Digital Tools in Clinical Trials – Regulatory Aspects

Cancer Drug Development Forum (CDDF) Workshop
27-28 September 2021

CDDF
Cancer Drug Development Forum

Ib Alstrup, Medicines Inspector GxP IT, Danish Medicines Agency

LÆGEMIDDELSTYRELSEN
DANISH MEDICINES AGENCY

Inspection of Digital Tools used in Clinical Trials

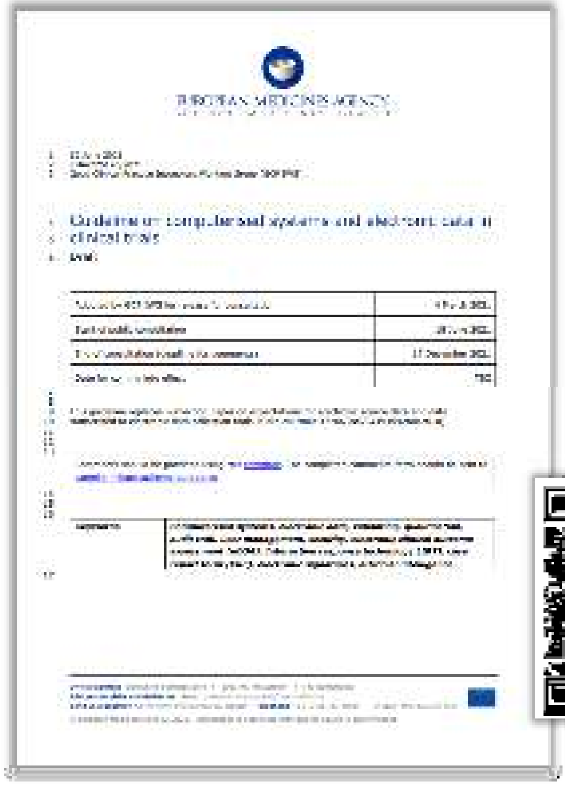
Examples of common focus areas

- Validation of specified (required) functionality
- Management of access rights (incl privileged accesses)
- Authentication method
- Electronic signatures
- Changes of data
- Edit checks and notifications
- Audit trail functionality
- Data buffering and acquisition
- Backup procedures
- IT security
- BYOD

Source: https://executiveeducation.hms.harvard.edu/industry-insights/building-better-digital-medicine-tools

New EMA Guidance on Computerised Systems in GCP

Public consultation from June to December, 2021



EUROPEAN MEDICINES AGENCY

Guideline on computerised systems and electronic data in clinical trials

Draft

- Drafted by the E-subgroup at the EMA GCP Inspectors Working Group (IWG)
- Will replace the Reflection paper on expectations for electronic source data… (2010)
- Vastly improves on the clarity of regulatory expectations to procedures and practices for validation and safe operation of systems used in clinical trials (compared to ICH GCP)

https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/draft-guideline-computerised-systems-electronic-data-clinical-trials_en.pdf

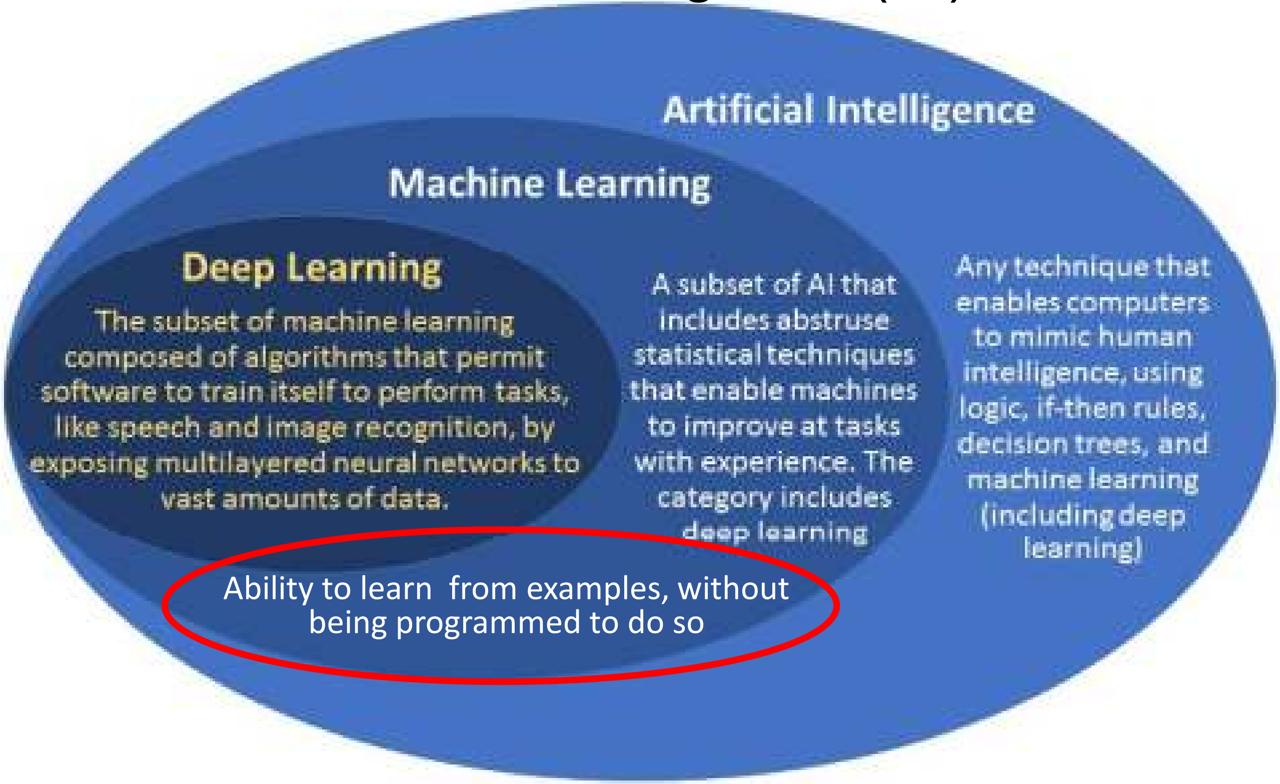
Digital Tools used in Clinical Trials

Recent examples including AI/ML

- Activity tracker
- MAA collects clinical endpoints
- Night-time sleep and scratch
- Static AI/ML algorithm
- Trained by supervised learning
- No human in the loop

A Few Terms

Related to Artificial Intelligence (AI) and Machine Learning (ML)



Artificial Intelligence

Machine Learning

Deep Learning

The subset of machine learning composed of algorithms that permit software to train itself to perform tasks, like speech and image recognition, by exposing multilayered neural networks to vast amounts of data.

A subset of AI that includes abstruse statistical techniques that enable machines to improve at tasks with experience. The category includes deep learning

Any technique that enables computers to mimic human intelligence, using logic, if-then rules, decision trees, and machine learning (including deep learning)

Ability to learn from examples, without being programmed to do so

Supervised learning (vs unsup.)

- A subcategory of ML using labeled datasets to train algorithms to classify data or predict outcomes

Static system (vs dynamic)

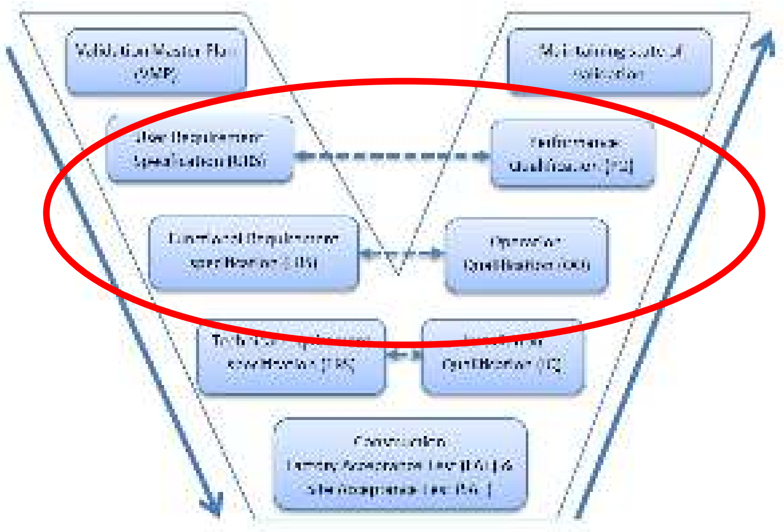
- Algorithm taught to specific performance level, then used as such

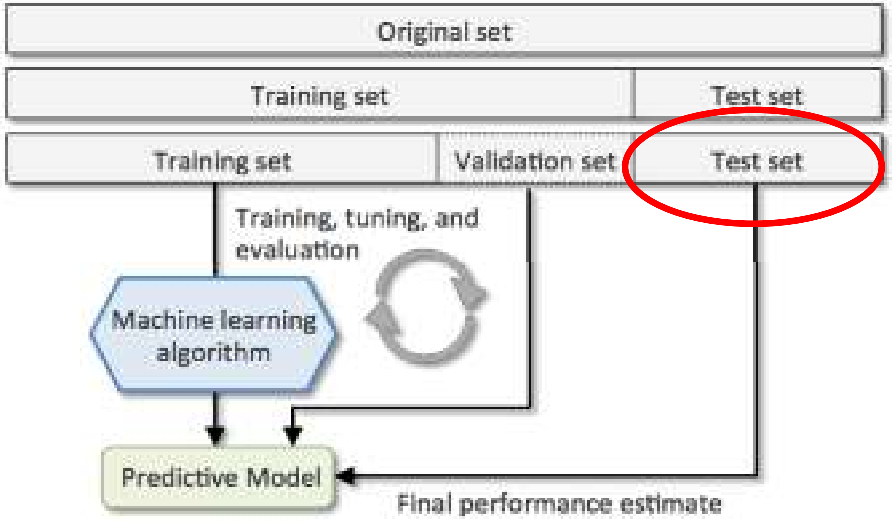
Human in the loop (vs none)

- Algorithm providing an input to a manual decision

Suggested Focus when inspecting AI/ML Systems

Traditional software vs AI/ML


Validation Master Plan (VMP)
Maintaining state of validation
User Requirement Specification (URS)
Performance Qualification (PQ)
Functional Requirement Specification (FRS)
Operation Qualification (OQ)
Technical Requirement Specification (TRS)
Installation Qualification (IQ)
Construction
Factory Acceptance Test (FAT) &
Site Acceptance Test (SAT)


Original set
Training set
Test set
Training set
Validation set
Test set
Training, tuning, and evaluation
Machine learning algorithm
Predictive Model
Final performance estimate

- Main focus should be on the test and test data used to document AI/ML performance
- Rather than on the design, training and validation (optimization) of the algorithm
- In essence, a black-box approach

Questions to critical GxP AI/ML applications (v. 0.9.4)

Posted on DKMA website and LinkedIn (March 2021)

Questions to critical GxP AI/ML applications

LÆGEMIDDELSTYRELSEN

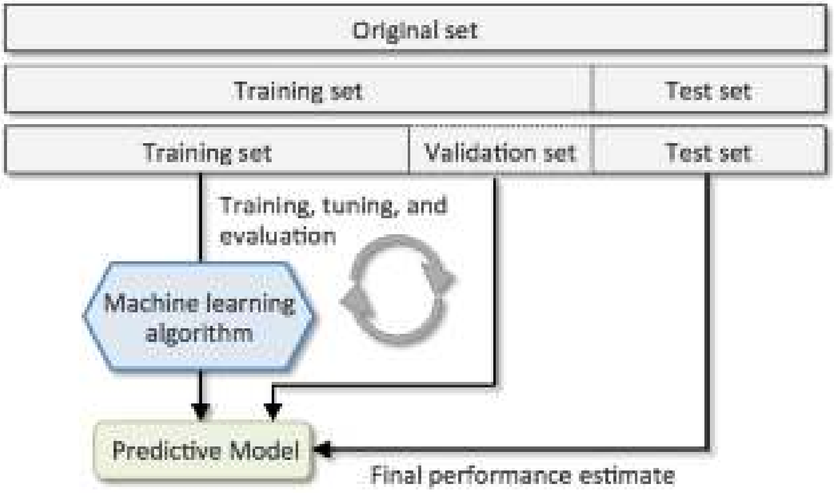
Suggested criteria for using AI/ML algorithms in GxP

Very good feedback with comments from pharma companies, organisations and individuals covering 15 or the top-20 pharma companies

Questions to critical GxP AI/ML applications (v. 0.9.5)

Datasets

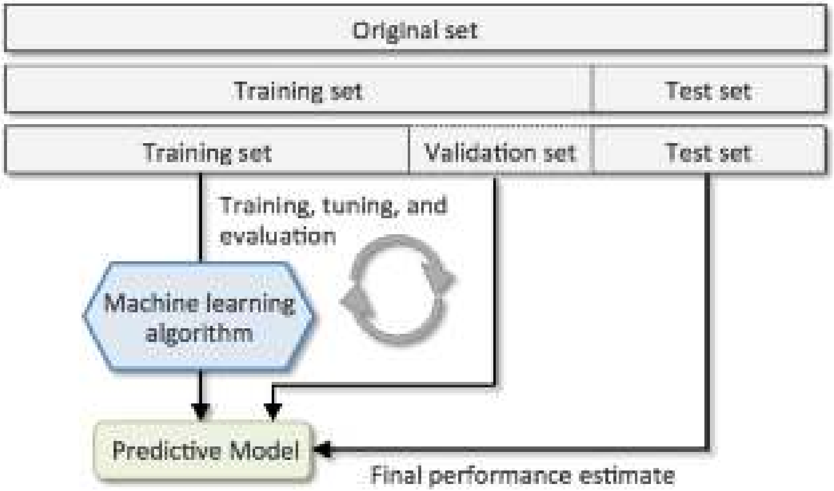
1. What was the process for training, validation (optimizing) and testing of the algorithm, who was involved in the different phases and which deviations were made from plan, if any?
2. How large are the datasets used to train, validate and test the algorithm, how is each individual data element within each group identified (named) and where are the datasets stored?
3. What evidence exists that no part of the dataset used to test the algorithm has previously been used to train or validate (optimize) the same algorithm, or originates from the same subject, unless features are independent?


Original set
Training set
Test set
Training set
Validation set
Test set
Training, tuning, and evaluation
Machine learning algorithm
Predictive Model
Final performance estimate

Questions to critical GxP AI/ML applications (v. 0.9.5)

Datasets

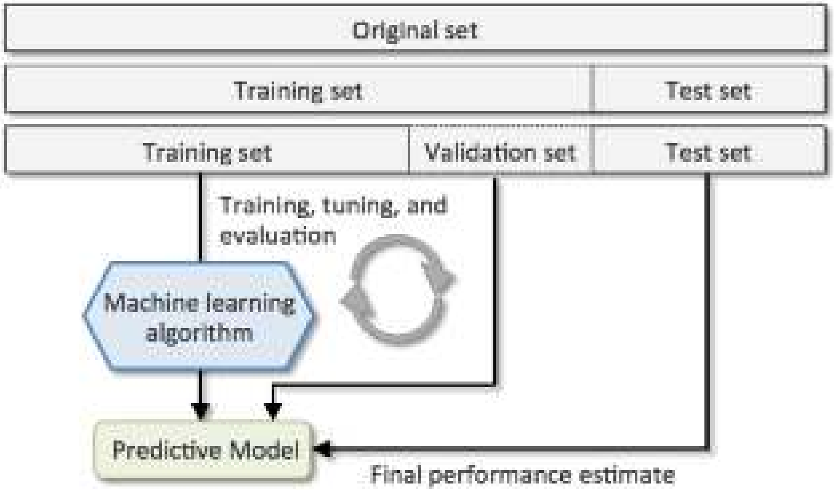
4. When were the data points used to test the algorithm separated from the complete pool of data points and what selection criteria were used?
5. What kind of data cleaning, normalization, homogenization, exclusion criteria, data synthetization or similar were the test data subject to and why?
6. How was it ensured that the test dataset is representative of real data from the intended scope of the application and throughout the intended lifecycle of the application?


Original set
Training set
Test set
Training set
Validation set
Test set
Training, tuning, and evaluation
Machine learning algorithm
Predictive Model
Final performance estimate

Questions to critical GxP AI/ML applications (v. 0.9.5)

Datasets

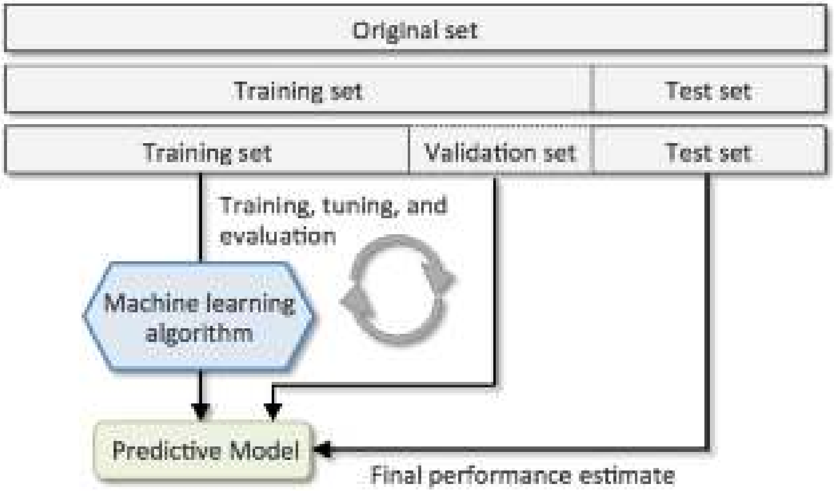
7. How was it ensured that the test dataset contains enough challenging data, e.g. that an algorithm which has been trained to distinguish between dogs and wolves, has not only been tested with dog pictures of dachshunds and poodles, but also with German shepherds and Siberian huskies?
8. What features in the training dataset have the highest effect on the output of the algorithm, how has that influenced the selection of, and how does it (if available) correspond to the test dataset?


Original set
Training set
Test set
Training set
Validation set
Test set
Training, tuning, and evaluation
Machine learning algorithm
Predictive Model
Final performance estimate

Questions to critical GxP AI/ML applications (v. 0.9.5)

Datasets

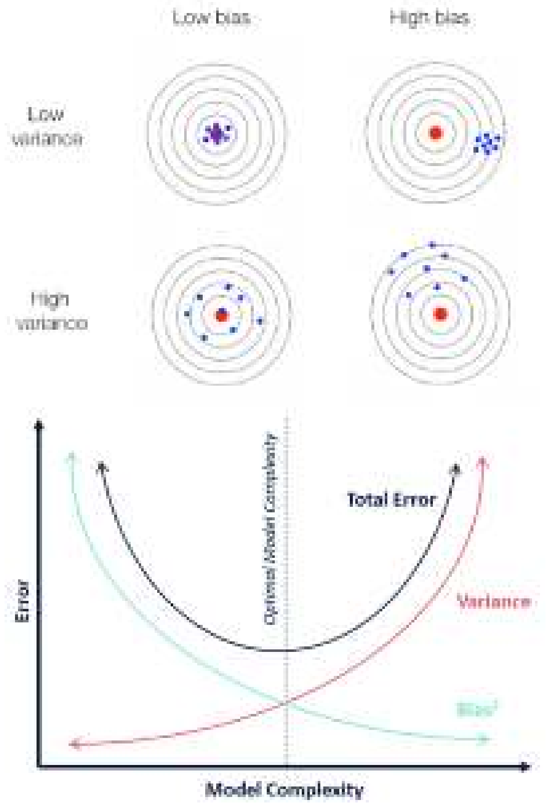
9. How was it ensured that the test dataset covers any technical differences (e.g. formatting) which may arise in real data due to differences in personnel, processes and equipment?
10. How was the correct annotation of the data used to test the algorithm verified, and has the annotation (where applicable) e.g. been verified by a second person or by laboratory tests?


Original set
Training set
Test set
Training set
Validation set
Test set
Training, tuning, and evaluation
Machine learning algorithm
Predictive Model
Final performance estimate

Questions to critical GxP AI/ML applications (v. 0.9.5)

Bias and Variance

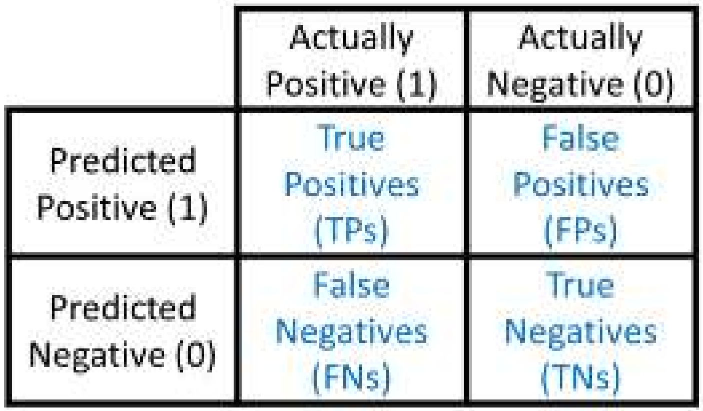
11. How was the algorithm optimized to deal with bias and variance (bias – variance tradeoff), what is the result of this optimization as seen in the test and how do bias – variance tradeoff graphs look?
    a. Bias is an error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).
    b. Variance is an error from sensitivity to small fluctuations in the training set. High variance can cause an algorithm to model the random noise in the training data, rather than the intended outputs (overfitting).



Questions to critical GxP AI/ML Applications (v. 0.9.5)

Confusion Matrix and Metrics

13. What are the values in the confusion matrix (TP/FP/FN/TN) and the following metrics?
    a. Sensitivity [TP/(TP+FN)] is the number of items correctly identified as positive out of total true positives. Sensitivity is also called Recall, Hit Rate or True Positive Rate (TPR)
    b. Specificity [TN/(TN+FP)] is the number of items correctly identified as negative out of total true negatives. Specificity is also called Selectivity or True Negative Rate (TNR)
    c. Precision [TP/(TP+FP)] is the number of items correctly identified as positive out of the total identified as positive. Precision is also called Positive Predictive Value (PPV).



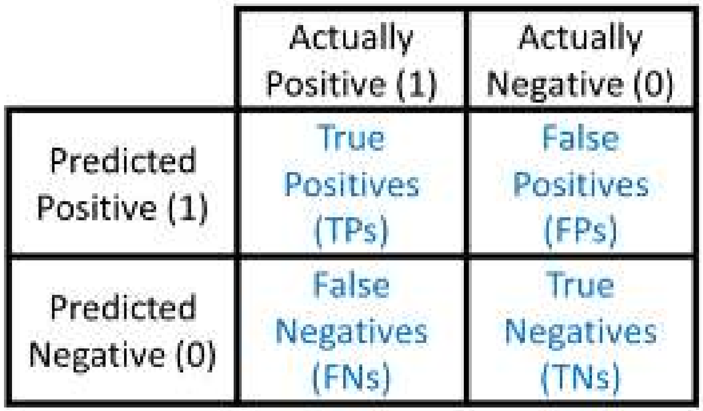
	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Questions to critical GxP AI/ML Applications (v. 0.9.5)

Confusion Matrix and Metrics

d. False Positive Rate [FP/(TN+FP)] is the number of items wrongly identified as positive out of total true negatives. E.g. a man being declared as pregnant.
   Is also called Type I error.

e. False Negative Rate [FN/(TP+FN)] is the number of items wrongly identified as negative out of total true positives. E.g. pregnant woman declared as not pregnant.
   Is also called Type II error.



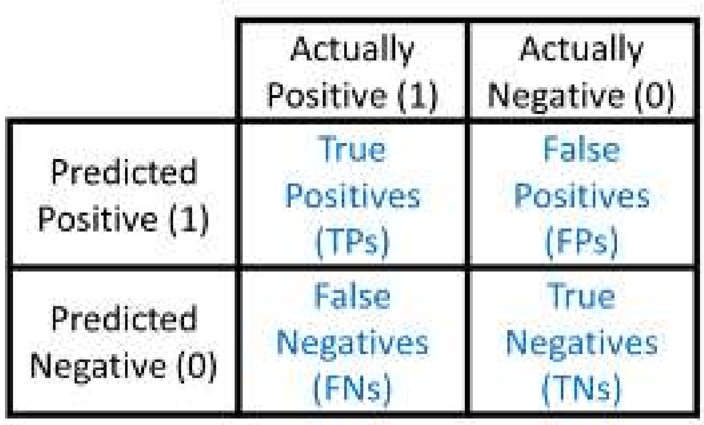
	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Questions to critical GxP AI/ML Applications (v. 0.9.5)

Confusion Matrix and Metrics

f. Accuracy [(TP+TN)/(N+P)] is the percentage of total items classified correctly. Should not be used with uneven sets of classes, as accuracy of one class can overpower the other.

g. F1 Score [2*(Precision*Sensitivity)/(Precision+Sensitivity)] is a harmonic mean of Precision and Sensitivity. The F1 Score has the advantage over accuracy that with uneven classes it gives a better metric to calculate the model performance.



	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Questions to critical GxP AI/ML Applications (v. 0.9.5)

Interpretation of Results

14. Which of the quadrants of the confusion matrix and which of the metrics are more important for the intended scope of the application and if low scores are seen, why may these be less important?
15. How was the intended scope of the application defined and limited based on both test data and test results including the confusion matrix and metrics?
16. How was the threshold defined for end results, e.g. is an outcome of ’50.01% it is a dog’ interpreted to the result ‘it is a dog’, and when, if ever, would an outcome require human interaction?

Thanks for your attention

For questions:

Ib Alstrup, Medicines Inspector GxP IT, Danish Medicines Agency
ibal@dkma.dk, www.linkedin.com/in/ib-alstrup-baa2542

LÆGEMIDDELSTYRELSEN
DANISH MEDICINES AGENCY