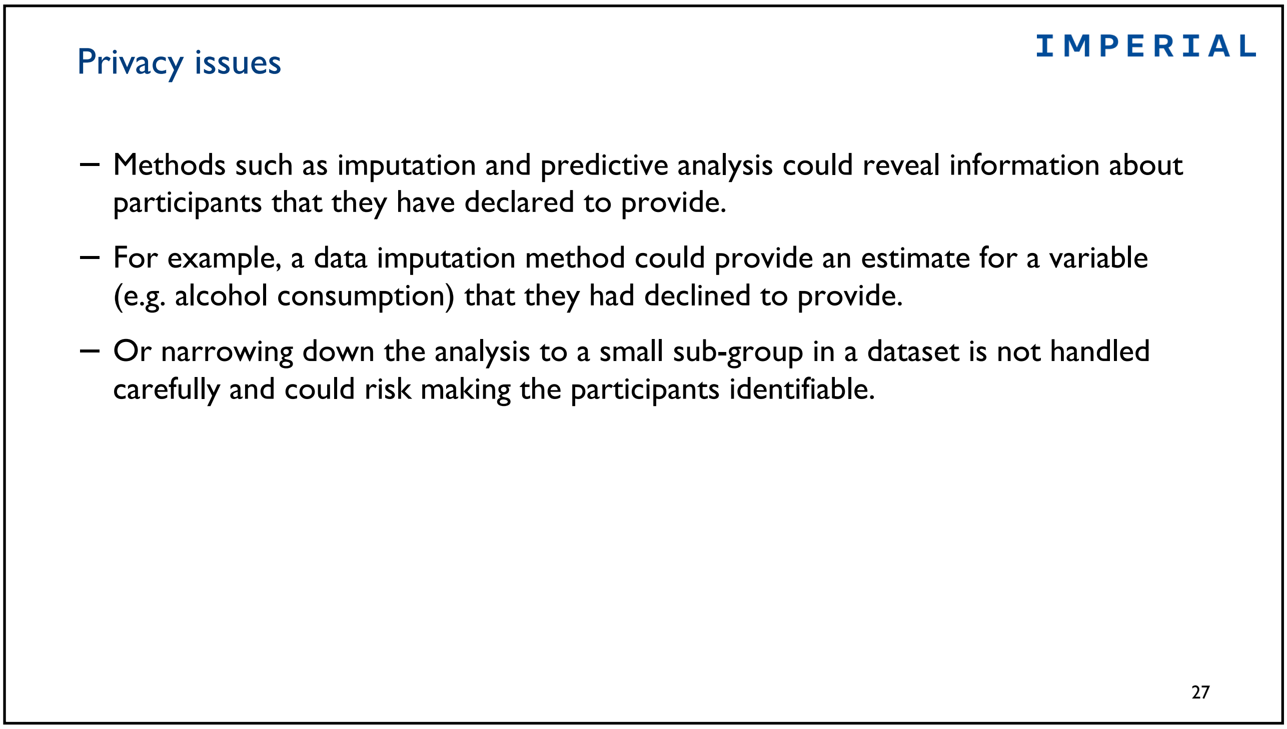
Explaining what and to whom

IMPERIAL

- For example, in some models designed for medical imaging, a heatmap or a region highlight is used to show the area that the model has been looking for to obtain the results/predictions.
- “However, the important question for users trying to understand an individual decision is not where the model was looking but instead whether it was reasonable that the model was looking in this region” [1].

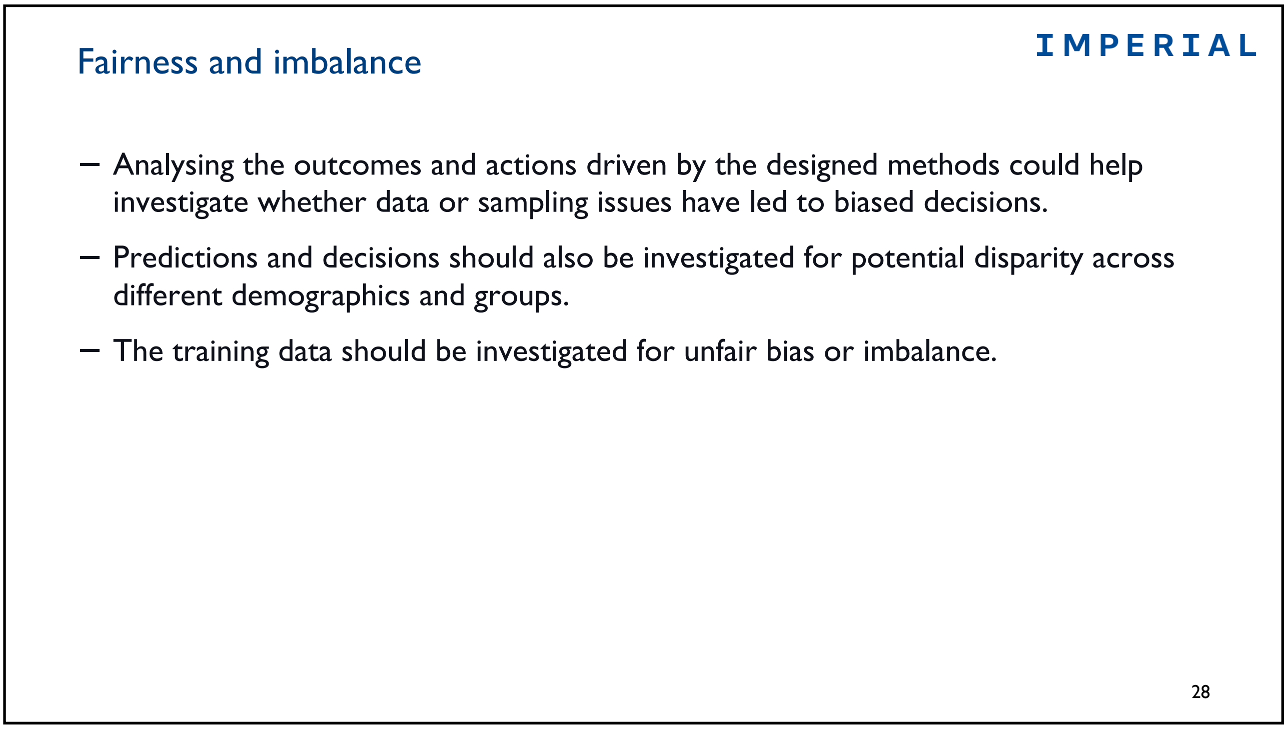
- A good article on this topic:

  [1] Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021 Nov;3(11):e745-e750. doi: 10.1016/S2589-7500(21)00208-9. PMID: 34711379.



Privacy issues

- Methods such as imputation and predictive analysis could reveal information about participants that they have declared to provide.
- For example, a data imputation method could provide an estimate for a variable (e.g. alcohol consumption) that they had declined to provide.
- Or narrowing down the analysis to a small sub-group in a dataset is not handled carefully and could risk making the participants identifiable.



Fairness and imbalance

- Analysing the outcomes and actions driven by the designed methods could help investigate whether data or sampling issues have led to biased decisions.
- Predictions and decisions should also be investigated for potential disparity across different demographics and groups.
- The training data should be investigated for unfair bias or imbalance.

Evaluation metrics

- Appropriate evaluation metrics and suitable/applicable targets for the intended outcomes should be considered.
- For example, what time window would be suitable for the input data to make a prediction, and what are the acceptable sensitivity and specificity thresholds for the intended applications?

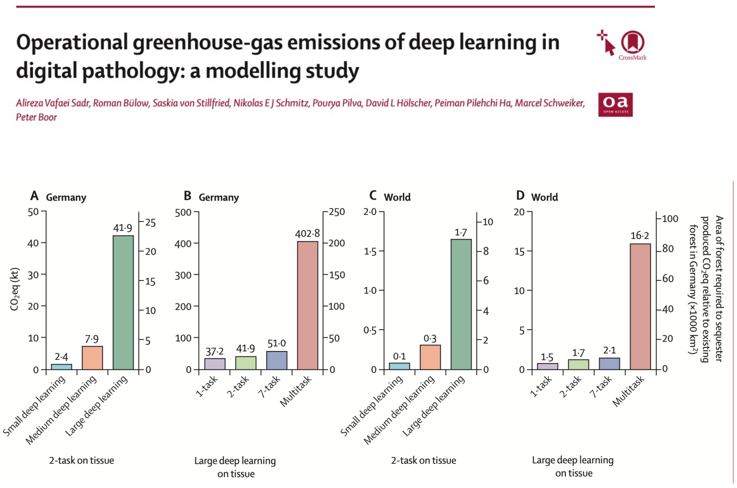
- In some applications, specificity and sensitivity or precision/recall may not carry the same weight, so the methods that look for an optimum balance between the two measures will not be suitable for this type of application.
- For example, a screening test may require higher recall, but lower precision could be tolerated.

Prediction and analysis errors

- In real-world deployments, you may need to identify suitable channels and procedures to investigate/report the prediction errors.
- Identifying procedures and steps to re-train/adjust the model to avoid potentially harmful errors.
- However, this should be done carefully and not add further inconsistencies or potential bias/errors to the model.

Sustainability and Environmental Impact

IMPERIAL



Operational greenhouse-gas emissions of deep learning in digital pathology: a modelling study

Alireza Vafaei Sadr, Roman Bülow, Saskia von Stillfried, Nikolas E J Schmitz, Pourya Pilva, David L Hölscher, Peiman Pilehchi Ha, Marcel Schweiker, Peter Boor

Vafaei Sadr, Alireza *et al.*, Operational greenhouse-gas emissions of deep learning in digital pathology: a modelling study, *The Lancet Digital Health*, Dec 2023.

31

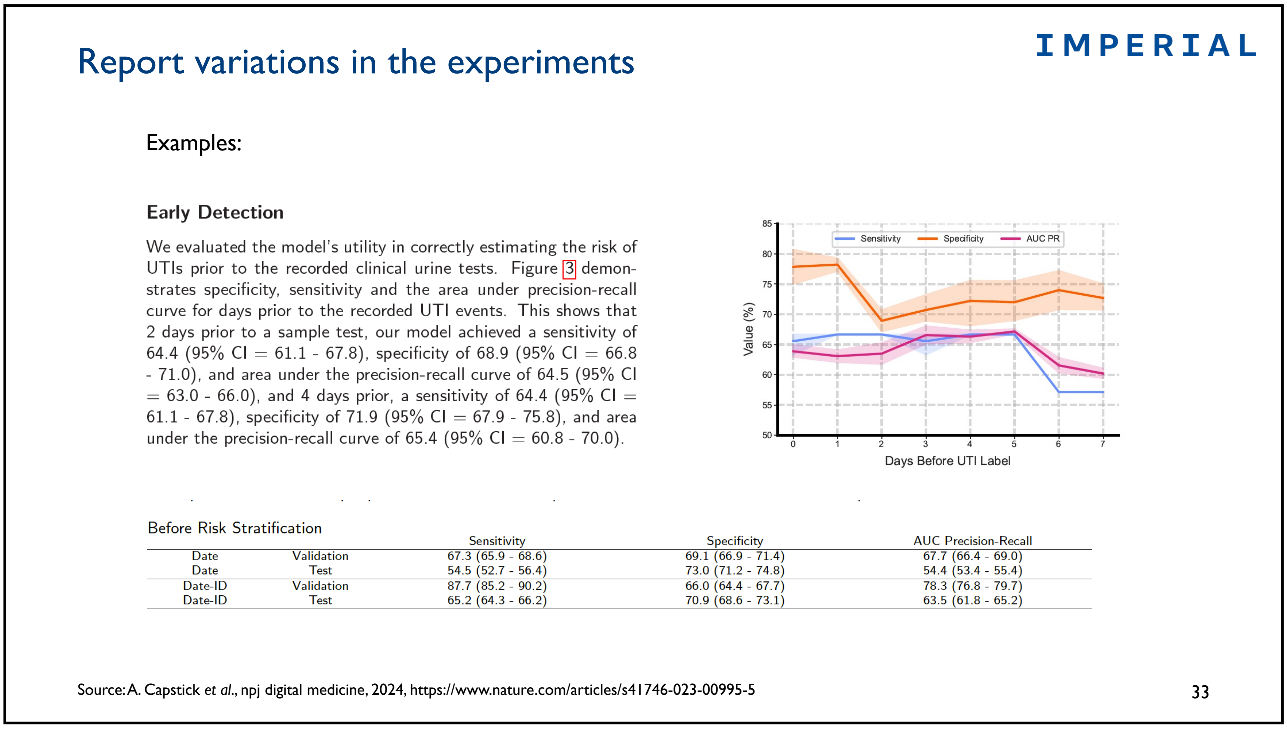
31

Reporting the results

IMPERIAL

- In the following slides, some of the good practices for reporting the results of machine learning models are discussed.
- I have used examples from our recent paper on UTI risk analysis in people living with dementia using remote monitoring data (A. Capstick *et al.*, npj digital medicine, 2024, https://www.nature.com/articles/s41746-023-00995-5).
- The evaluations are based on ID and ID-Date splits
  - In the ID split, the test data includes data from patients that the model has never seen before.
  - In the ID-Date split, the test data include data from a period of time and from patients that the model has not seen before.

32



Report variations in the experiments

IMPERIAL

Examples:

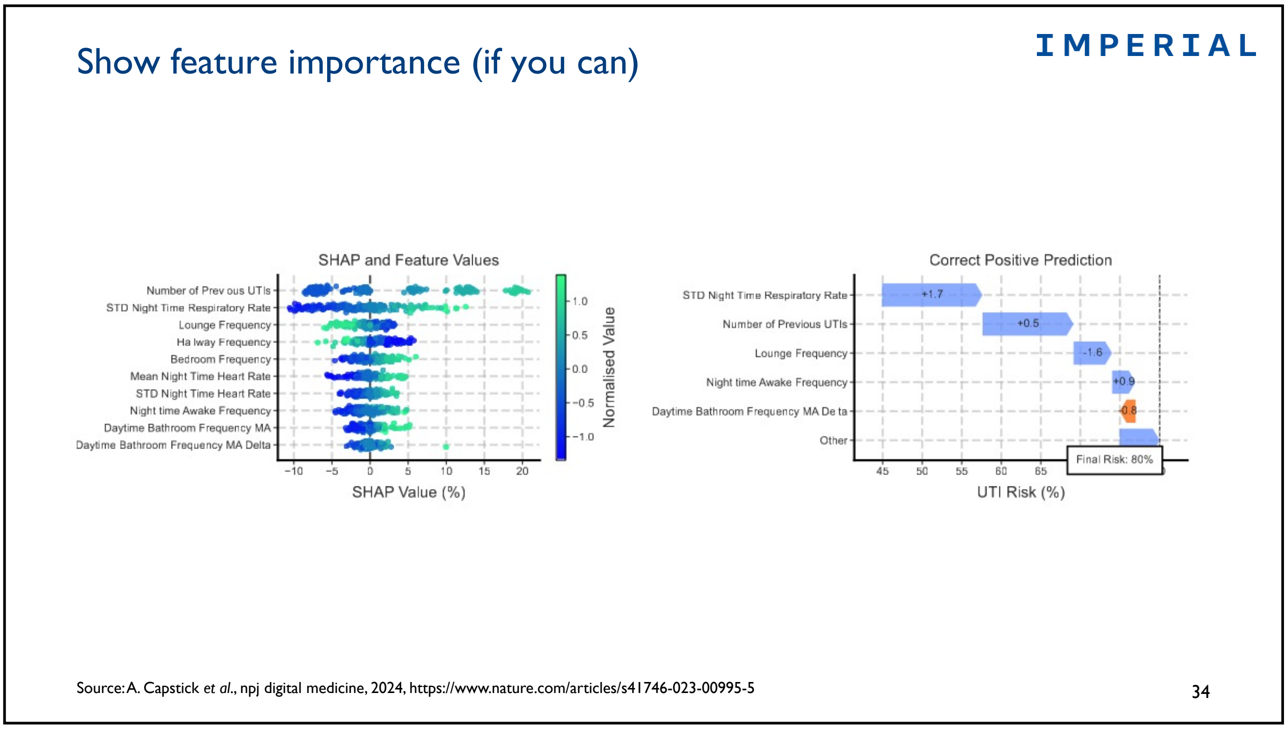
Early Detection

We evaluated the model's utility in correctly estimating the risk of UTIs prior to the recorded clinical urine tests. Figure 3 demonstrates specificity, sensitivity and the area under precision-recall curve for days prior to the recorded UTI events. This shows that 2 days prior to a sample test, our model achieved a sensitivity of 64.4 (95% CI = 61.1 - 67.8), specificity of 68.9 (95% CI = 66.8 - 71.0), and area under the precision-recall curve of 64.5 (95% CI = 63.0 - 66.0), and 4 days prior, a sensitivity of 64.4 (95% CI = 61.1 - 67.8), specificity of 71.9 (95% CI = 67.9 - 75.8), and area under the precision-recall curve of 65.4 (95% CI = 60.8 - 70.0).

Before Risk Stratification

		Sensitivity	Specificity	AUC Precision-Recall
Date	Validation	67.3 (65.9 - 68.6)	69.1 (66.9 - 71.4)	67.7 (66.4 - 69.0)
Date	Test	54.5 (52.7 - 56.4)	73.0 (71.2 - 74.8)	54.4 (53.4 - 55.4)
Date-ID	Validation	87.7 (85.2 - 90.2)	66.0 (64.4 - 67.7)	78.3 (76.8 - 79.7)
Date-ID	Test	65.2 (64.3 - 66.2)	70.9 (68.6 - 73.1)	63.5 (61.8 - 65.2)

Source: A. Capstick *et al*., npj digital medicine, 2024, https://www.nature.com/articles/s41746-023-00995-5



Show feature importance (if you can)

IMPERIAL

SHAP and Feature Values
Number of Prev ous UTIs
STD Night Time Respiratory Rate
Lounge Frequency
Ha lway Frequency
Bedroom Frequency
Mean Night Time Heart Rate
STD Night Time Heart Rate
Night time Awake Frequency
Daytime Bathroom Frequency MA
Daytime Bathroom Frequency MA Delta
−10
−5
0
5
10
15
20
SHAP Value (%)
1.0
0.5
0.0
−0.5
−1.0
Normalised Value
Correct Positive Prediction
STD Night Time Respiratory Rate
Number of Previous UTIs
Lounge Frequency
Night time Awake Frequency
Daytime Bathroom Frequency MA De ta
Other
+1.7
+0.5
-1.6
+0.9
-0.8
Final Risk: 80%
45
50
55
60
65
UTI Risk (%)

Source: A. Capstick *et al*., npj digital medicine, 2024, https://www.nature.com/articles/s41746-023-00995-5

Report the demographics and other data characteristics

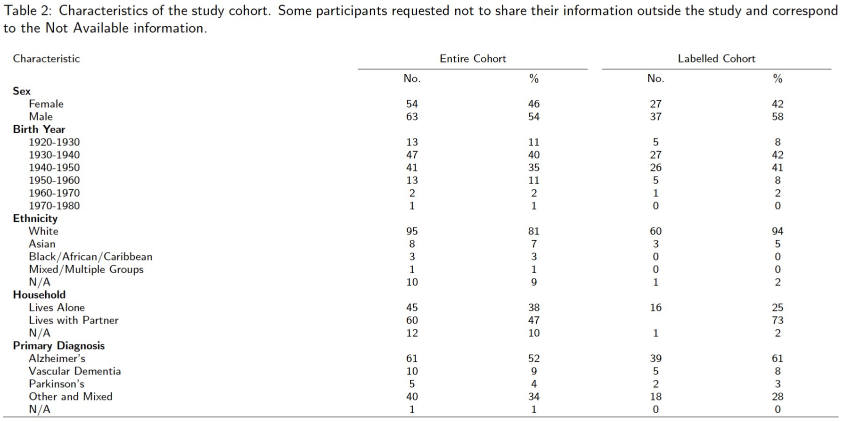
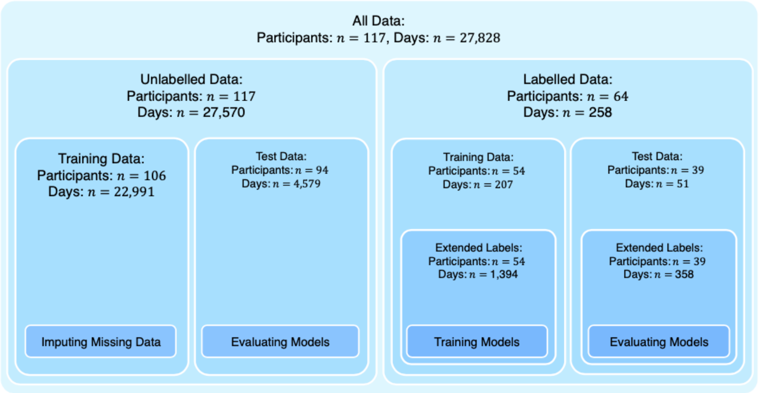


Table 2: Characteristics of the study cohort. Some participants requested not to share their information outside the study and correspond to the Not Available information.

Characteristic	Entire Cohort		Labelled Cohort	
	No.	%	No.	%
Sex				
Female	54	46	27	42
Male	63	54	37	58
Birth Year				
1920-1930	13	11	5	8
1930-1940	47	40	27	42
1940-1950	41	35	26	41
1950-1960	13	11	5	8
1960-1970	2	2	1	2
1970-1980	1	1	0	0
Ethnicity				
White	95	81	60	94
Asian	8	7	3	5
Black/African/Caribbean	3	3	0	0
Mixed/Multiple Groups	1	1	0	0
N/A	10	9	1	2
Household				
Lives Alone	45	38	16	25
Lives with Partner	60	47		73
N/A	12	10	1	2
Primary Diagnosis				
Alzheimer's	61	52	39	61
Vascular Dementia	10	9	5	8
Parkinson's	5	4	2	3
Other and Mixed	40	34	18	28
N/A	1	1	0	0

Source: A. Capstick *et al.*, npj digital medicine, 2024, https://www.nature.com/articles/s41746-023-00995-5

Report training/test splits



All Data:
Participants: $n = 117$, Days: $n = 27{,}828$

Unlabelled Data:
Participants: $n = 117$
Days: $n = 27{,}570$

Training Data:
Participants: $n = 106$
Days: $n = 22{,}991$

Imputing Missing Data

Test Data:
Participants: $n = 94$
Days: $n = 4{,}579$

Evaluating Models

Labelled Data:
Participants: $n = 64$
Days: $n = 258$

Training Data:
Participants: $n = 54$
Days: $n = 207$

Extended Labels:
Participants: $n = 54$
Days: $n = 1{,}394$

Training Models

Test Data:
Participants: $n = 39$
Days: $n = 51$

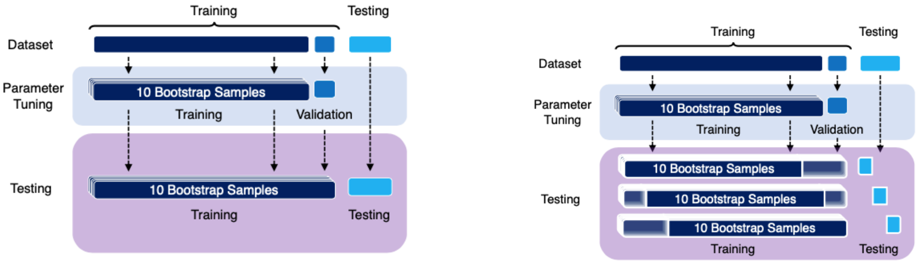
Extended Labels:
Participants: $n = 39$
Days: $n = 358$

Evaluating Models

Source: A. Capstick *et al.*, npj digital medicine, 2024, https://www.nature.com/articles/s41746-023-00995-5

Report validation and repeats

IMPERIAL



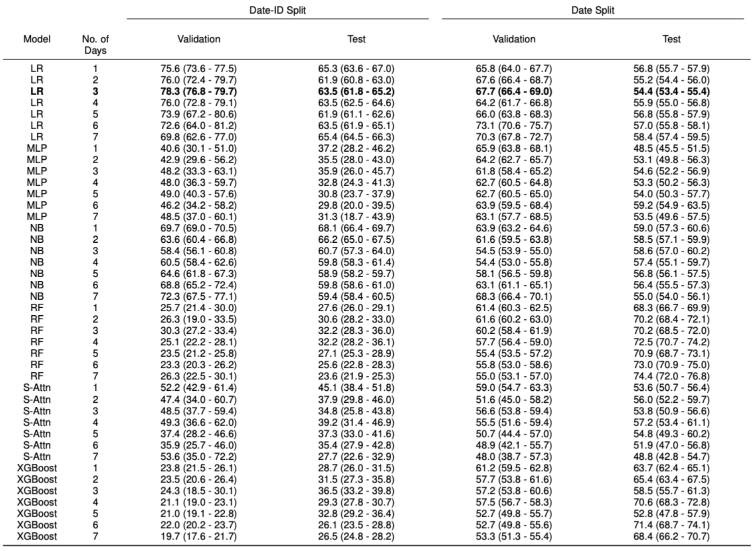
(a) Date Split

(b) Date-ID Split

Source: A. Capstick *et al.*, npj digital medicine, 2024, https://www.nature.com/articles/s41746-023-00995-5

Report comparisons against baseline models

IMPERIAL

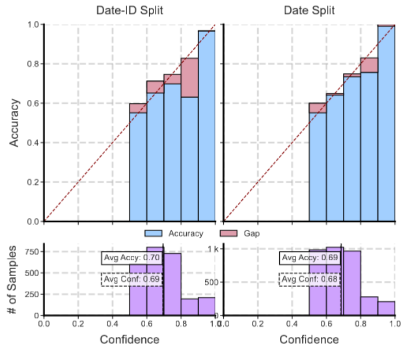


Model	No. of Days	Date-ID Split		Date Split	
		Validation	Test	Validation	Test
LR	1	75.6 (73.6 - 77.5)	65.3 (63.6 - 67.0)	65.8 (64.0 - 67.7)	56.8 (55.7 - 57.9)
LR	2	76.0 (72.4 - 79.7)	61.9 (60.8 - 63.0)	67.6 (66.4 - 68.7)	55.2 (54.4 - 56.0)
LR	**3**	**78.3 (76.8 - 79.7)**	**63.5 (61.8 - 65.2)**	**67.7 (66.4 - 69.0)**	**54.4 (53.4 - 55.4)**
LR	4	76.0 (72.8 - 79.1)	63.5 (62.5 - 64.6)	64.2 (61.7 - 66.8)	55.9 (55.0 - 56.8)
LR	5	73.9 (67.2 - 80.6)	61.9 (61.1 - 62.6)	66.0 (63.8 - 68.3)	56.8 (55.8 - 57.9)
LR	6	72.6 (64.0 - 81.2)	63.5 (61.9 - 65.1)	73.1 (70.6 - 75.7)	57.0 (55.8 - 58.1)
LR	7	69.8 (62.6 - 77.0)	65.4 (64.5 - 66.3)	70.3 (67.8 - 72.7)	58.4 (57.4 - 59.5)
MLP	1	40.6 (30.1 - 51.0)	37.2 (28.2 - 46.2)	65.9 (63.8 - 68.1)	48.5 (45.5 - 51.5)
MLP	2	42.9 (29.6 - 56.2)	35.5 (28.0 - 43.0)	64.2 (62.7 - 65.7)	53.1 (49.8 - 56.3)
MLP	3	48.2 (33.3 - 63.1)	35.9 (26.0 - 45.7)	61.8 (58.4 - 65.2)	54.6 (52.2 - 56.9)
MLP	4	48.0 (36.3 - 59.7)	32.8 (24.3 - 41.3)	62.7 (60.5 - 64.8)	53.3 (50.2 - 56.3)
MLP	5	49.0 (40.3 - 57.6)	30.8 (23.7 - 37.9)	62.7 (60.5 - 65.0)	54.0 (50.3 - 57.7)
MLP	6	46.2 (34.2 - 58.2)	29.8 (20.0 - 39.5)	63.9 (59.5 - 68.4)	59.2 (54.9 - 63.5)
MLP	7	48.5 (37.0 - 60.1)	31.3 (18.7 - 43.9)	63.1 (57.7 - 68.5)	53.5 (49.6 - 57.5)
NB	1	69.7 (69.0 - 70.5)	68.1 (66.4 - 69.7)	63.9 (63.2 - 64.6)	59.0 (57.3 - 60.6)
NB	2	63.6 (60.4 - 66.8)	66.2 (65.0 - 67.5)	61.6 (59.5 - 63.8)	58.5 (57.1 - 59.9)
NB	3	58.4 (56.1 - 60.8)	60.7 (57.3 - 64.0)	54.5 (53.9 - 55.0)	58.6 (57.0 - 60.2)
NB	4	60.5 (58.4 - 62.6)	59.8 (58.3 - 61.4)	54.4 (53.0 - 55.8)	57.4 (55.1 - 59.7)
NB	5	64.6 (61.8 - 67.3)	58.9 (58.2 - 59.7)	58.1 (56.5 - 59.8)	56.8 (56.1 - 57.5)
NB	6	68.8 (65.2 - 72.4)	59.8 (58.6 - 61.0)	63.1 (61.1 - 65.1)	56.4 (55.5 - 57.3)
NB	7	72.3 (67.5 - 77.1)	59.4 (58.4 - 60.5)	68.3 (66.4 - 70.1)	55.0 (54.0 - 56.1)
RF	1	25.7 (21.4 - 30.0)	27.6 (26.0 - 29.1)	61.4 (60.3 - 62.5)	68.3 (66.7 - 69.9)
RF	2	26.3 (19.0 - 33.5)	30.6 (28.2 - 33.0)	61.6 (60.2 - 63.0)	70.2 (68.4 - 72.1)
RF	3	30.3 (27.2 - 33.4)	32.2 (28.3 - 36.0)	60.2 (58.4 - 61.9)	70.2 (68.5 - 72.0)
RF	4	25.1 (22.2 - 28.1)	32.2 (28.2 - 36.1)	57.7 (56.4 - 59.0)	72.5 (70.7 - 74.2)
RF	5	23.5 (21.2 - 25.8)	27.1 (25.3 - 28.9)	55.4 (53.5 - 57.2)	70.9 (68.7 - 73.1)
RF	6	23.3 (20.3 - 26.2)	25.6 (22.8 - 28.3)	55.8 (53.0 - 58.6)	73.0 (70.9 - 75.0)
RF	7	26.3 (22.5 - 30.1)	23.6 (21.9 - 25.3)	55.0 (53.1 - 57.0)	74.4 (72.0 - 76.8)
S-Attn	1	52.2 (42.9 - 61.4)	45.1 (38.4 - 51.8)	59.0 (54.7 - 63.3)	53.6 (50.7 - 56.4)
S-Attn	2	47.4 (34.0 - 60.7)	37.9 (29.8 - 46.0)	51.6 (45.0 - 58.2)	56.0 (52.2 - 59.7)
S-Attn	3	48.5 (37.7 - 59.4)	34.8 (25.8 - 43.8)	56.6 (53.8 - 59.4)	53.8 (50.9 - 56.6)
S-Attn	4	49.3 (36.6 - 62.0)	39.2 (31.4 - 46.9)	55.5 (51.6 - 59.4)	57.2 (53.4 - 61.1)
S-Attn	5	37.4 (28.2 - 46.6)	37.3 (33.0 - 41.6)	50.7 (44.4 - 57.0)	54.8 (49.3 - 60.2)
S-Attn	6	35.9 (25.7 - 46.0)	35.4 (27.9 - 42.8)	48.9 (42.1 - 55.7)	51.9 (47.0 - 56.8)
S-Attn	7	53.6 (35.0 - 72.2)	27.7 (22.6 - 32.9)	48.0 (38.7 - 57.3)	48.8 (42.8 - 54.7)
XGBoost	1	23.8 (21.5 - 26.1)	28.7 (26.0 - 31.5)	61.2 (59.5 - 62.8)	63.7 (62.4 - 65.1)
XGBoost	2	23.5 (20.6 - 26.4)	31.5 (27.3 - 35.8)	57.7 (53.8 - 61.6)	65.4 (63.4 - 67.5)
XGBoost	3	24.3 (18.5 - 30.1)	36.5 (33.2 - 39.8)	57.2 (53.8 - 60.6)	58.5 (55.7 - 61.3)
XGBoost	4	21.1 (19.0 - 23.1)	29.3 (27.8 - 30.7)	57.5 (56.7 - 58.3)	70.6 (68.3 - 72.8)
XGBoost	5	21.0 (19.1 - 22.8)	32.8 (29.2 - 36.4)	52.7 (49.8 - 55.7)	52.8 (47.8 - 57.9)
XGBoost	6	22.0 (20.2 - 23.7)	26.1 (23.5 - 28.8)	52.7 (49.8 - 55.6)	71.4 (68.7 - 74.1)
XGBoost	7	19.7 (17.6 - 21.7)	26.5 (24.8 - 28.2)	53.3 (51.3 - 55.4)	68.4 (66.2 - 70.7)

Source: A. Capstick *et al.*, npj digital medicine, 2024, https://www.nature.com/articles/s41746-023-00995-5

Analyse reliability of your model

IMPERIAL

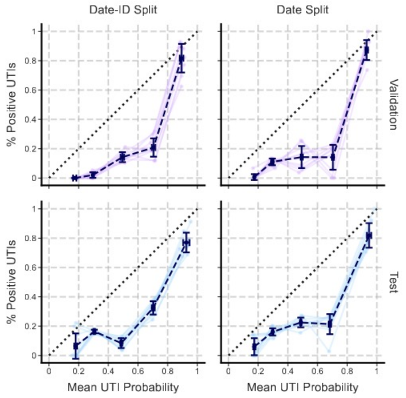


Supplementary Figure 6: Reliability plot. Top shows the model confidence (on positive and negative UTI cases) against accuracy on the test set of “Date Split” and “Date-ID Split”. The gap shows the difference between the average accuracy and confidence of a bin, which would ideally be 0. Bottom shows the histogram of confidences reported by the model on the test set of “Date Split” and “Date-ID Split”.

Source: A. Capstick *et al.*, npj digital medicine, 2024, https://www.nature.com/articles/s41746-023-00995-5

Investigate calibration

IMPERIAL



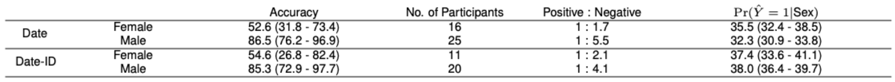
Supplementary Figure 7: Calibration plot. The mean UTI predicted risk against the proportion of positive UTI cases for the best model. This is plotted for the validation and test set from the “Date Split” and “Date-ID Split” data. The error bars represent the standard deviation of the values from the 10 bootstrap repeats.

Source: A. Capstick *et al.*, npj digital medicine, 2024, https://www.nature.com/articles/s41746-023-00995-5

Investigate for bias

IMPERIAL

Table 3: Mean (95% CI) % of the accuracy on the Male and Female participants on the test set of the different data splits with 10 bootstrap repeats. We also show the ratio of positive and negative labels in each demographic and the likelihood of a positive model prediction.



		Accuracy	No. of Participants	Positive : Negative	$\Pr(\hat{Y} = 1 \mid \text{Sex})$
Date	Female	52.6 (31.8 - 73.4)	16	1 : 1.7	35.5 (32.4 - 38.5)
	Male	86.5 (76.2 - 96.9)	25	1 : 5.5	32.3 (30.9 - 33.8)
Date-ID	Female	54.6 (26.8 - 82.4)	11	1 : 2.1	37.4 (33.6 - 41.1)
	Male	85.3 (72.9 - 97.7)	20	1 : 4.1	38.0 (36.4 - 39.7)

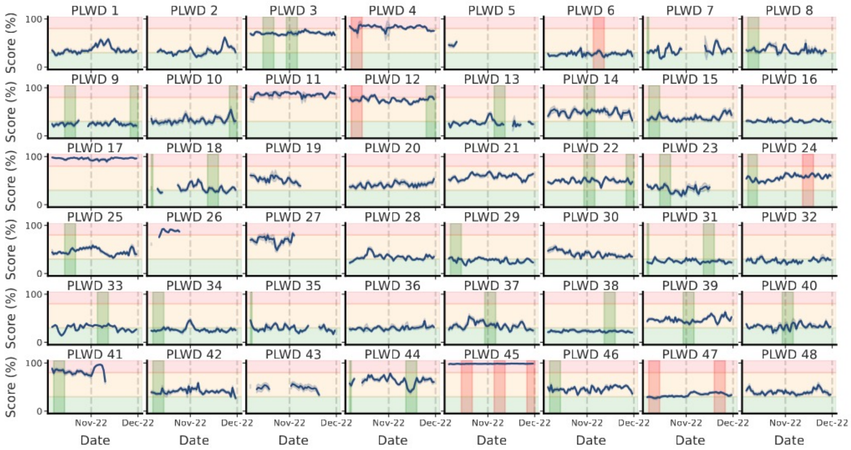
Source: A. Capstick *et al.*, npj digital medicine, 2024, https://www.nature.com/articles/s41746-023-00995-5

41

41

Decision-support vs. current practice

IMPERIAL



Source: A. Capstick *et al.*, npj digital medicine, 2024, https://www.nature.com/articles/s41746-023-00995-5

42

42

Important notes in reporting

- Discuss ethical considerations and potential risks
- Learn about clinical context and use (needs to start before technical development); for example, UTIs are recurrent, and a person with a past history of UTI could have a higher risk of having another UTI.
- Report ethical approvals and regulatory requirements/approvals (if applicable)
- Describe data availability
- Capture and report code, libraries used and versions (or create a reproducible environment)
- Report any pre-processing, imputation, scaling and normalisation
- Report libraries and hyperparameters
- Avoid any leap of magic!

Responsible and ethical machine learning

- Anticipate risks and potential issues
- Actively engage and think of solutions
- Work with domain experts and learn about the context and conditions of data collection, and the use of system/model.
- Create feedback loops (before and after deployment)
- Plan for update and maintenance

Responsible and ethical machine learning

IMPERIAL

- Consider all issues, including bias, fairness, and the risk of harm, throughout your design, development, evaluation, and deployment.
- Consider sustainability requirements and scaling/generalisation.
- Considering these requirements is not somebody else's responsibility but the ML model designers'.

Regulatory framework and existing guidelines

IMPERIAL

- In designing and deploying ML models for healthcare and medical applications, you need to investigate the regulatory requirements.
- e.g., UKCA (UK), CE (Europe), FDA (USA) requirements and medical device registration.
- The existing guidelines for treatment and interventions. For example, the NICE guidelines: https://www.nice.org.uk/guidance
- EU Regulatory framework proposal on Artificial Intelligence, https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Further reading

- A good article on this topic:
  - 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

If you have any questions

- Please feel free to arrange a meeting or email (p.barnaghi@imperial.ac.uk).
- My office: 928, Sir Michael Uren Research Hub, White City Campus.