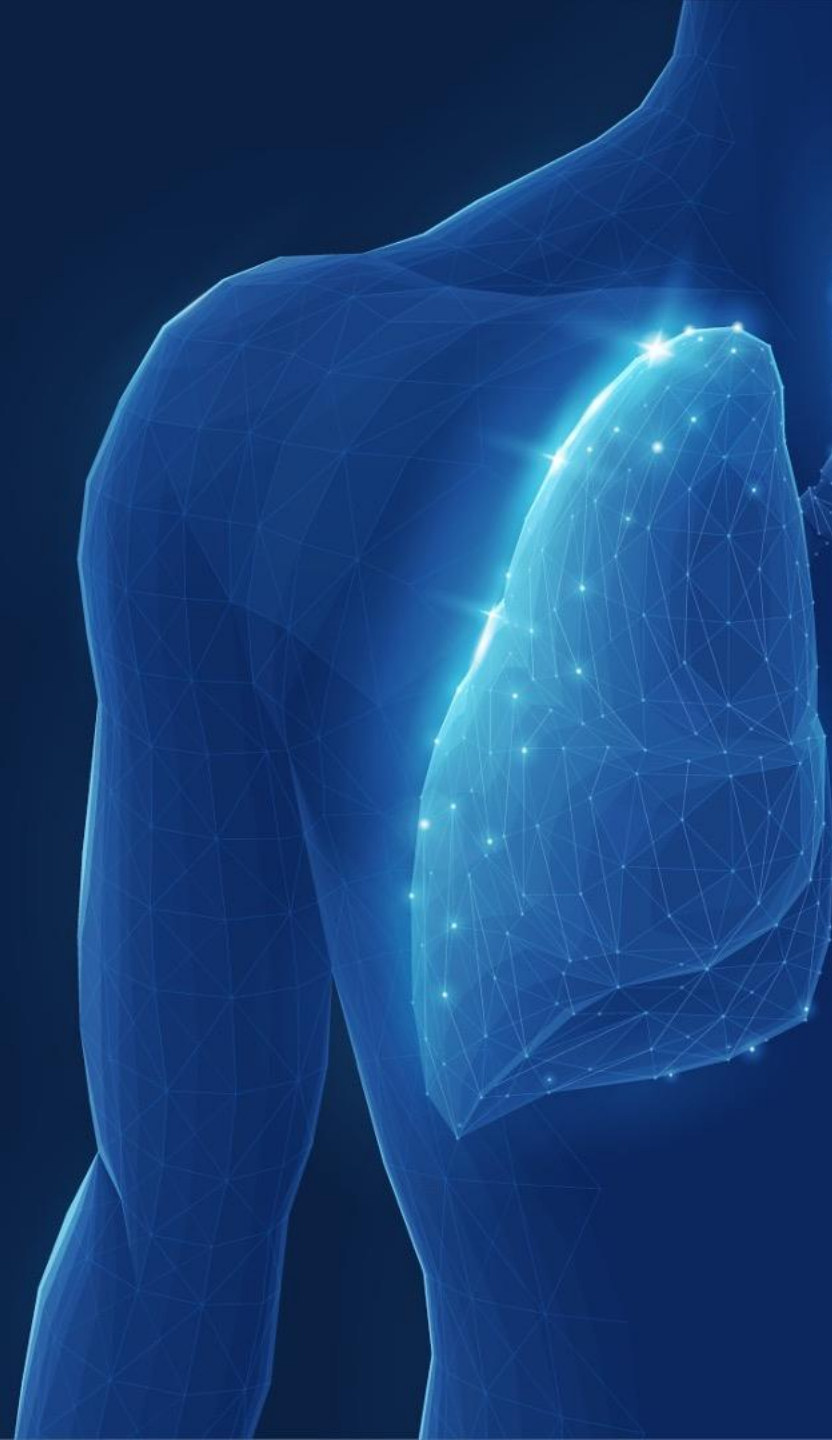
Lenus

Adopting ethical AI insights to improve respiratory care.



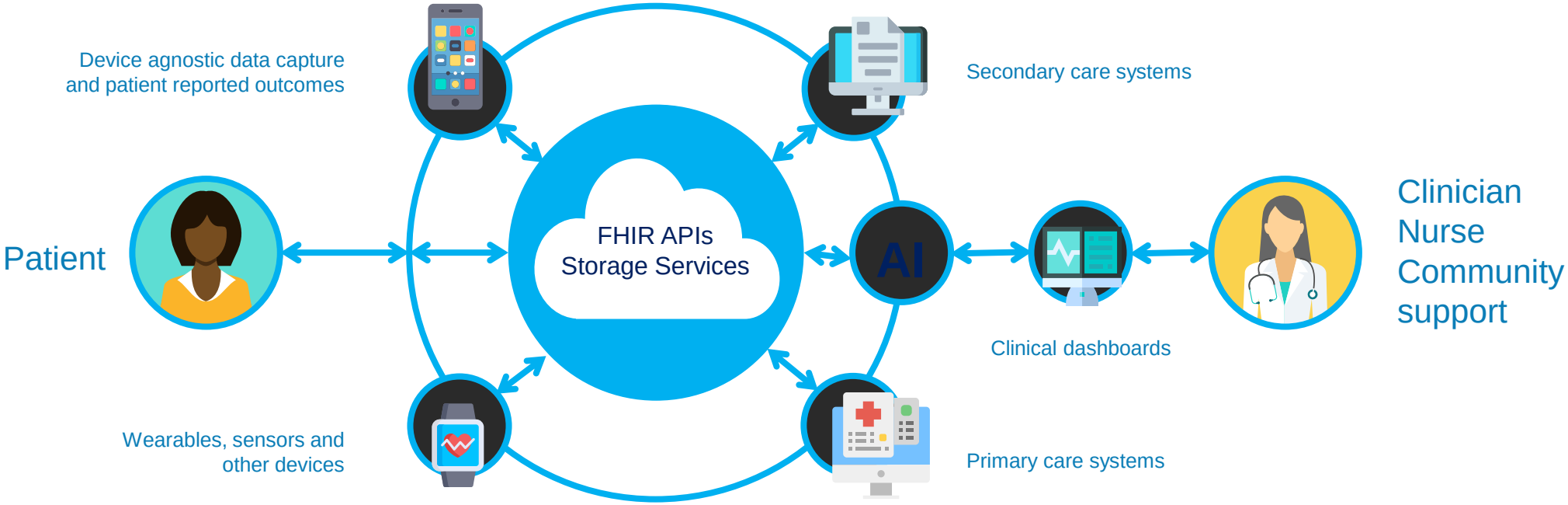
Shane Burns

Ampersand Advocates seminar September 2023

Overview

1. Lenus Health: Some background context
2. What is AI/ML?
3. Key aspects to consider for model approval
4. COPD AI insights: Lessons learned so far
5. Conclusions

Lenus Health Platform - secure healthcare data exchange



A virtual care platform providing a system of interoperability for secure health data exchange developed on Azure PaaS

Infrastructure Layer

Offers identity, consent, security, data capture, curation, storage and integration with systems of record.

Analytics Layer

Exploits structured datasets across patient and clinical systems to create AI-driven actionable insights

Services Layer

Provides an open environment for co-development of digital care pathways enabled through platform APIs

What is Artificial Intelligence (AI)?

- AI is a broad term which describes the use of software/machines that mimic human cognition to perform tasks of varying complexity.
- A simple AI could be a rule-based risk calculator and may not involve any machine learning solution.
- Machine Learning (ML) is a subset of AI that uses algorithms to find patterns and relationships in large datasets.
- At Lenus, we utilize ML to develop risk prediction models. The developed model is then used to make predictions on unseen data.
- For example, a certain model may take as input recent lab test results and as output provide a risk score that the patient has a certain disease.

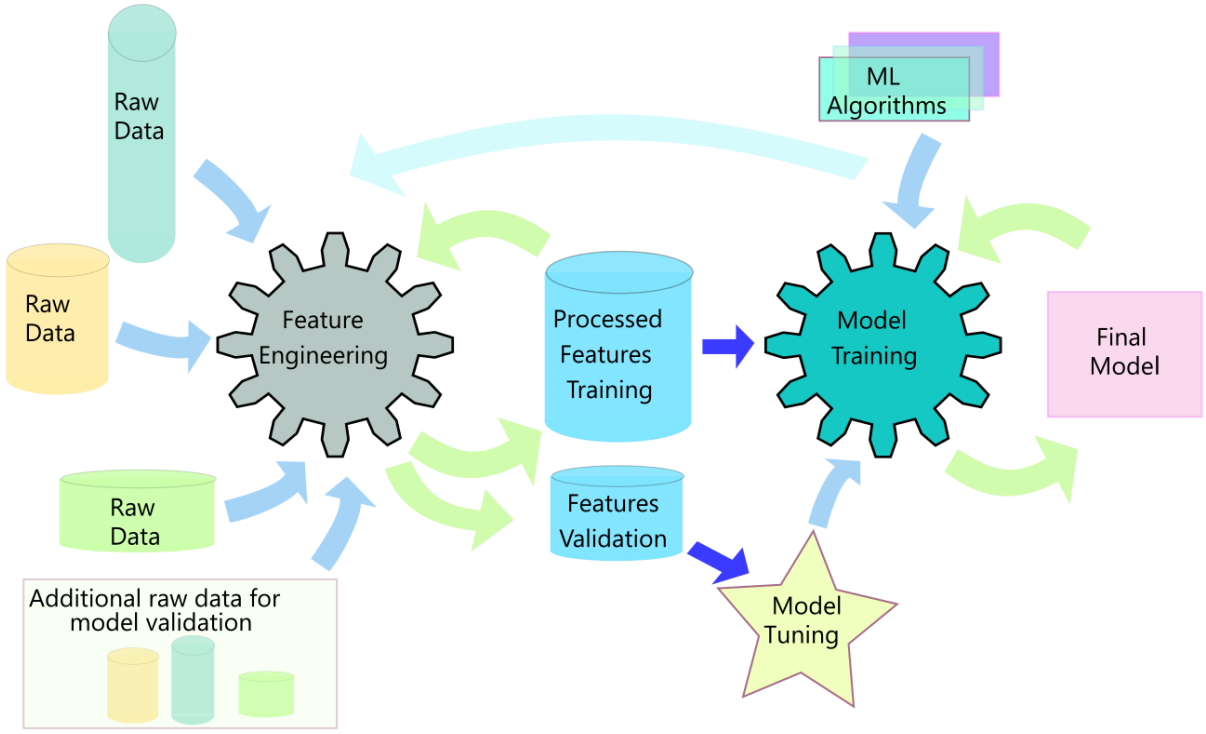
What type of data do we use to develop these models?

- PROs data, COPD status data, lung function, smoking
- Wearable device data
- Home NIV data
- Demographics
- Laboratory
- Prescribing
- Hospital admissions
- Comorbidities
- Labels

ML Model development overview

- Problem identification and formulation.
  - For example, 1-year mortality prediction formulated as a classification problem (1 deceased in 1 year, 0 not deceased)
  - One row of data per patient per year
- Identification of relevant/available data
  - Feature engineering
- Model training
  - Algorithm choice + labels
  - Parameter tuning
- Model testing
  - Model explainability
  - Fairness evaluation

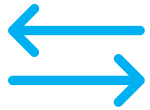
ID	AGE	admissions	…	BMI	last_albumin	label
1	77	2	….	17	29	1
2	68	0	….	22	45	0
3	80	3	….	21	32	0



Lenus Stratify™ Model Suite

Model	Identifies patients based on	Patient Prioritization within Pathway
PLAN	▪ high risk of mortality in 12 months	▪ Initiate patient review and anticipatory care planning. ▪ Case finding tool for advanced therapies.
ACT	▪ high risk of hospital readmission in 3 months	▪ Identify patients for therapy review to prevent downstream admission ▪ Case find for advanced therapies.
ALERT	▪ high risk of having an exacerbation event in next 3-5 days	▪ Contact immediately to initiate rescue medication and care
CLASS	▪ three common clusters in a population	▪ determine if patients are receiving guideline directed therapy based on their risk profile ▪ prioritise those in need of review

Our approach to AI



Bias-free

- Routinely collected healthcare data is full of biases.
- Building models without mitigating against these can have detrimental outcomes.

Explainable

- Explainable model predictions are essential to ensure both clinician and patient trust in the AI
- Explainability is also a powerful technique for bias hunting

Fair

- Evaluation of model performance, selection rate, and explainability across protected demographics.
- Fairness-aware model training
- Representation/inclusion criteria

What is needed to approve a model for use in a clinical setting?

- This methodology has been co-developed with our clinical collaborators at GGC (Greater Glasgow and Clyde).

- Model approval session occurs anytime there is a significant change to an existing model, or if a new model is being added.

- The key aspects of this are:
  1. Explainability
  2. Validation performance
  3. Calibration
  4. Fairness

Model approval checklist and notes

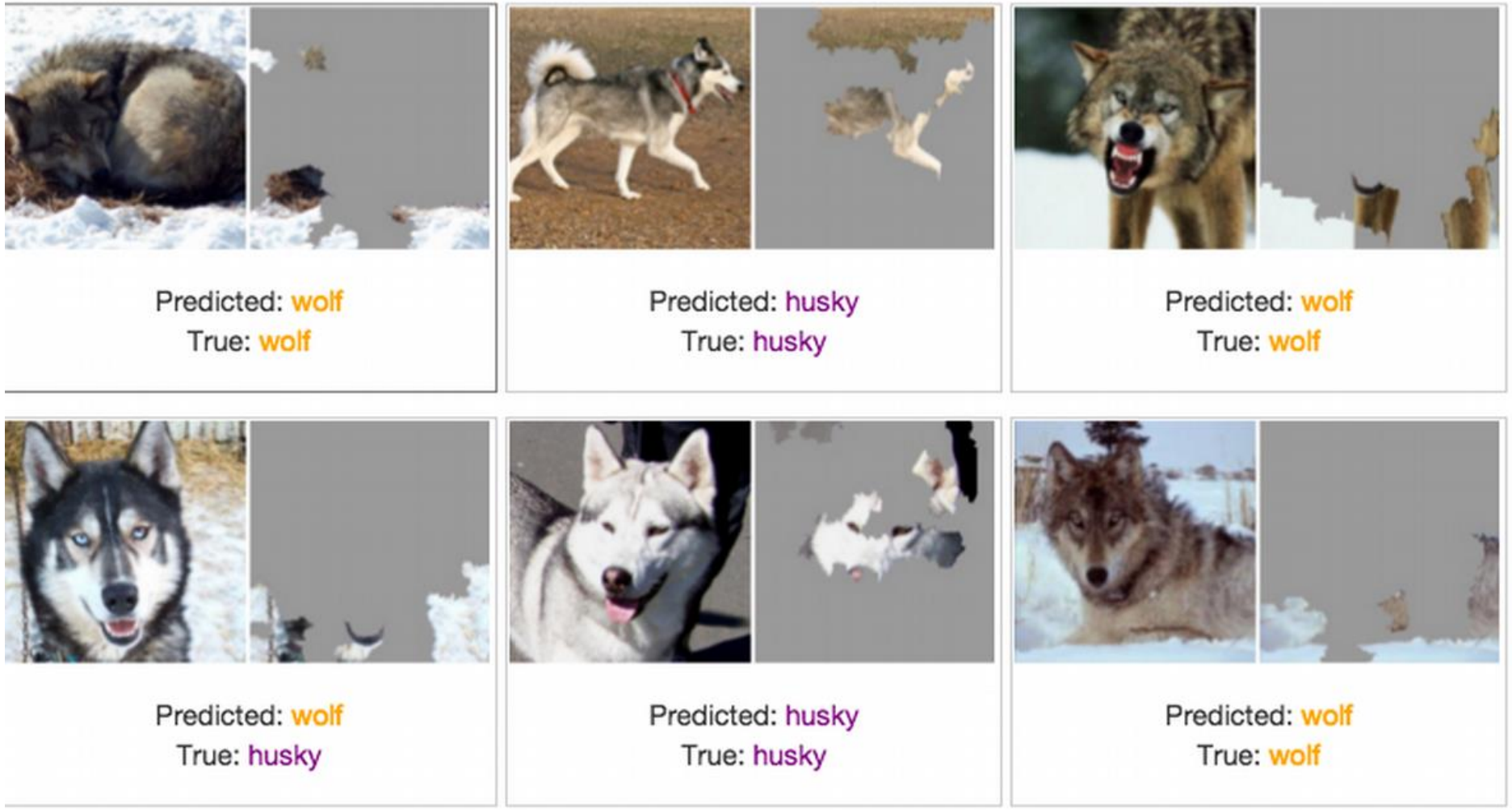
Model approval meeting notes

Model approval checklist

- Has the model gone through Lenus Engineering QA? ✓
- Are the training and holdout test cohorts independent and comparable? ✓
- Is the model formulation and algorithm suitable? ✓
- Are the model features: ✓
  - Appropriate for the intended use-case? ✓
  - Engineered appropriately? ✓
- Is the model calibration satisfactory? ✓
- Is the model performance satisfactory? In particular: ✓
  - The area under the precision recall curve (PR-AUC)? ✓
  - Expected numbers of patients brought forward correctly/incorrectly and missed? ✓
- Is the global model explainability bio-plausible? ✓
- Is the model fairness on different population sub-groups satisfactory? ✓

Summary
Model version: v2.2.0
Training date: 02/03/2023
Model approval meeting date: 20/3/2023
Model approval meeting outcome (approve): Yes

Model Explainability – why is it important?



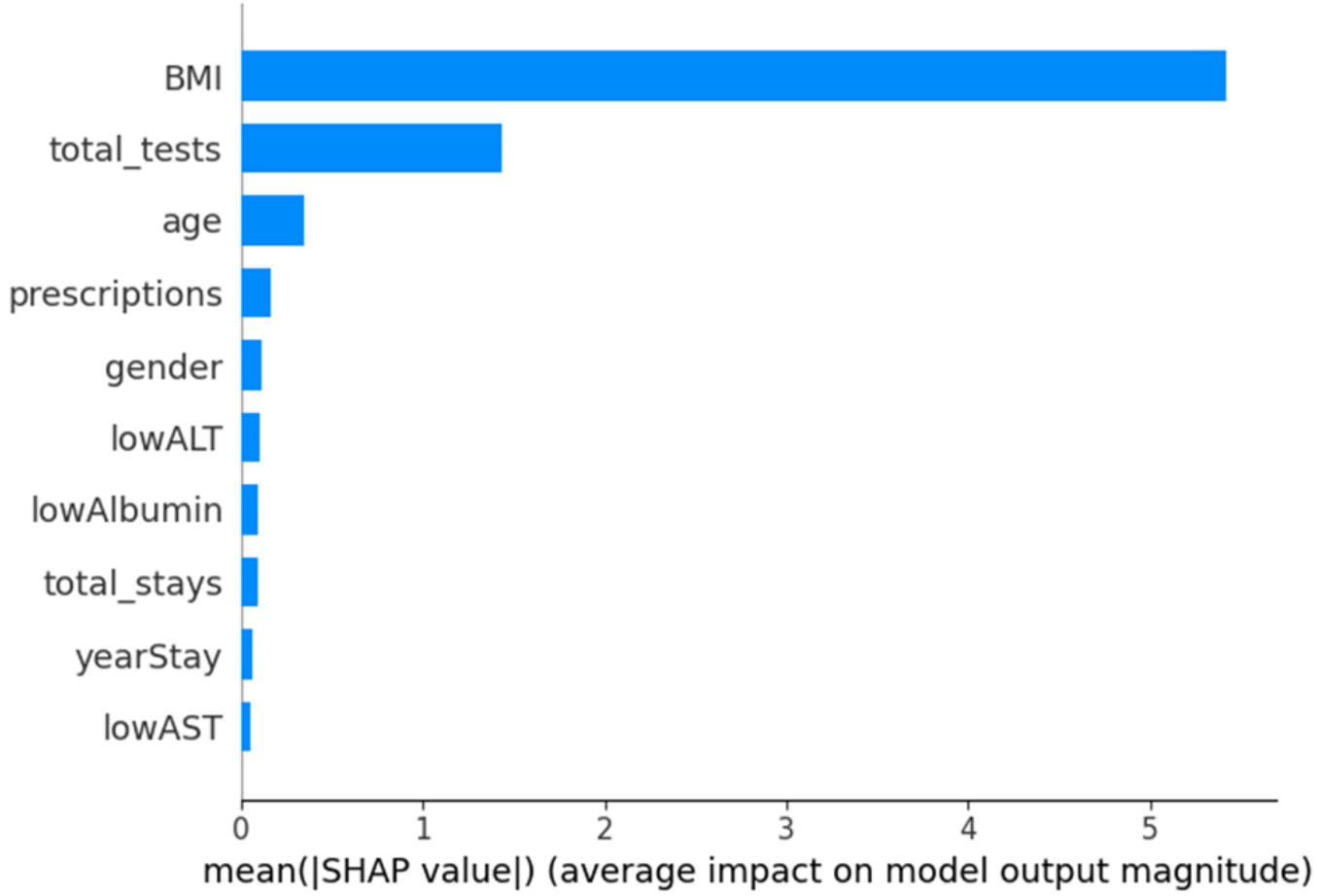
Source: Explaining Black-Box Machine Learning Predictions – Sameer Singh

Model Explainability – why is it important?

- Melanoma prediction model achieved an accuracy greater than humans
- Explainability techniques similar to the one used in the previous slide were thankfully applied to the model
- The analysis showed that the model was mostly identifying measuring rulers
- Blindly using this model would have a detrimental impact on patients

Model Explainability – why is it important?

- Explainability and feature importance can be calculated using techniques of varying complexity.
- Global explainability describes what data features are important to the model overall
- and local explainability describes what is important on an individual patient prediction level.
- This allows for interrogation of the model prediction and interpret the bio-plausability of the model prediction.



Model explainability - implications

- R. Caruana, Y. L. (2015). Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission.
Proceedings of www.annualreviews.org • Ethical Machine Learning in Health Care 21, 1721-1730.

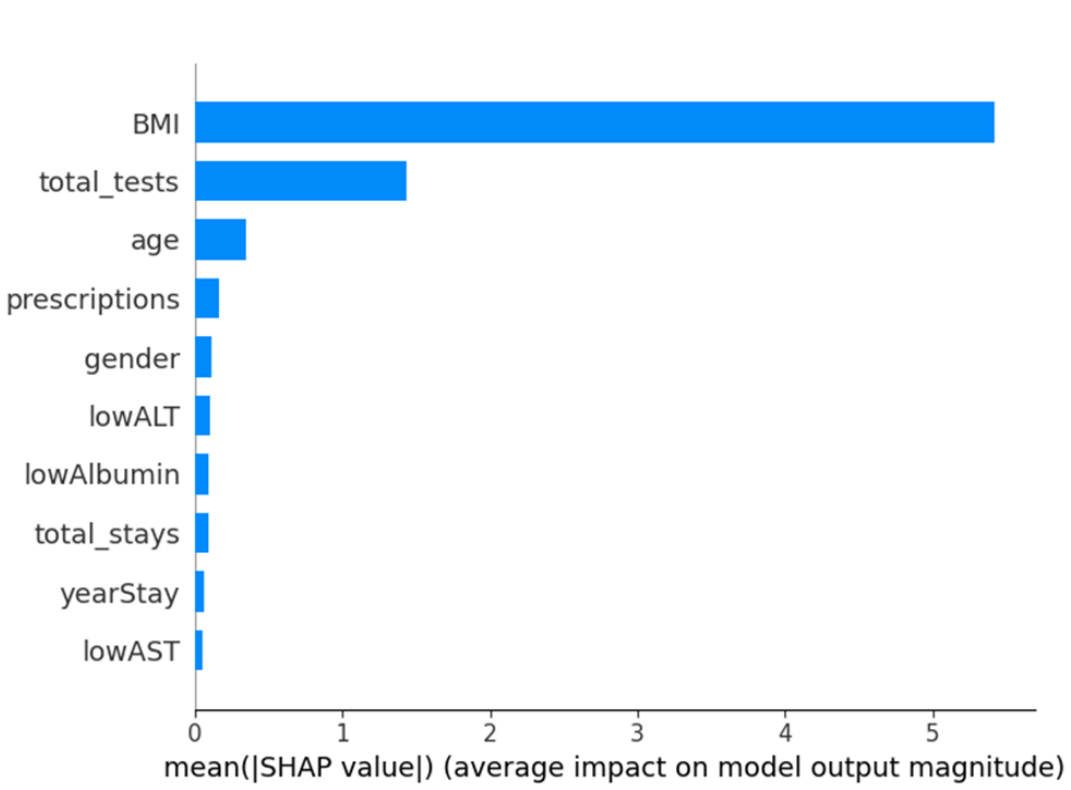
- Authors found that patients with asthma who presented with pneumonia had a **greater** survival chance than those without asthma. They found this using model explainability.

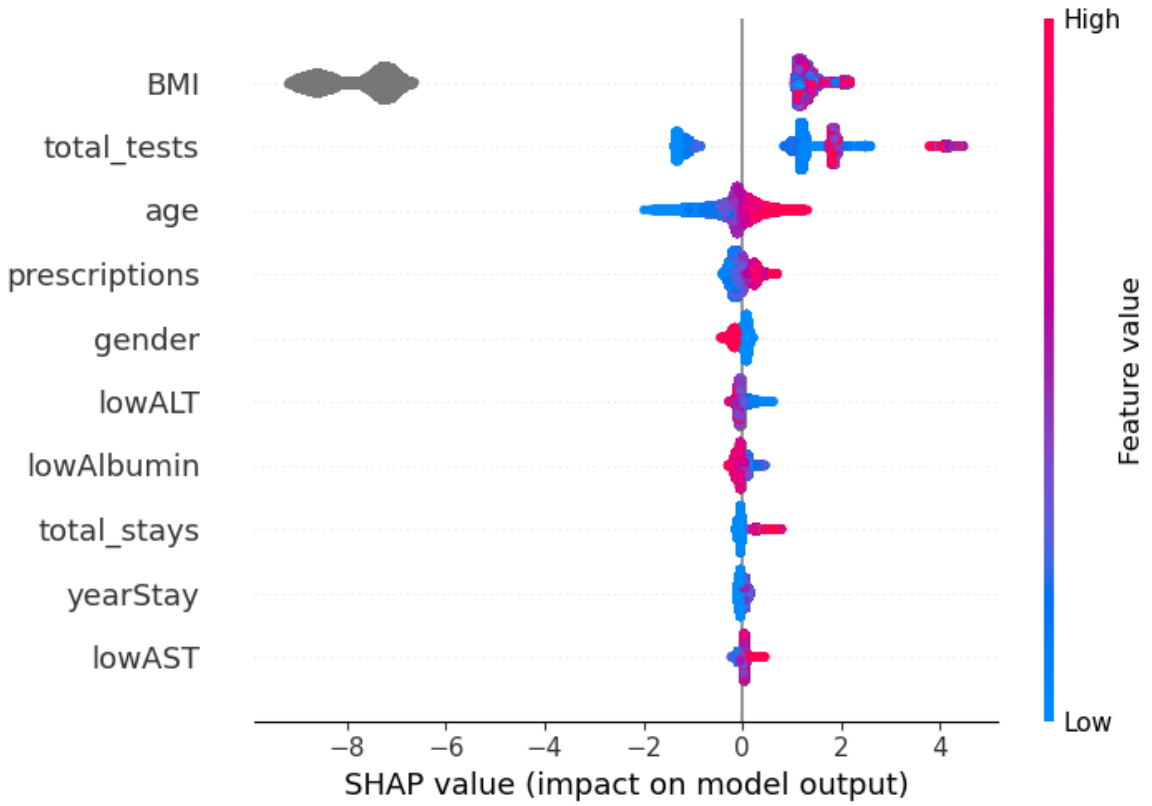
- The authors further identified the confounding effect which lead to this observation.

- Interrogation of the model found that the hospital's treatment policy was different for patients
- presenting with pneumonia and asthma.

- Without such rigorous considerations, models may be blindly deployed and could have the potential to do harm.

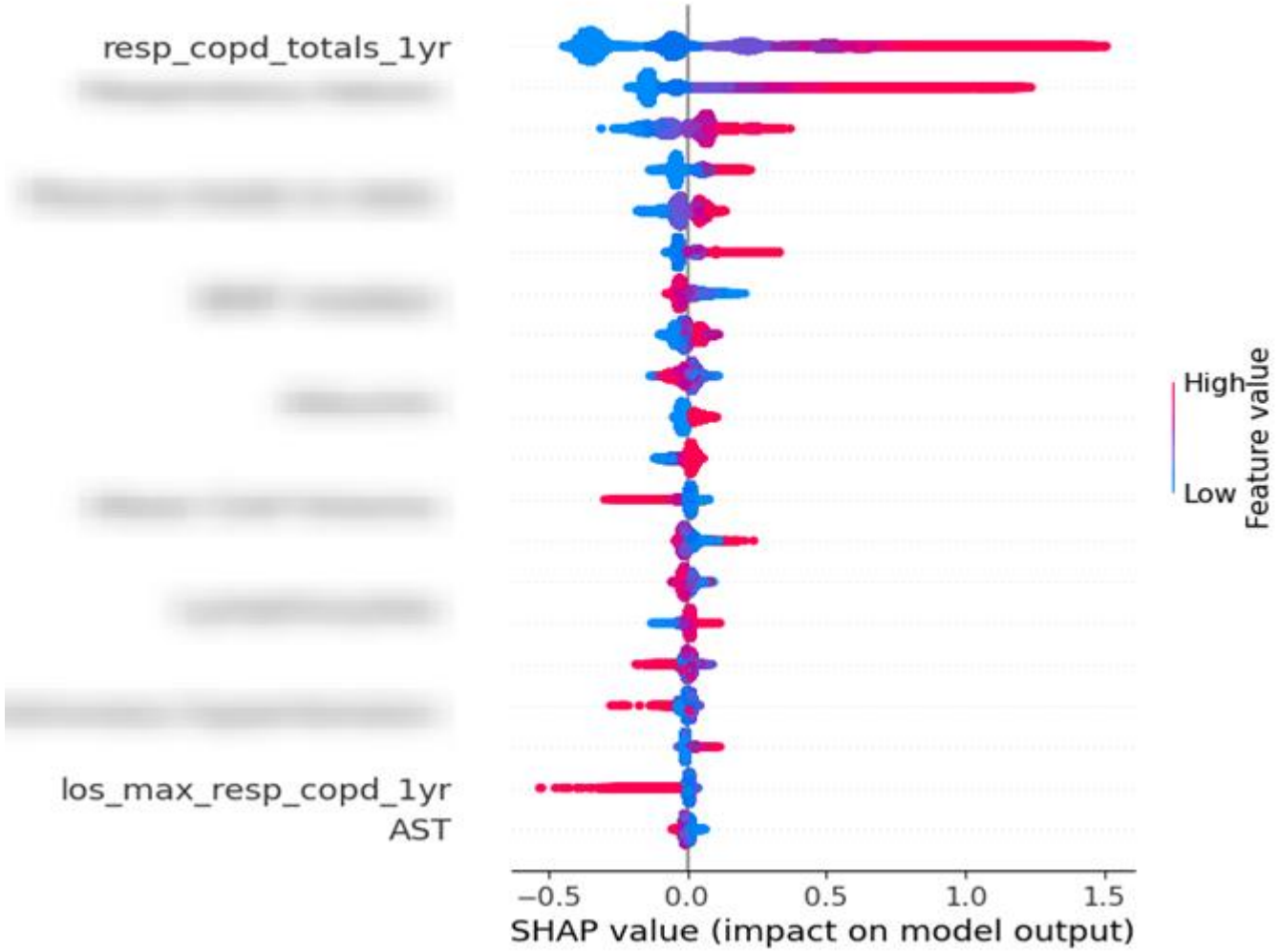
Using explainability to find biases





SHAP (SHapley Additive exPlanations) summary plot. This particular bias highlights the issue with understanding the context of missing data, naively using a sparsity aware algorithm, and the consequence for generalizability.

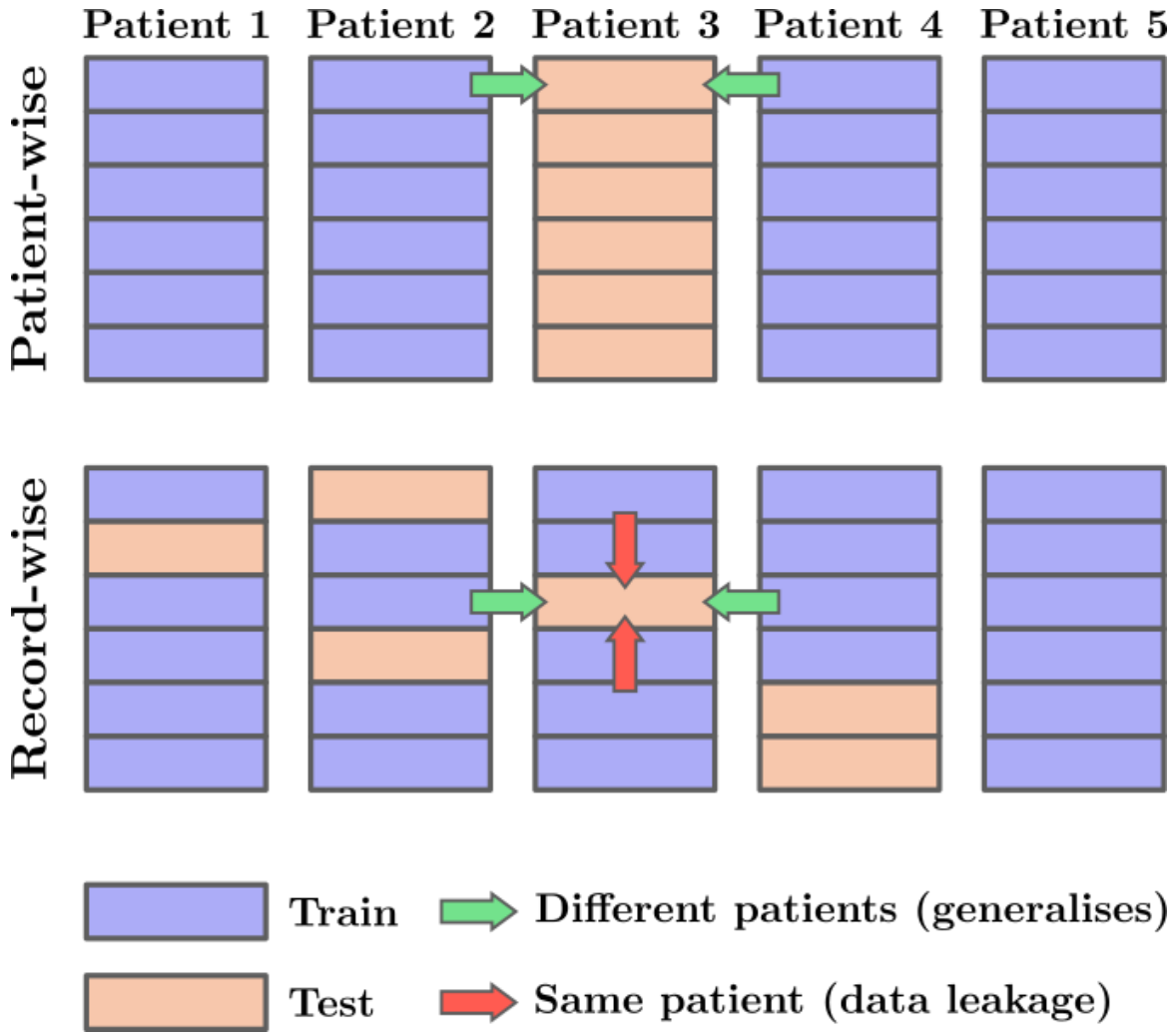
Using explainability to find biases– readmission prediction example



- Most features bio-plausible but some potential biased features visible
- Los_max_resp_copd_1yr for example
  - Long stays corresponding to no re-admission is capturing people who are still in hospital 3-months after the admission index which is not the correct context

How do you validate performance?

- Sohrab Saeb, L. L. (2017). The need to approximate the use-case in clinical machine learning. *GigaScience*.
- The authors reviewed 369 articles using patient-wearable data for clinical prediction.
- Of the articles that met the study criteria, almost half incorrectly used record-wise validation.
- The results give misleadingly high accuracies

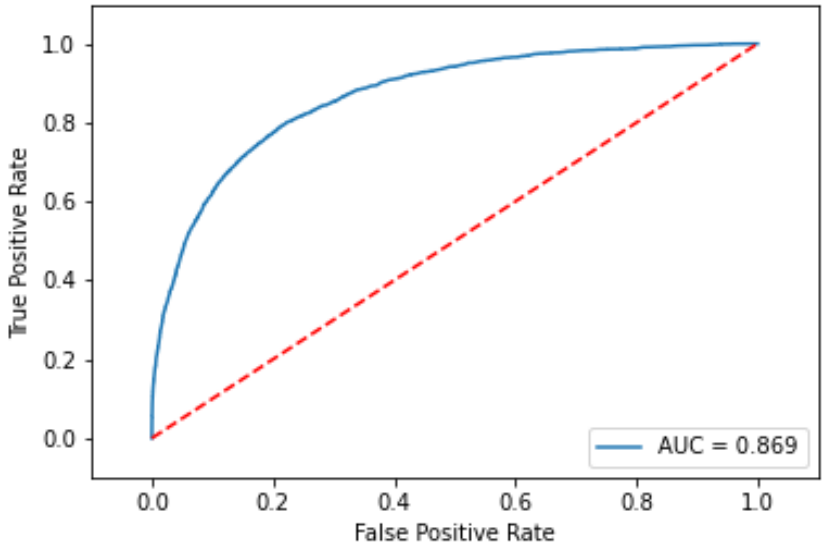


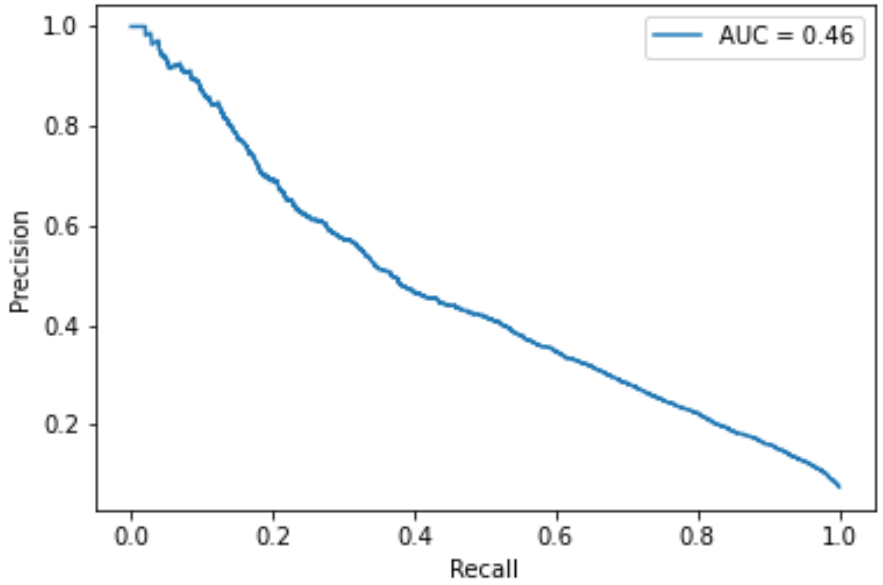
Model performance

- In healthcare settings we typically deal with large class imbalance. Metrics such as ROC-AUC and accuracy alone are not appropriate to report on model performance. For example, if a dataset contains 99% *no-disease* and 1% *disease*, a model that simply predicts everyone as having no-disease will be 99% accurate.

Metric	Definition	Interpretation
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	How many patients were classified correctly?
Precision	$\frac{TP}{TP + FP}$	How many positive patients identified are relevant?
Recall	$\frac{TP}{TP + FN}$	How many relevant patients were brought forward as positive ?
F1 Score	$\frac{2(precision \times recall)}{precision + recall}$	Harmonic mean of precision and recall
ROC-AUC		Area under the Receiver Operating characteristic curve
PR-AUC		Area under the precision recall curve
Brier loss		Measure of how well calibrated the model is

Model performance – mortality prediction example





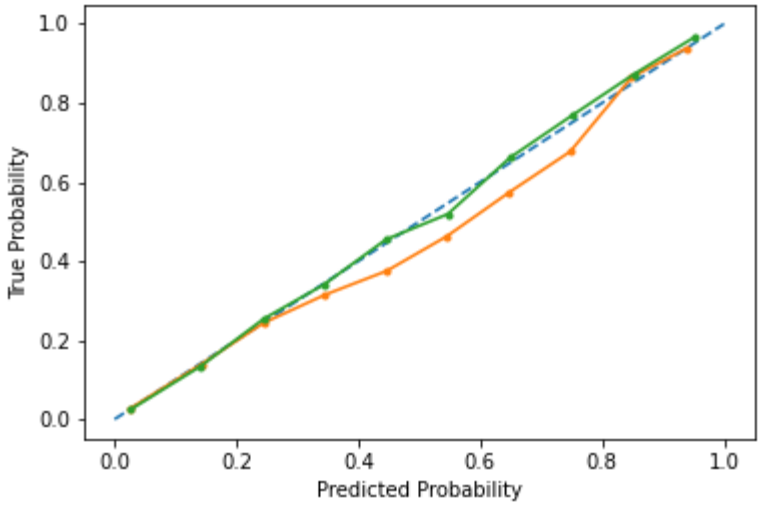
ROC curve, left, and PR curve, right. A perfect classifier would have both values = 1.

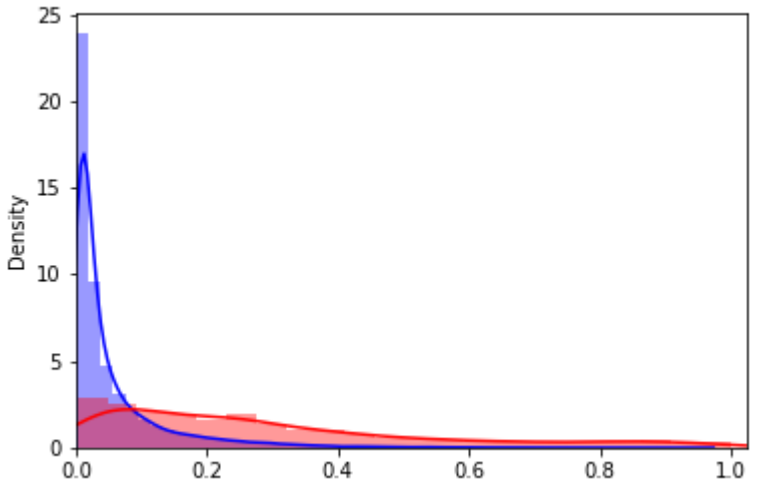
$$Confusion\ matrix\ scaled\ to\ 100\ patients = \begin{pmatrix} 4 & 3 \\ 4 & 89 \end{pmatrix}, \begin{pmatrix} TP & FN \\ FP & TN \end{pmatrix}$$

The model correctly predicts 89 (out of every 93) surviving the next 12 months (bottom row).
The model correctly predicts 4 (out of 7) patients (top row) of being deceased in the next 12 months but misses 3.

Why model calibration and thresholding are important

- If a model is perfectly calibrated, the inference probabilities will match the true probability of event, e.g., half of patients with a score of 0.5 will truly be deceased within 12 months.





- Model calibration curve (left). Blue dashed line; perfectly calibrated, orange line; uncalibrated, green line; calibrated using isotonic regression. The uncalibrated model is over-confident in prediction.

- Calibrated inference probability distribution. Blue alive at 12 months, red deceased.

Model calibration and thresholding

- To action a ML model in a clinical setting it is necessary to select a threshold probability to act on.

- If you actioned patients with a risk of 25% and above, the model would correctly classify 3 out of 7 true mortality cases but incorrectly bring forward 4 False Positives.

- If you actioned patients with a risk of 40% and above, the model would correctly classify 2 out of 7 true mortality cases and only incorrectly bring forward 1 False Positive.

- This interpretation is only possible with a calibrated model.

Probability threshold	Number of patients predicted mortality scaled to 100 patients		
	Correct	Incorrect	Missed
0.25	4	4	3
0.4	2	1	5
0.8	1	0	6

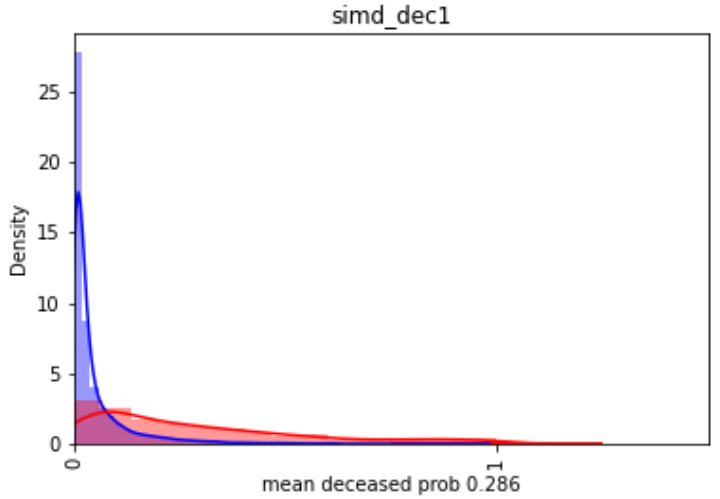
The numbers in the table represent the hold out test dataset scaled to 100 patients. 93 patients were alive after 12 months and 7 were deceased

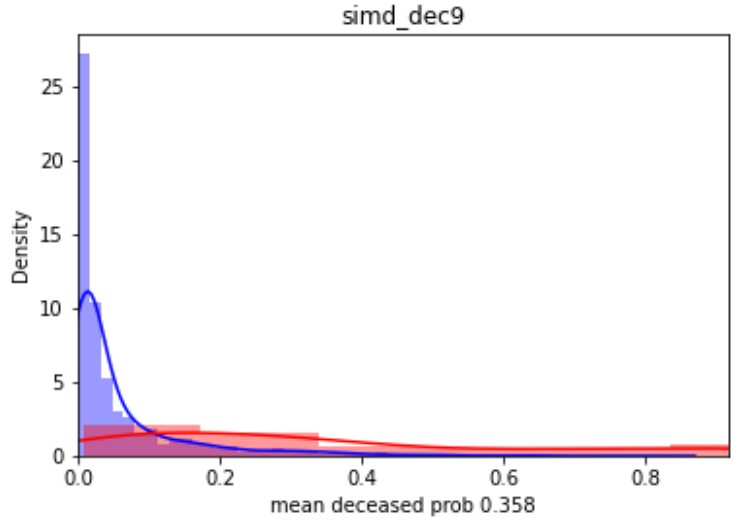
Why fairness is important

- Fairness is important to assess during model development and once the model is deployed
- We evaluate model performance, explainability, and selection rate for protected demographics
- Global explainability can vary a lot across demographics
- Selection rate can vary a lot across demographics
- The above nuances need to be identified and addressed during model development

Model fairness by SIMD (Scottish index of Multiple deprivation)

Group	Class	Precision	Recall	Positive selection	Validation % counts	Training % counts
SIMD 1	Deceased	0.45	0.42	6.4%	36.0	35.9
SIMD 2	Deceased	0.42	0.47	8.4%	18.9	18.9
SIMD 3	Deceased	0.44	0.42	7.1%	10.4	10.7
SIMD 4	Deceased	0.46	0.47	7.2%	8.4	8.1
SIMD 5	Deceased	0.41	0.45	8.8%	6.2	6.3
SIMD 6	Deceased	0.47	0.47	8.0%	4.7	4.7
SIMD 7	Deceased	0.44	0.56	13%	4.6	4.5
SIMD 8	Deceased	0.43	0.52	8.5%	3.7	4.0
SIMD 9	Deceased	0.41	0.51	9.6%	3.7	3.9
SIMD 10	Deceased	0.46	0.51	8.0%	3.3	3.1





- Model performance by SIMD at a threshold of 0.25
- Model is more performant in less deprived areas
- This is exacerbated when choosing a higher threshold
- Model is more confident in less deprived areas (0.358 vs 0.286)

Steps taken to help ensure clinical safety

- It is important that the model has been formulated correctly and that we understand the data. If not, we run the risk of unknowingly introducing bias.

- Domain expertise is key. We rely on our clinical collaborations when choosing appropriate features and examining feature bio-plausibility.

- It isn’t enough to evaluate a model in terms of performance. Explainability and fairness need to be considered.

- TSET – University of Swansea collaboration

- BS30440: Validation Framework for the Use of AI in Healthcare

NHS SCOTLAND COPD

COPD Insights

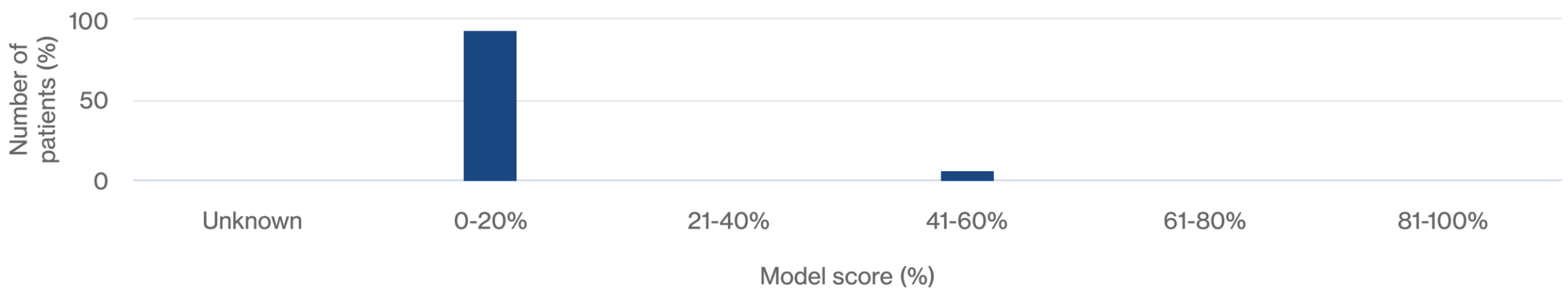
Sign in to the COPD Insights service.

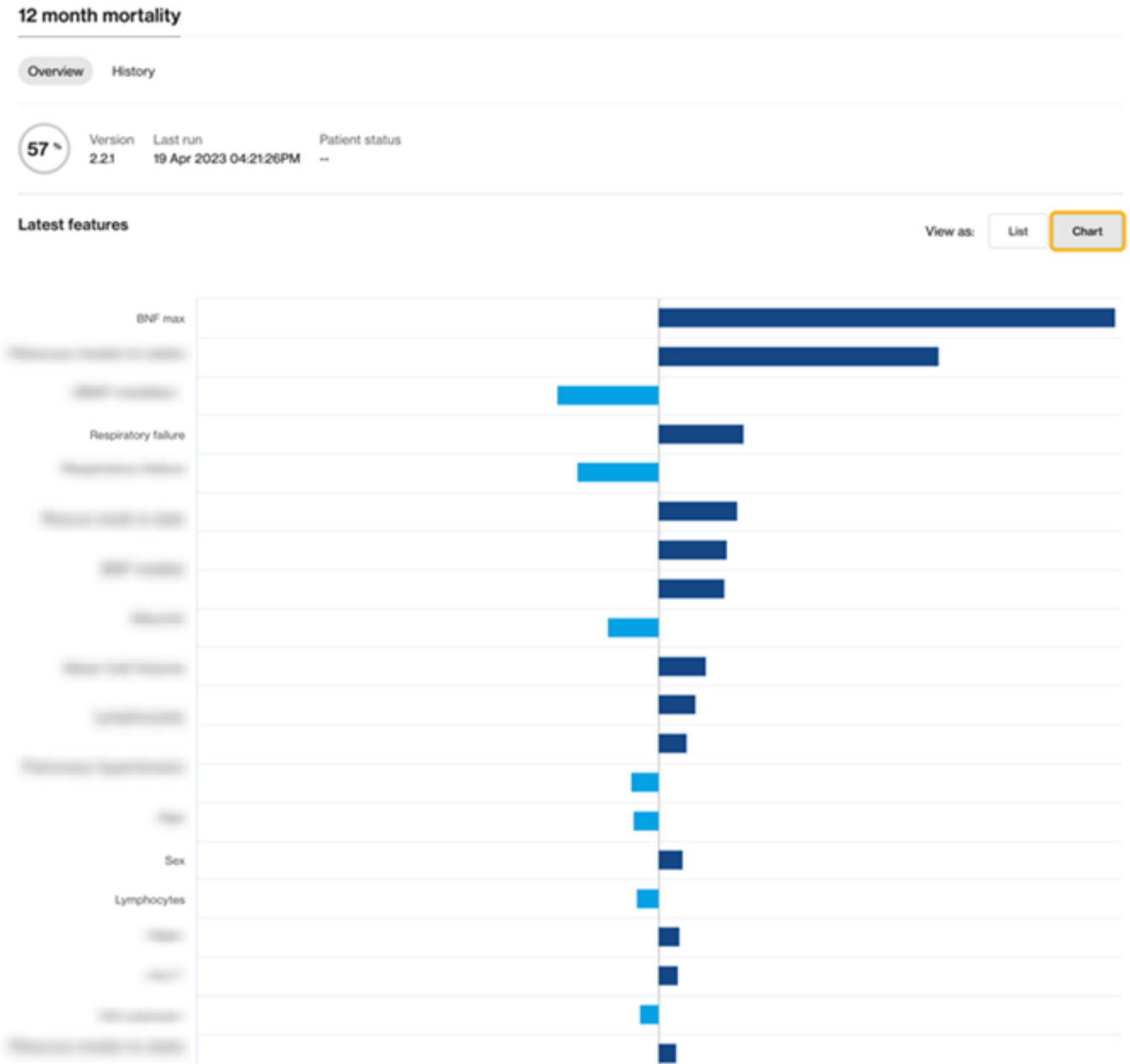
Log In

Patients

Patient cohort % of risk

Model run: 19 Apr 2023 04:21:23PM
Model version: 2.2.1
Number of patients: 17





Hi [redacted]

Thanks for joining the DYNAMIC-AI trial. We got the first run of data through from it, and it's really interesting, with nothing worrying. It did flag one thing - that it might be worth checking your overnight breathing or blood gases again at some point. We could have a chat about that - not urgent, but I'd have clinic space + could give you a call tomorrow or thursday sometime, if any time either day would suit you? Chris

Chris Carlin - 25 April 2023 13:28

Thank you

shane.burns@lenushealth.com

lenushealth.com

@lenushealth

@Lenus Health

Shane Burns

September 2023