How to evaluate Machine Learning models in Laboratory Medicine?

International Federation of Clinical Chemistry and Laboratory Medicine

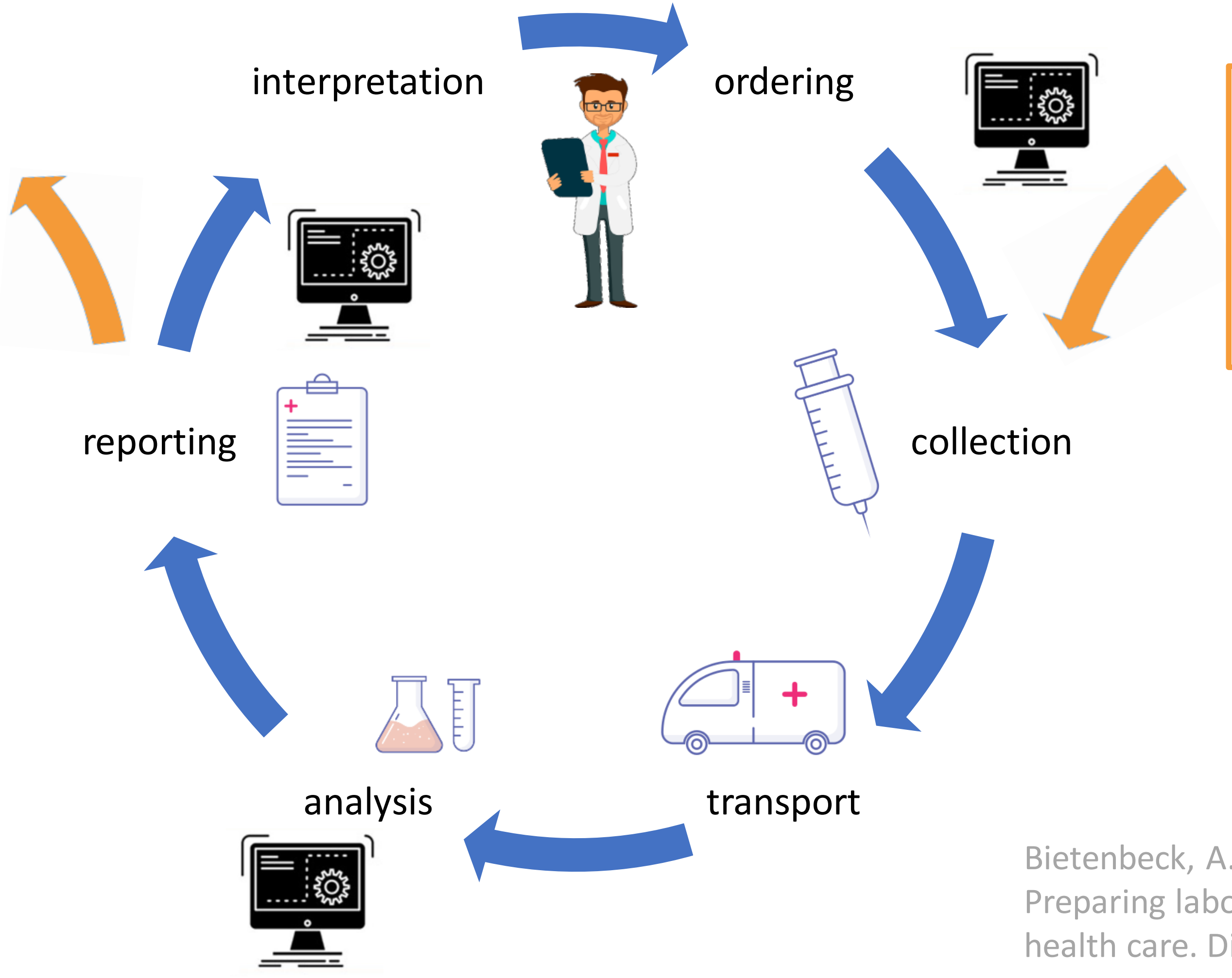
Advancing excellence in laboratory medicine for better healthcare worldwide

13.02.2024
Andreas Bietenbeck (lab@bietenbeck.net)
Germany

ifcc.org

Machine Learning applications can facilitate in all phases of Laboratory Medicine

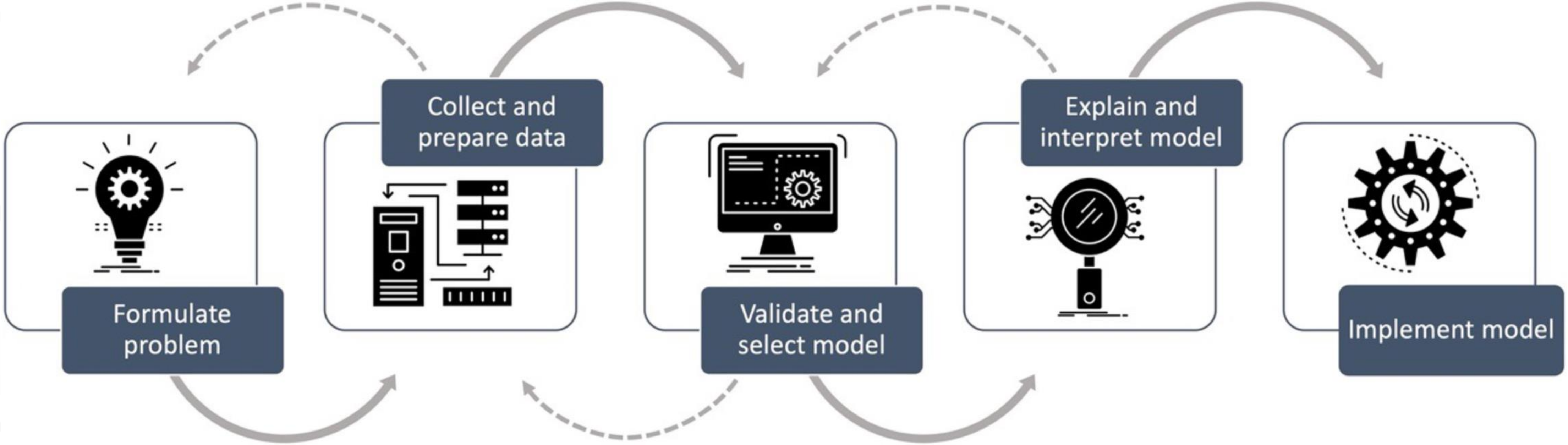
- Electronic Health Records
- patient
- Clinical decision support systems
- research
- …



- Clinical information
- Results from other laboratories

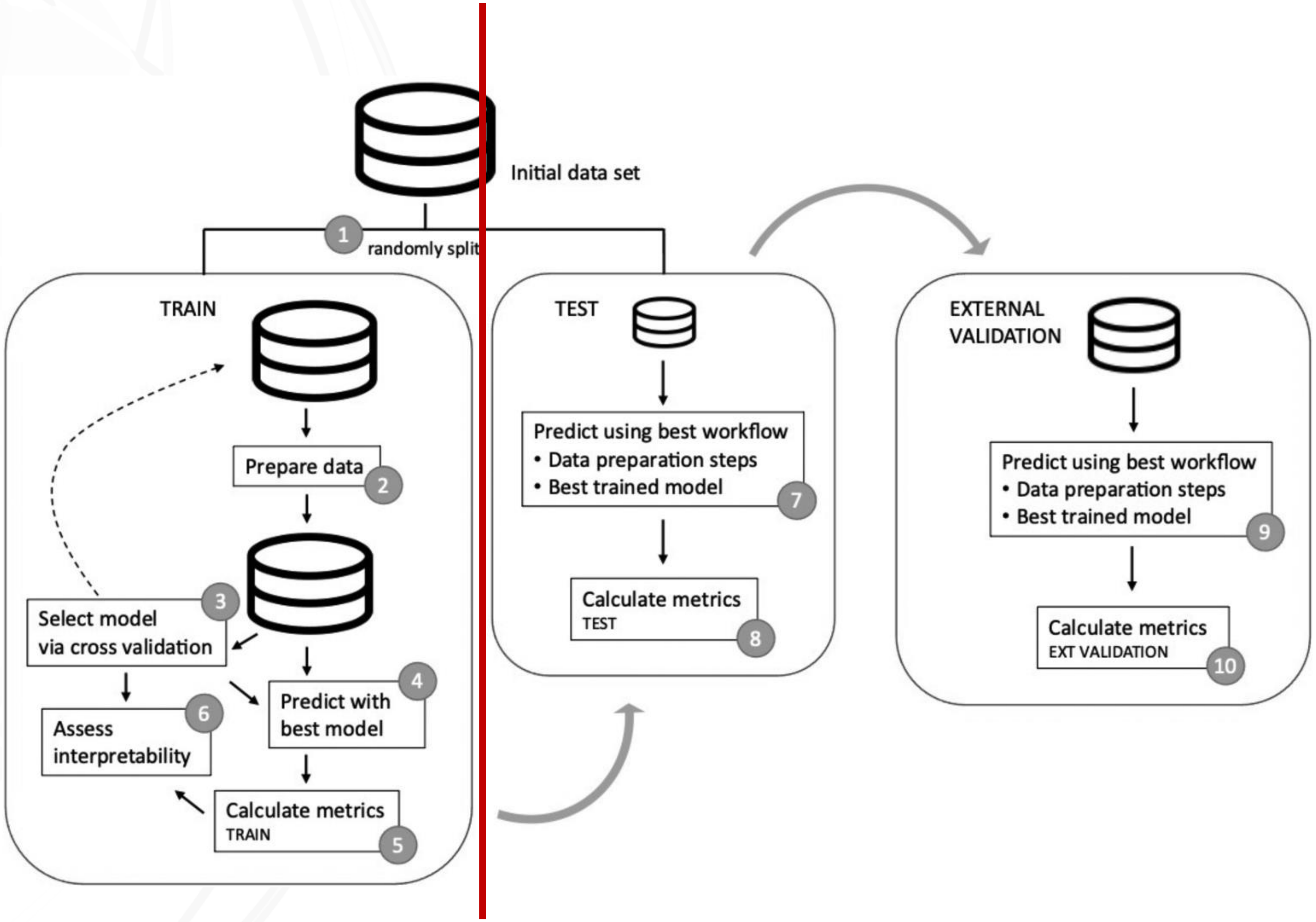
Bietenbeck, A., & Streichert, T. (2021). Preparing laboratories for interconnected health care. Diagnostics, 11(8), 1487.

Overview



Collect and prepare data
Explain and interpret model
Formulate problem
Validate and select model
Implement model

Training, testing, external validation



In the manuscript:

- 30.000 features from novel Omics method
- 60 samples (30 healthy, 30 ill)
- Principal component analysis
- Support vector machine for classification
- Leave-one-out cross validation
- No external validation
- Code available
- AUC: .69

Insufficient cases for number of features

Weak validation

What might have happened…

- Nobody understands the data so let’s do machine learning
- First approach (e.g. removal of correlated features, random forest): AUC .56
- Next approaches: try out other pre-processing pipelines, algorithms… (> 100 permutations …)
- Report only the best result

When not to use machine learning

1. New cases not similar enough to any of the training examples – failure to generalize
2. Similar inputs associated with different outputs
3. Defined outcomes are controversial because of an ill-defined gold standard
4. Insufficient infrastructure or resources (data scientists) for machine learning
5. Unreliable outcome labelling, lack of in-house expertise to provide training diagnoses.
6. No clear strategy or understanding of the operational context
7. Traditional rule-based software methods are equivalent/better (simple or well-characterized problem)
8. Insufficient data (quantity or quality)

Features:

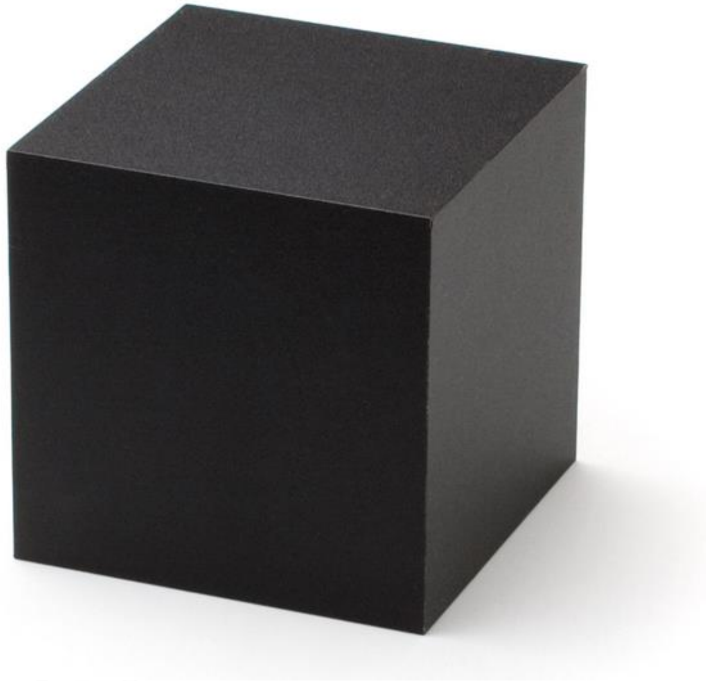
MCV

HB

Ferritin

Reticulocytes

Haptoglobin

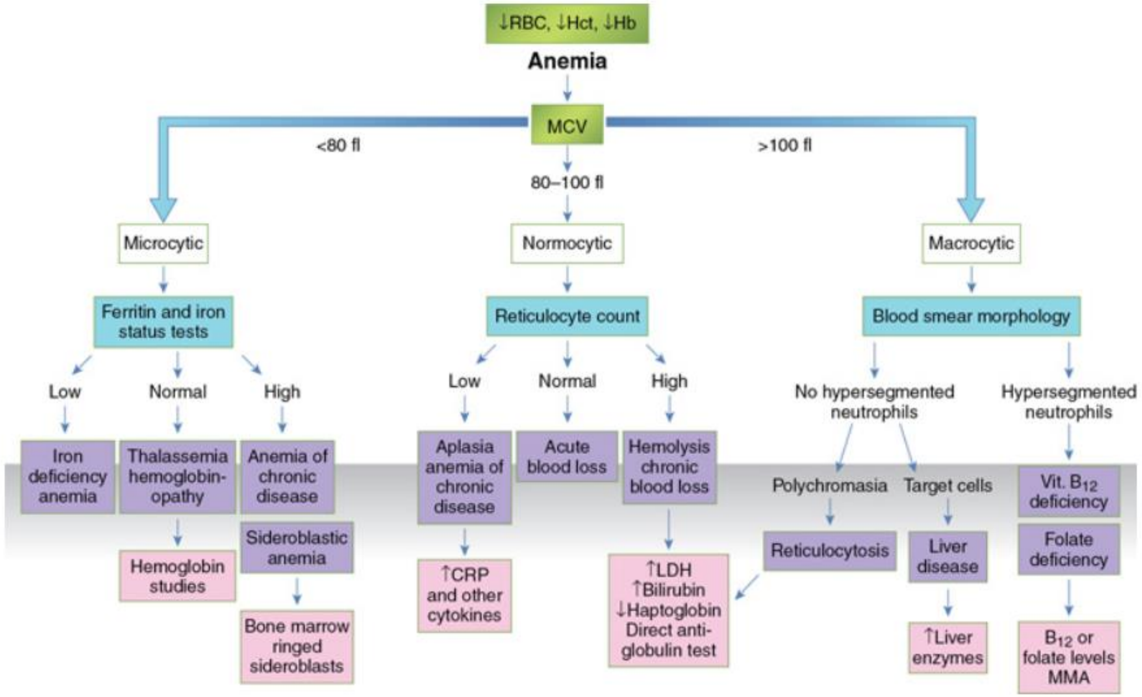


"deep neural network"

Outcome:

- Iron deficiency anemia
- Renal anemia
- Hemolytic anemia
- Other forms of anemia

Simple rules for anaemia classification

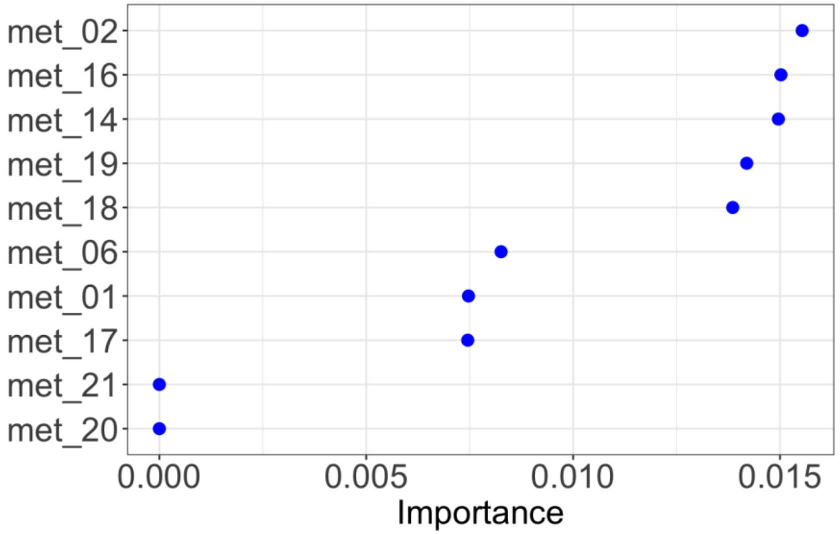


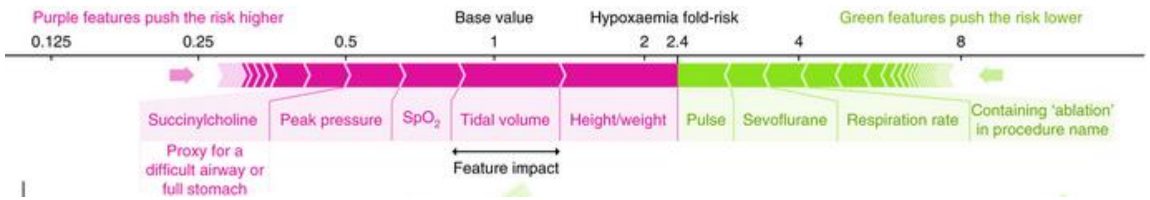
Rifai, Nader. Tietz textbook of clinical chemistry and molecular diagnostics-e-book. Elsevier Health Sciences, 2017.

When <u>not</u> to use machine learning

1. New cases not similar enough to any of the training examples – failure to generalize
2. Similar inputs associated with different outputs
3. Defined outcomes are controversial because of an ill-defined gold standard
4. Insufficient infrastructure or resources (data scientists) for machine learning
5. Unreliable outcome labelling, lack of in-house expertise to provide training diagnoses.
6. No clear strategy or understanding of the operational context
7. Traditional rule-based software methods are equivalent/better (simple or well-characterized problem)
8. Insufficient data (quantity or quality)

Feature importance analysis

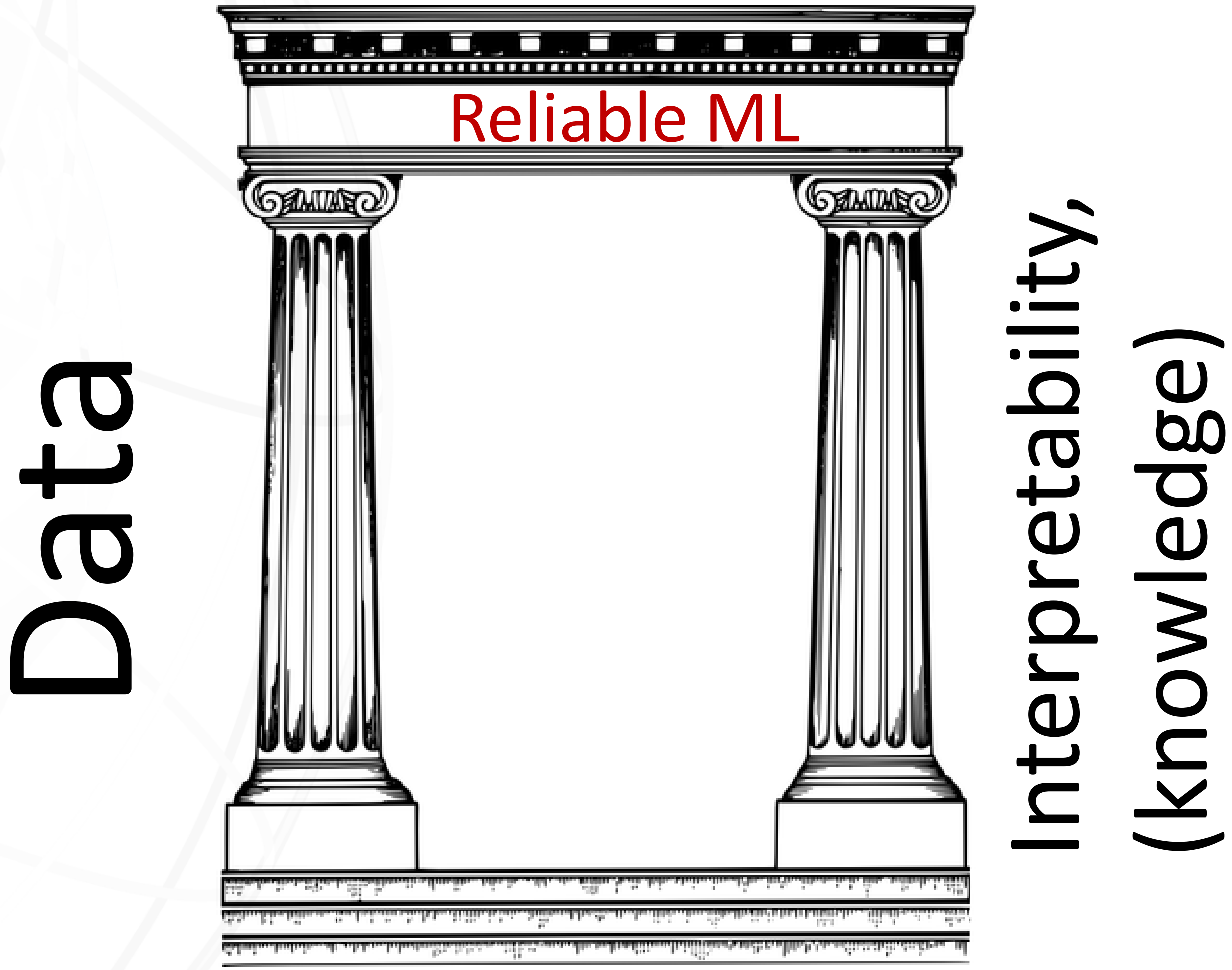




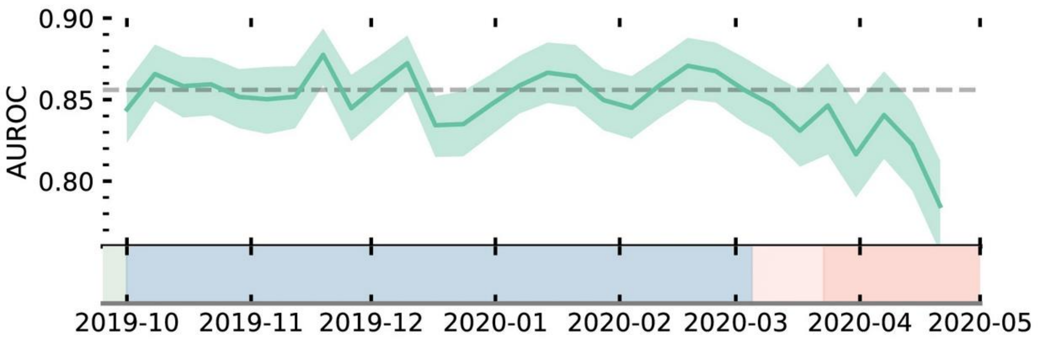
Lundberg, Scott M., et al. "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery." Nature biomedical engineering 2.10 (2018): 749-760.

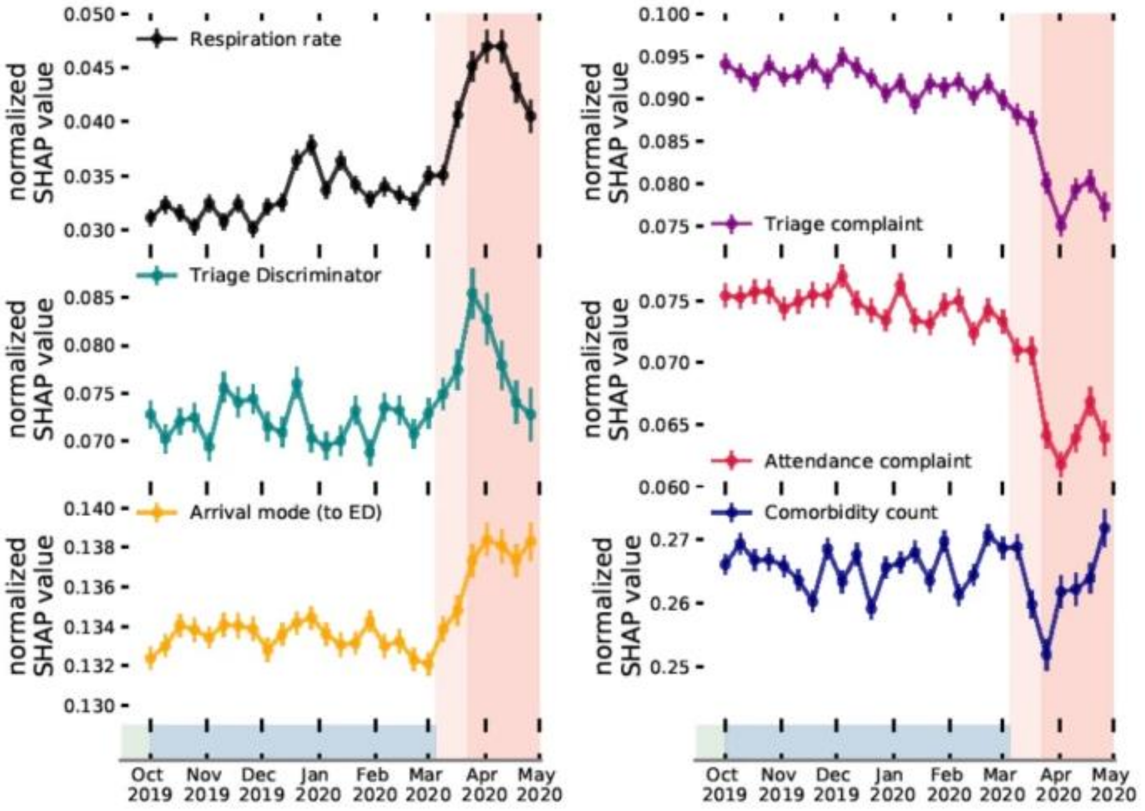
Recommendation 13: Interpret the results and performance of the selected model using suitable global and/or local interpretability methods. Address performance and potential harms in relevant subgroups and clinical scenarios.

Reliable ML models with reproducible results



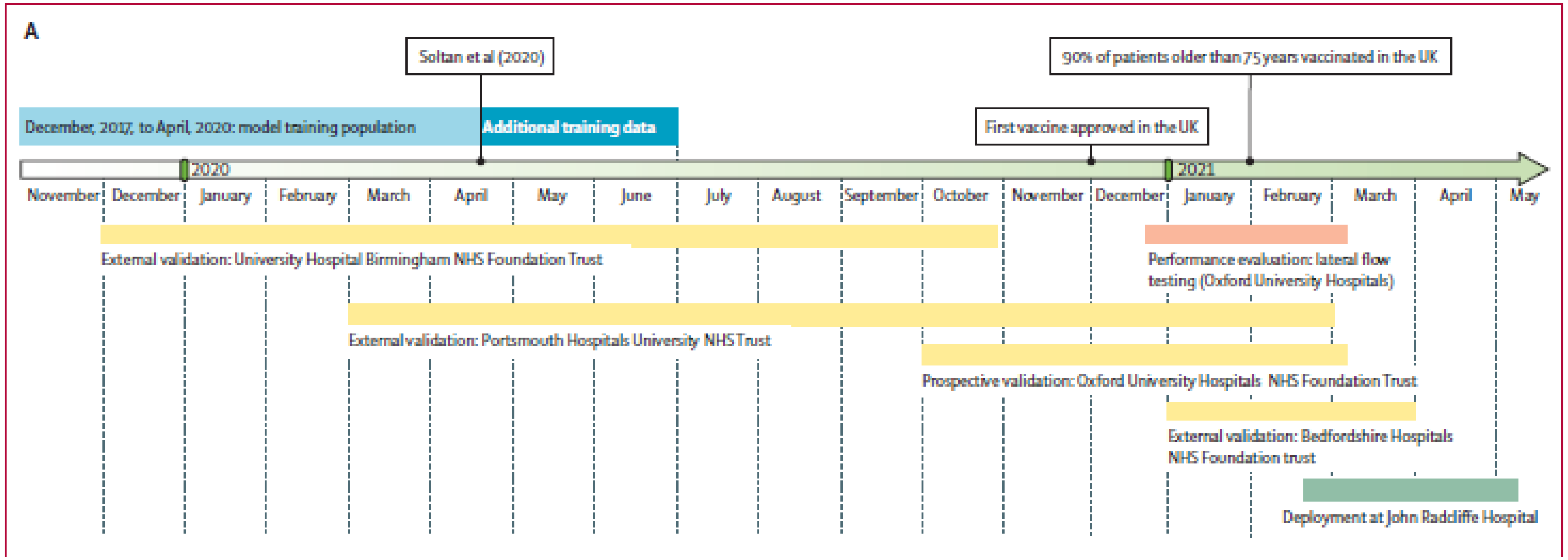
Clinical settings are dynamic environments

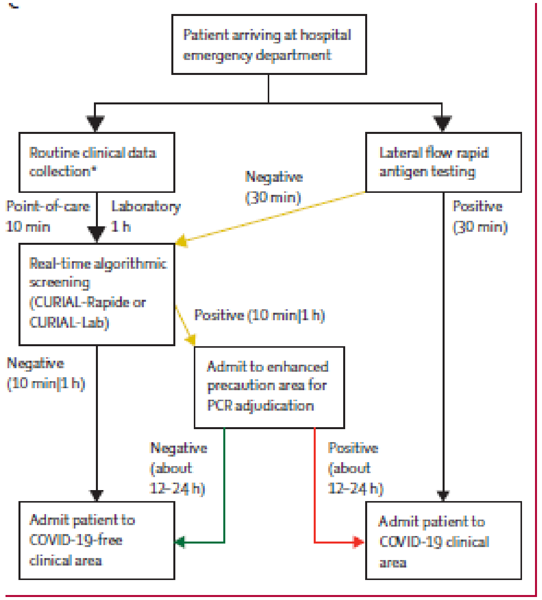




Duckworth, Christopher, et al. "Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during COVID-19." Scientific reports 11.1 (2021): 23017.

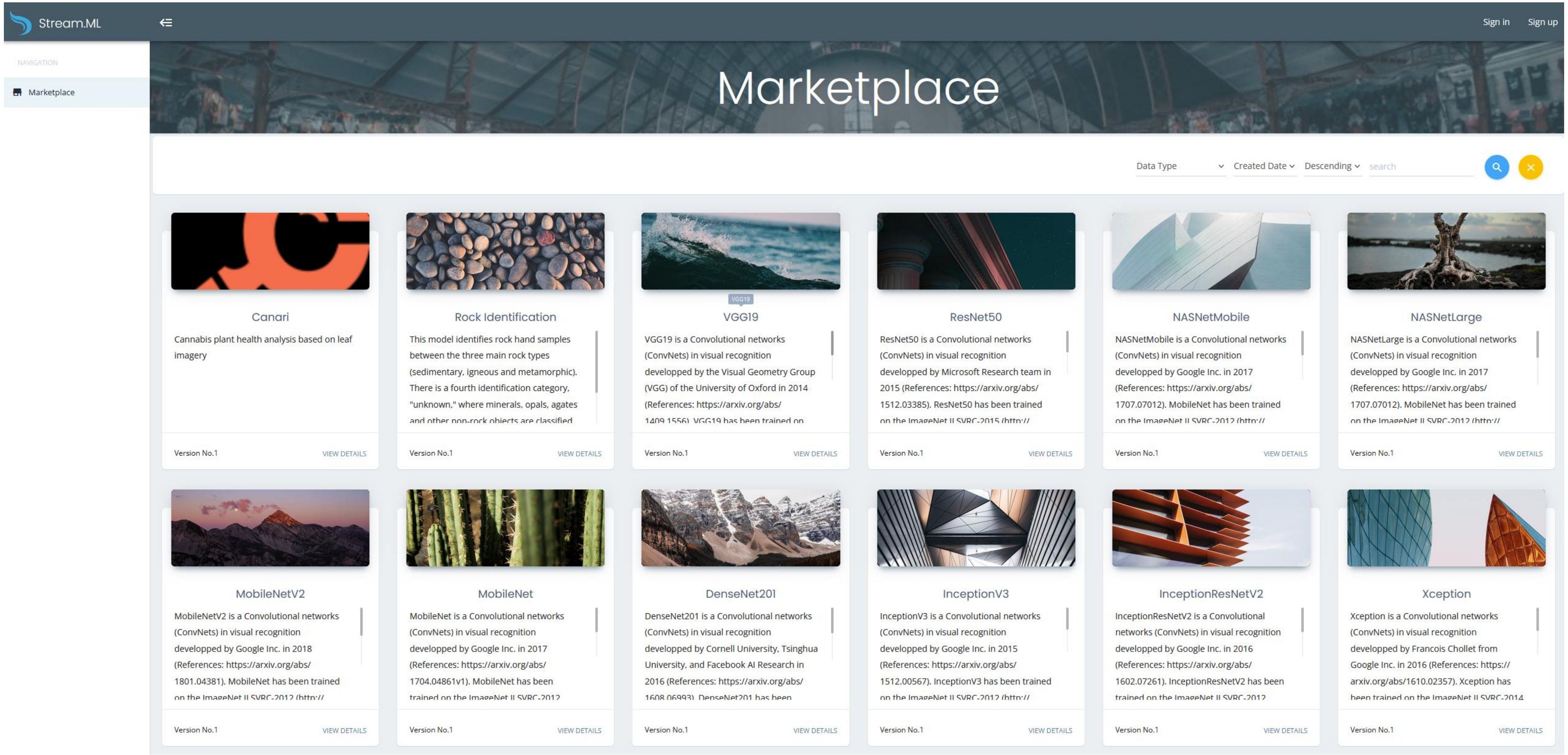
Implementation adjusted to clinical settings



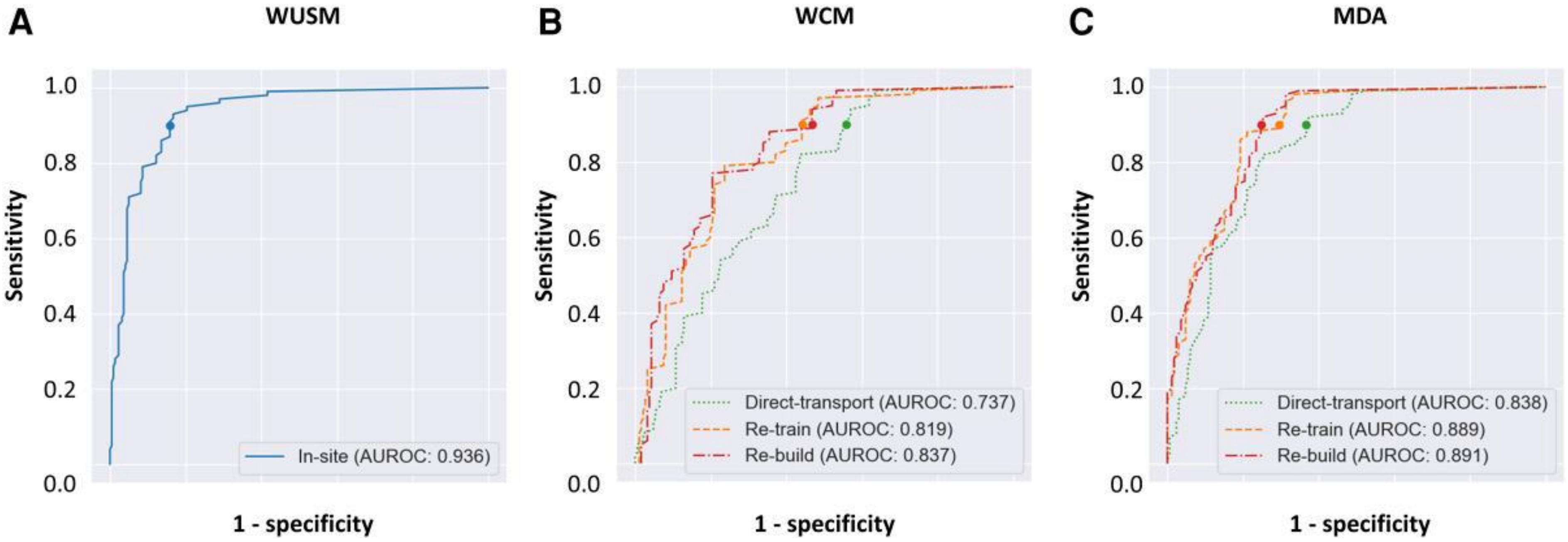


Soltan, Andrew AS, et al. The Lancet Digital Health 2022

Can you buy a ML model for Laboratory Medicine?



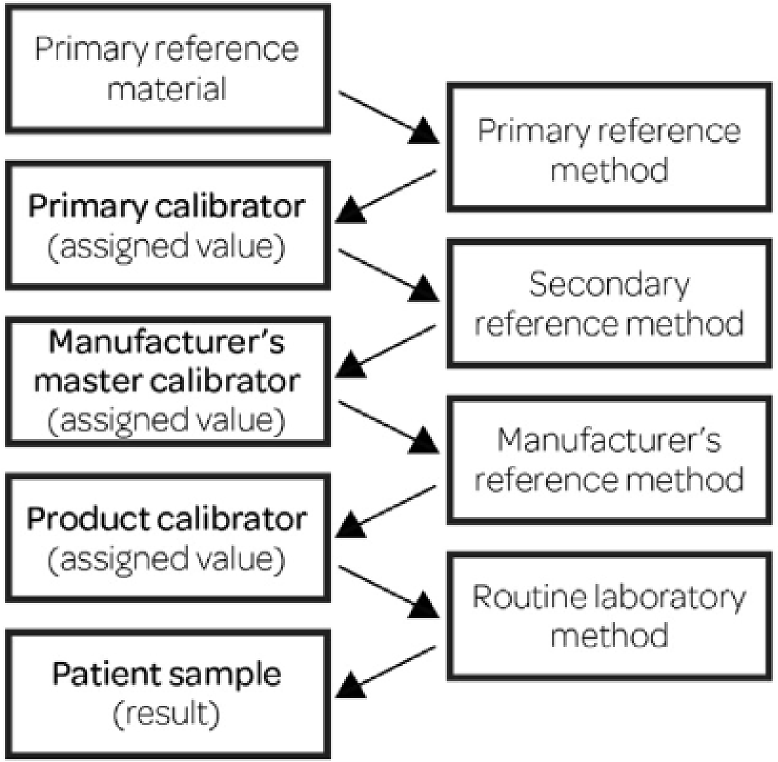
Generalizability of a Machine Learning Models is lacking

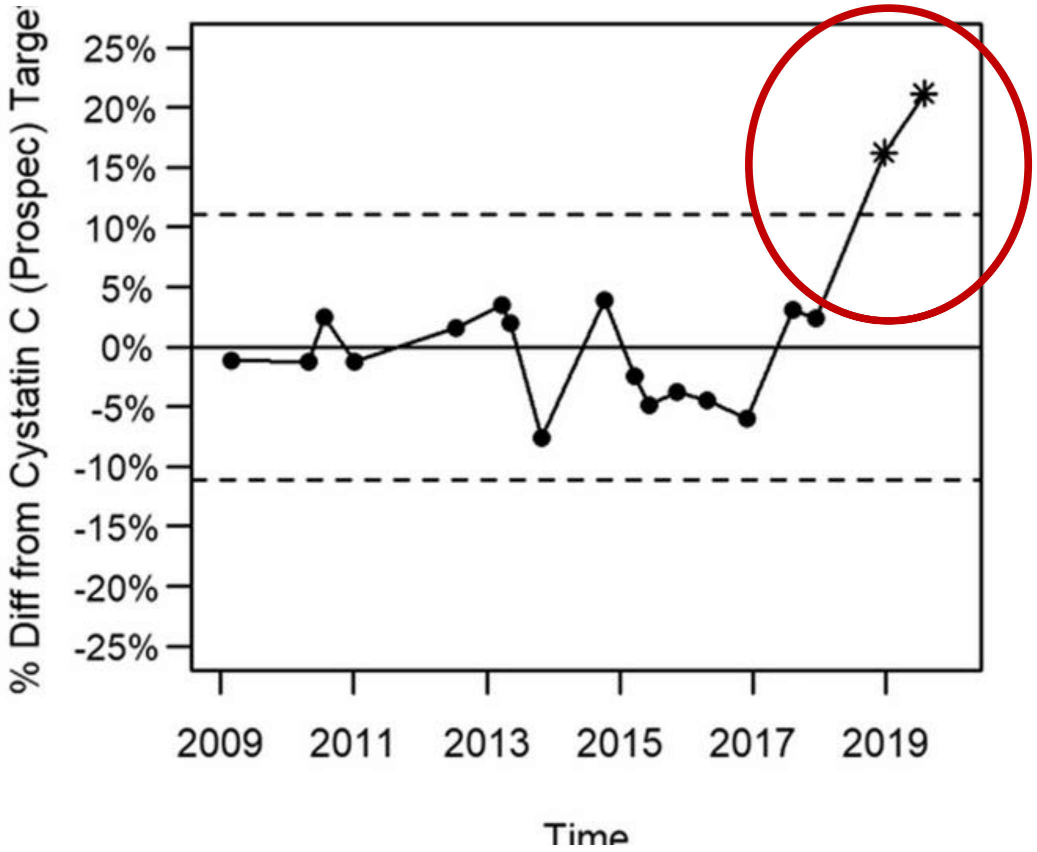


“This difference could be partially attributed to the fact that both WUSM and MDA laboratories use the same analyzers to conduct routine chemistry tests […]”

Yang, He S., et al. "Generalizability of a Machine Learning Model for Improving Utilization of Parathyroid Hormone-Related Peptide Testing across Multiple Clinical Centers." Clinical chemistry 69.11 (2023): 1260-1269.

Traceable and stable measurements prevent model deterioration





Karger, Amy B., et al. "Long-term longitudinal stability of kidney filtration marker measurements: Implications for epidemiological studies and clinical care." Clinical chemistry 67.2 (2021): 425-433.

- Keep it simple: Only use Machine Learning when you have to.
- Beware of data leakage!
- Machine Learning is no “magic bullet”: Insufficient data (quantity and quality) cannot lead to convincing results.
- Play to your strengths: Use interpretability methods to evaluate machine learning models.
- Only stable, traceable measurements can guarantee stable, transferable ML models.

Resources

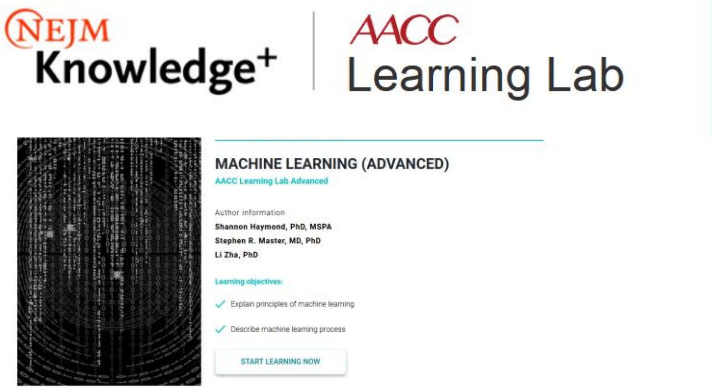
Clinical Chemistry 69:7
690–698 (2023)

Special Report

Machine Learning in Laboratory Medicine: Recommendations of the IFCC Working Group

Stephen R. Master,[a,b,*] Tony C. Badrick,[c] Andreas Bietenbeck,[d] and Shannon Haymond[e,f,*]

- https://area9lyceum.com/laboratorymedicine/
- IFCC Webinar part 2
- lab@bietenbeck.net



For further information, visit
www.ifcc.org | eacademy.ifcc.org

eAcademy

International Federation
of Clinical Chemistry
and Laboratory Medicine