tetris

Risk assessment *Tools* for severe side *Effects* after *breas*T *R*adiotherapy: radiation safety through biological extended models and d*I*gital twin*S*

Models to estimate the risk of normal-tissue complications after radiotherapy

Filippo Schiavo (postdoc), Karolinska University Hospital, Stockholm, Sweden

TETRIS webinar, 22 January 2026

This project has received funding under Euratom Research and Training Programme (EURATOM) Grant ID 101166699

EU Grant Agreement n. 101166699

Karolinska Institutet

KAROLINSKA UNIVERSITETSSJUKHUSET

Karolinska Comprehensive Cancer Center

NTCP model selection strategy

The risk scores to be communicated to the breast cancer patients will derive from the NTCP models. These may be:

1) Selected from the literature and tested in the TETRIS cohort, or
2) Developed in-house using a meta-analytic approach

for example, from

$$NTCP_{dyspnoea} = \frac{1}{1 + e^{4\gamma_{50}\left(1-\frac{MLD}{D_{50}}\right)}}$$

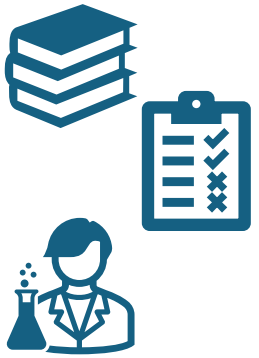
to

$$NTCP_{dyspnoea} = \frac{1}{1 + OR_1 OR_2 \dots OR_k e^{4\gamma_{50}^{0}\left(1-\frac{MLD}{D_{50}^{0}}\right)}}$$

Odds-ratios for various Predictors retrieved from the literatue

Selection of models from the literature

1) Comprehensive review
2) Qualitative assessment
3) Comparative analysis on the TETRIS cohort
    - Same performance metrics
    - Unified performance metrics
4) Assessment of the effect of clinical predictors across models
5) Model testing and comparison on individual patients



$$NTCP = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots)}}$$

β_0 = intercept
$\beta_{1\text{-}n}$ = coefficients (log of odds ratios)
$x_{1\text{-}n}$ = predictors

Comprehensive review & Qualitative assessment

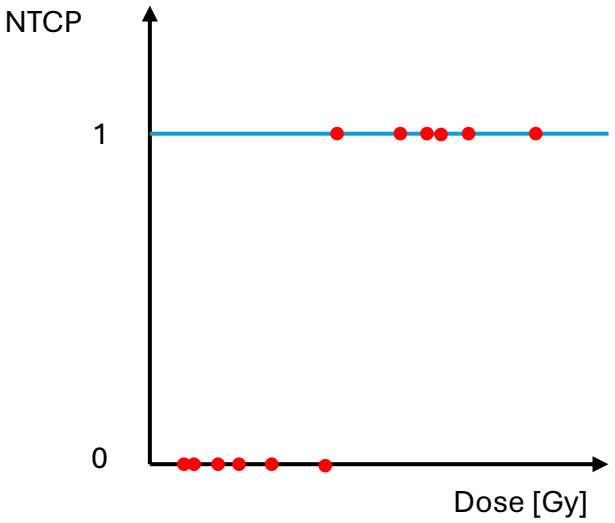
For each clinical endpoint a ranked list of models will be produced.

The endpoints regard organs at risk of developing adverse effects from the treatement with radiation, such as:

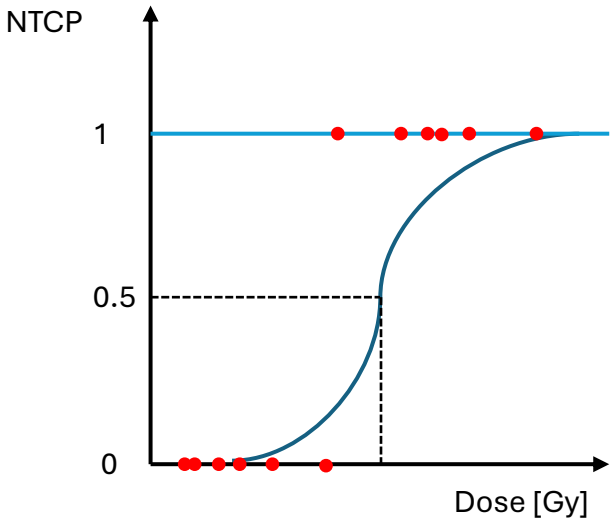
- Lung fibrosis
- Coronary artery disease
- Second primary cancers

So far 23 studies and even more NTCP models have been identified and a qualitative assessment is being performed, based on factors such as cohort sizes, type of model used, incidences, predictors, etc.

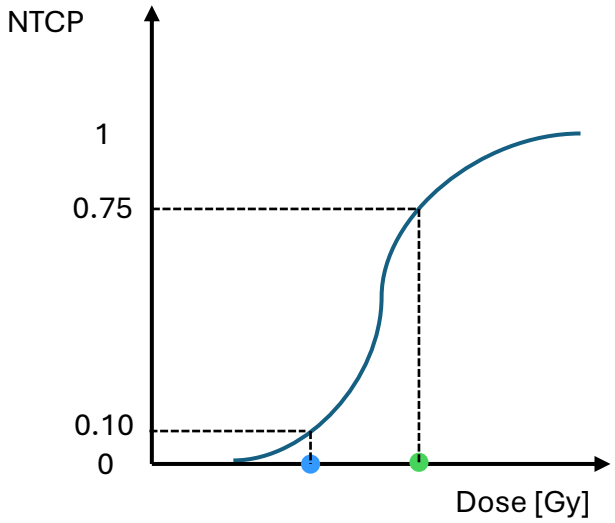
Comparative analysis of models



Observed outcomes (0,1)
Training dataset



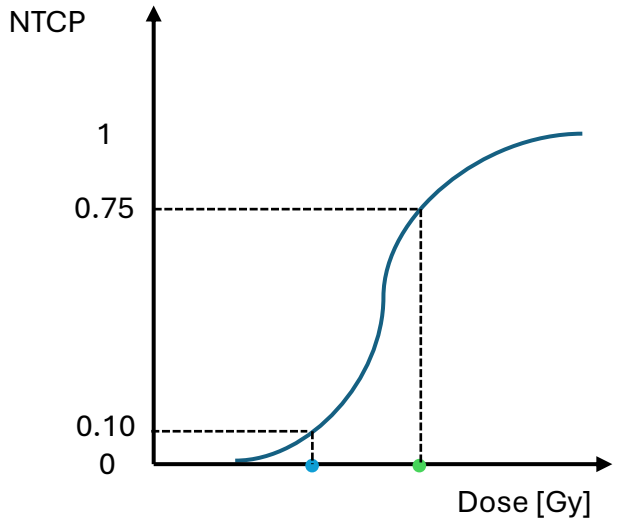
Prediction model [0,1]



Prediction on two new patients
→ pt1: 10%; pt2 : 75%
The observed events are 0 in both
(no adverse event)

How do we quantify model performance?

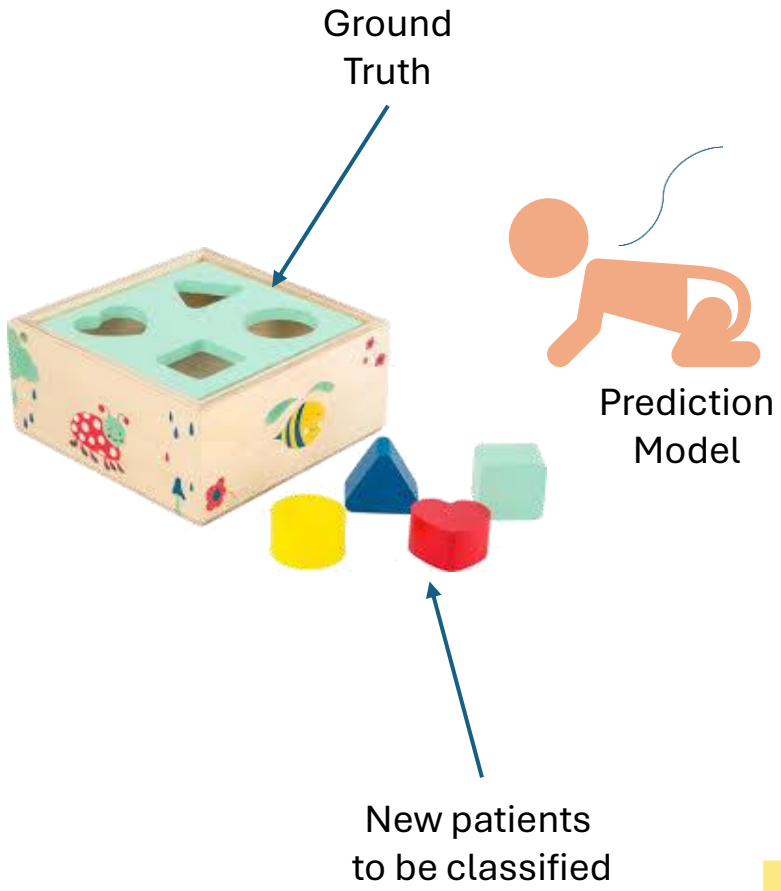
Problem



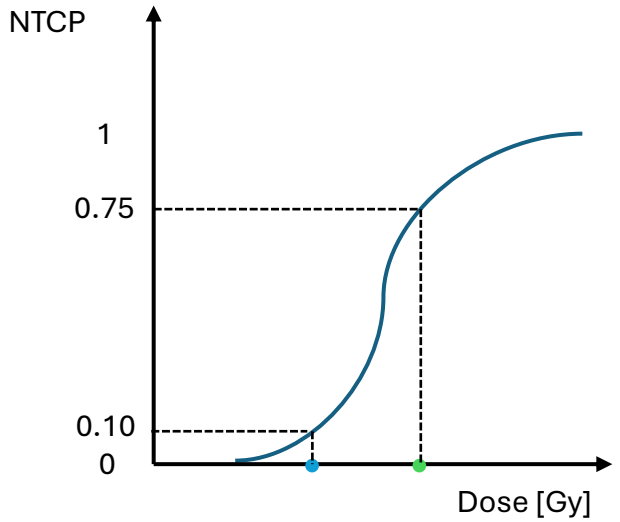
Prediction on two new patients
→ pt1: 10%; pt2 : 75%
The observed events are 0 in both
(no adverse event)

How do we quantify model performance?

- DISCRIMINATION
- CALIBRATION



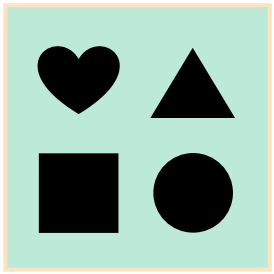
Problem



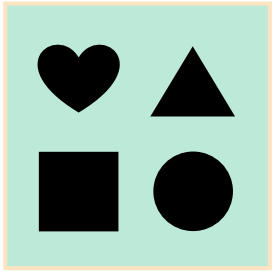
Prediction on two new patients
→ pt1: 10%; pt2 : 75%
The observed events are 0 in both
(no adverse event)

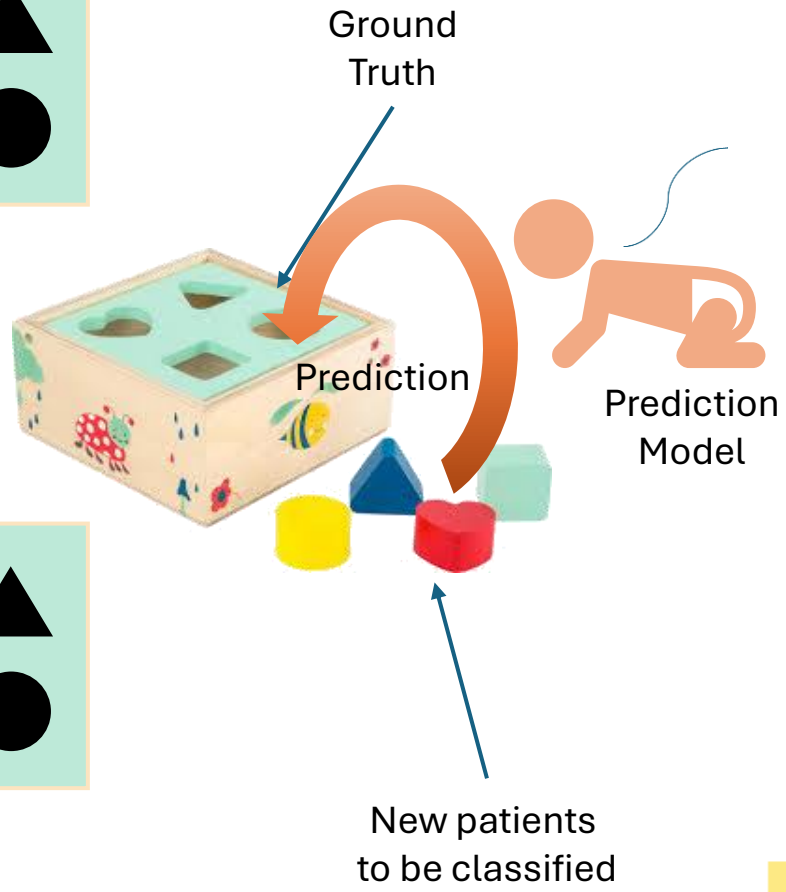
How do we quantify model performance?

DISCRIMINATION

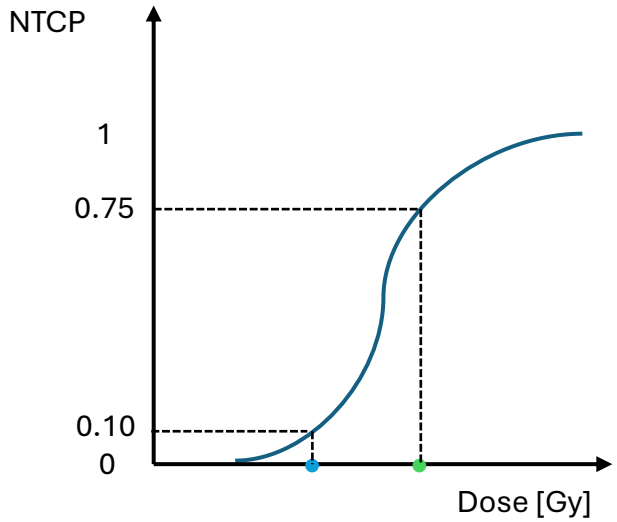


CALIBRATION





Problem

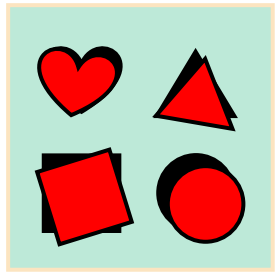


Prediction on two new patients
→ pt1: 10%; pt2 : 75%
The observed events are 0 in both
(no adverse event)

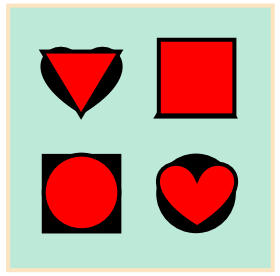
How do we quantify model performance?

DISCRIMINATION

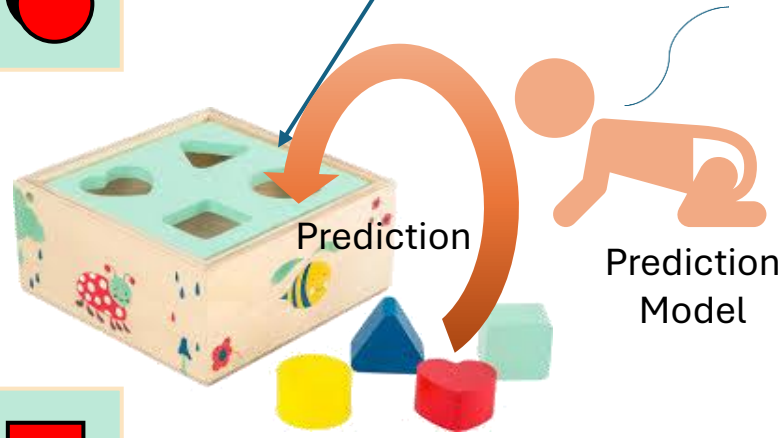
How often does the model assign a block to the correct outcome?



CALIBRATION

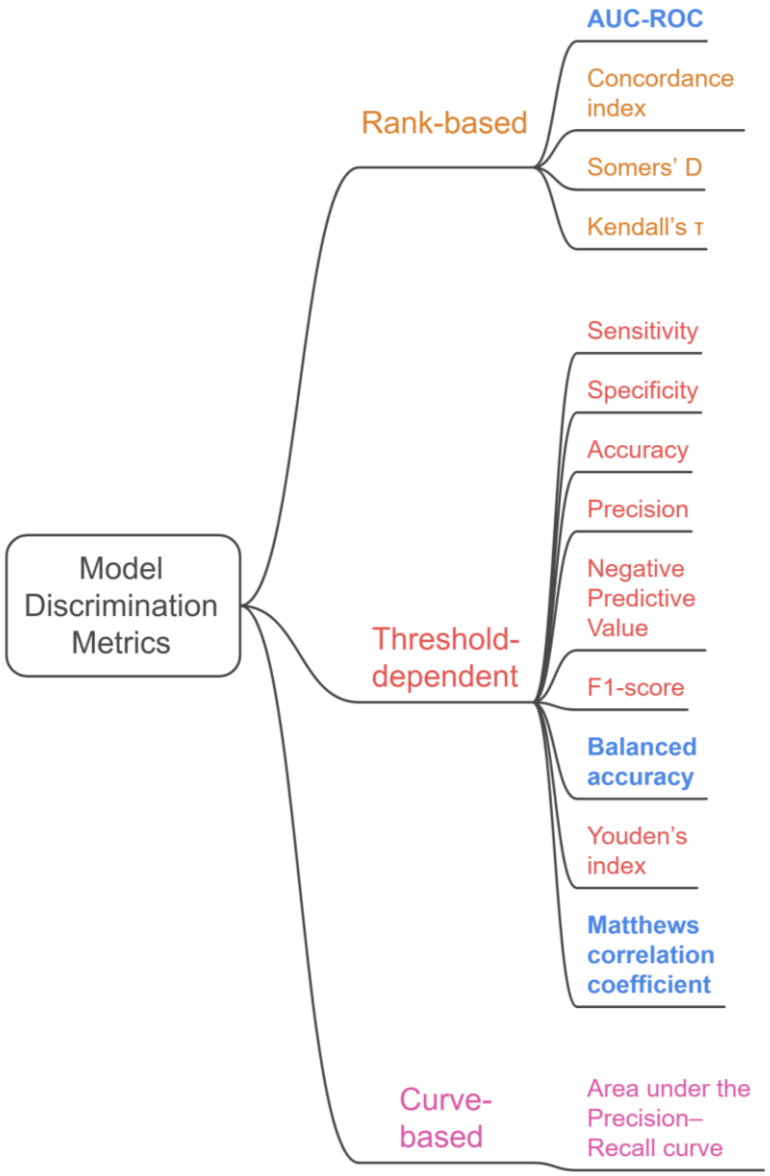


How well do the model's predictions overlap with the true outcomes?



Discrimination

Ability of the model to distinguish patients with vs without toxicity



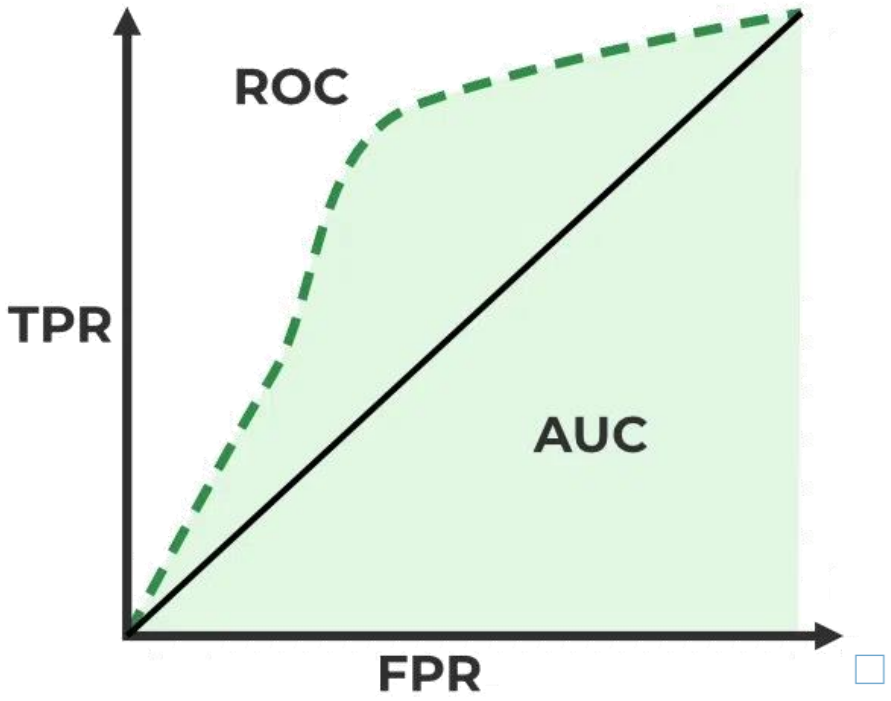
This project has received funding under Euratom Research and Training Programme (EURATOM) Grant ID 101166699

TETRIS – Risk assessment tools for severe side effects after breast radiotherapy: radiation safety through biological extended models and digital twins
EU Grant Agreement n. 101166699

Discrimination metrics: rank-based

AUC-ROC (C-statistic)

- The standard metric for goodness of fit for binary outcomes (→classification) in a logistic regression model
- Interpreted as the probability that a randomly selected patient having the adverse event had a higher risk score than a patient who had not experienced the event
- Threshold independent, as the curve represents all the possible NTCP thresholds
- Problematic in the case of rare events



Discrimination metrics: threshold-dependent

Balanced accuracy → $\frac{TPR + TNR}{2}$

- Better than accuracy for rare events
- But it doesn't penalize PPV and NPV symmetrically

Matthews correlation coefficient

- It considers all the positive and negative metrics of the confusion matrix, by rewarding TPR, TNR, PPV, NPV, while penalizing FNR, FPR, FOR, FDR
- Best for rare events

Calibration

Agreement between NTCP and observed toxicity rates

NB: In a clinical setting, calibration is preferred over discrimination, due to the

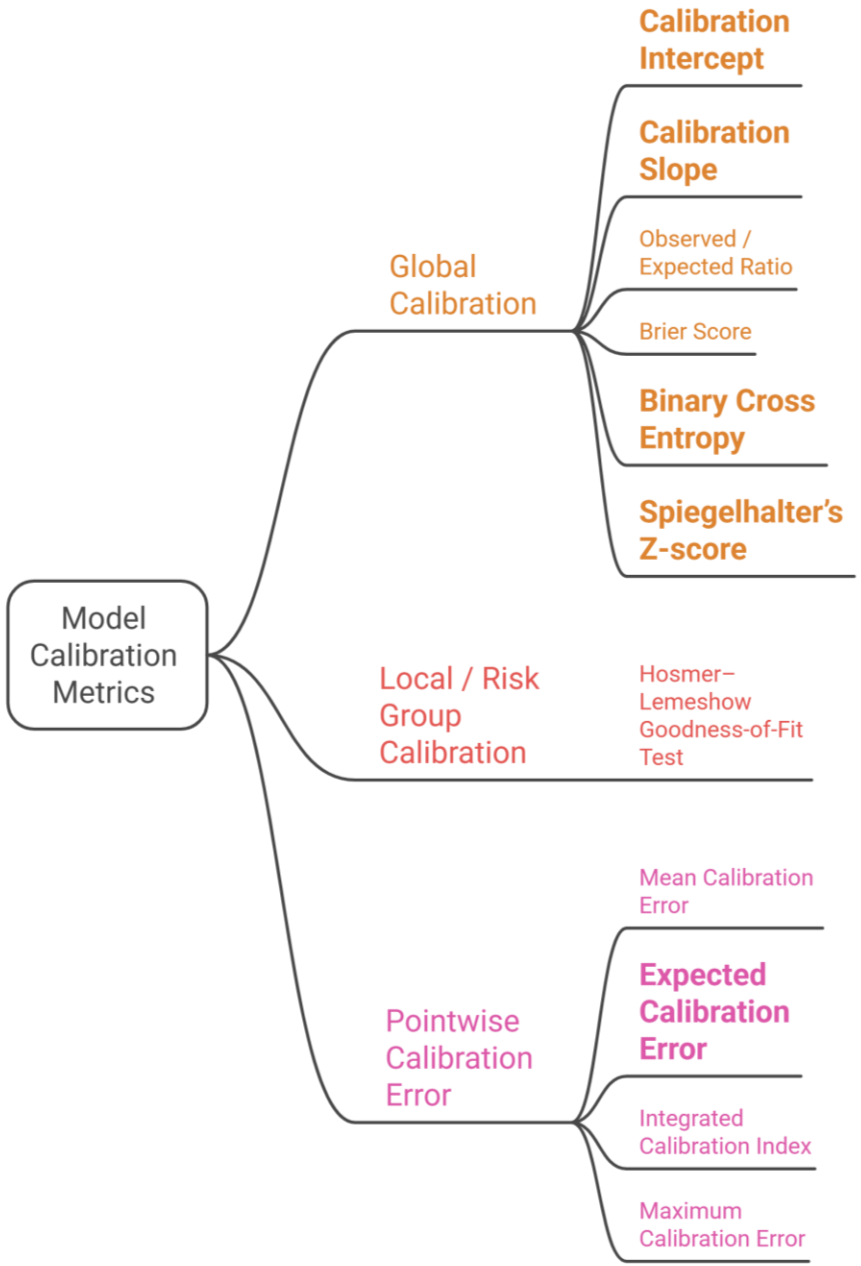
→ communication of an absolute risk to the patient

vs

→ ranking the patient's risk relatively to another

Calibration

Agreement between NTCP and observed toxicity rates

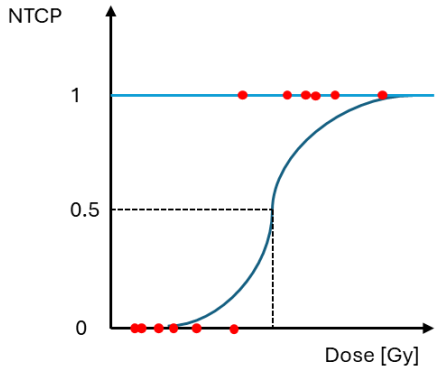


This project has received funding under Euratom Research and Training Programme (EURATOM) Grant ID 101166699

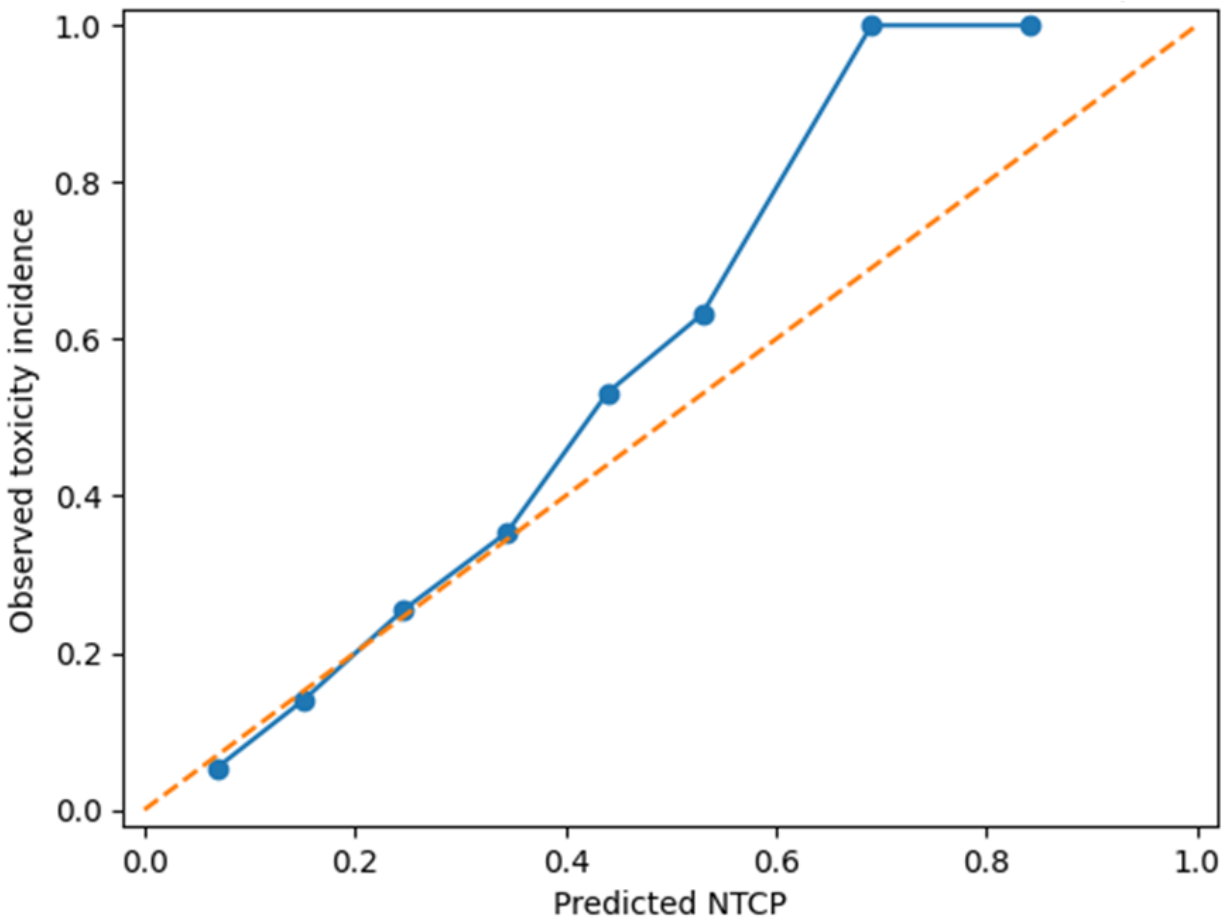
TETRIS – Risk assessment tools for severe side effects after breast radiotherapy: radiation safety through biological extended models and digital twins
EU Grant Agreement n. 101166699

Calibration metrics: global calibration

- **Calibration Intercept** ("calibration in-the-large")
- **Calibration slope**



Prediction model [0,1]



Calibration metrics: global calibration

Binary cross entropy, or Log-loss

$$BCE = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)\right]$$

- = - log(likelyhood). Lower values are more favourable
- A metric mixing calibration and discrimination
- It accounts for confidence of predictions
  (e.g. $y_i = 0$, $p_i = 0.9$: model too confident in predicting 1, wrong value
  → higher loss due to the logarithm)

Calibration metrics: global calibration

Spiegelhalter's Z-score

$$Z = \frac{\sum_{i=1}^{N}(y_i - p_i)(1-2p_i)}{\sum_{i=1}^{N}\sqrt{p_i(1-p_i)(1-2p_i)^2}}$$

Spiegelhalter. Stat Med, 1986, 5(5):421-33

- It is a normalized test statistic, providing a significance test for deviation from perfect calibration
- a Z-score close to zero paired with a p-value > 0.05 indicates an overall well-calibrated model
- It detects calibration bias
- Useful for benchmarking a model, not for model comparison or ranking

General model descriptors

Akaike Information Criterion (AIC)

$$AIC = 2K - 2\ln(L)$$

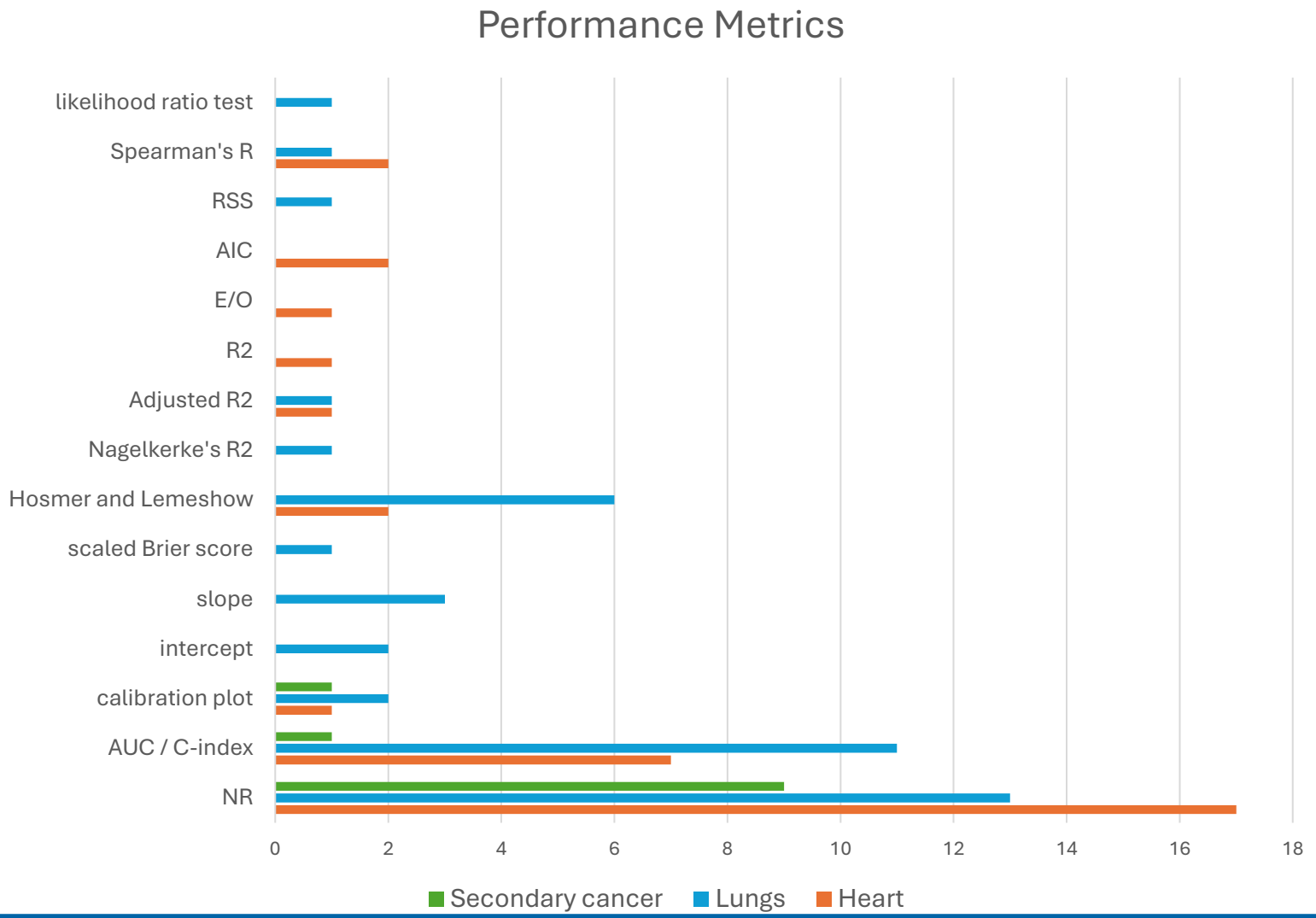
- It accounts for model complexity (K: number of parameters) and goodness of fit ($\ln(L)$: log-Likelihood)
- … giving a performance measure relative to other models
  → useful for model selection (within study)

Bayesian Information Criterion (BIC)

$$BIC = K\,ln(n) - 2\ln(L)$$

- It is similar to AIC, penalizing more heavily the number of parameters for larger cohort sizes n

NTCP Model selection in TETRIS



TETRIS – Risk assessment tools for severe side effects after breast radiotherapy: radiation safety through biological extended models and digital twins
EU Grant Agreement n. 101166699

TRIPOD (Transparent Reporting of a multivariable prediction model)

Guidelines for uniform reporting of summary scores

RESEARCH METHODS AND REPORTING

TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods

Gary S Collins,[1] Karel G M Moons,[2] Paula Dhiman,[1] Richard D Riley,[3,4] Andrew L Beam,[5] Ben Van Calster,[6,7] Marzyeh Ghassemi,[8] Xiaoxuan Liu,[9,10] Johannes B Reitsma,[2] Maarten van Smeden,[2] Anne-Laure Boulesteix,[11] Jennifer Catherine Camaradou,[12,13] Leo Anthony Celi,[14,15,16] Spiros Denaxas,[17,18] Alastair K Denniston,[4,9] Ben Glocker,[19] Robert M Golub,[20] Hugh Harvey,[21] Georg Heinze,[22] Michael M Hoffman,[23,24,25,26] André Pascal Kengne,[27] Emily Lam,[12] Naomi Lee,[28] Elizabeth W Loder,[29,30] Lena Maier-Hein,[31] Bilal A Mateen,[17,32,33] Melissa D McCradden,[34,35] Lauren Oakden-Rayner,[36] Johan Ordish,[37] Richard Parnell,[12] Sherri Rose,[38] Karandeep Singh,[39] Laure Wynants,[40] Patricia Logullo[1]

For numbered affiliations see end of the article

Correspondence to: G S Collins gary.collins@csm.ox.ac.uk (or @GSCollins on Twitter; ORCID 0000-0002-2772-2316)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2024;385:e078378 http://dx.doi.org/10.1136/bmj-2023-078378

Accepted: 17 January 2024

The TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) statement was published in 2015 to provide the minimum reporting recommendations for studies developing or evaluating the performance of a prediction model. Methodological advances in the field of prediction have since included the widespread use of artificial intelligence (AI) powered by machine learning methods to develop prediction models. An update to the TRIPOD statement is thus needed. TRIPOD+AI provides harmonised guidance for reporting prediction model studies, irrespective of whether regression modelling or machine learning methods have been used. The new checklist supersedes the TRIPOD 2015 checklist, which should no longer be used. This article describes the development of TRIPOD+AI and presents the expanded 27 item checklist with more detailed explanation of each reporting recommendation, and the TRIPOD+AI for Abstracts checklist. TRIPOD+AI aims to promote the complete, accurate, and transparent reporting of studies that develop a prediction model or evaluate its performance. Complete reporting will facilitate study appraisal, model evaluation, and model implementation.

SUMMARY POINTS

There has been considerable interest and financial investment in developing prediction models by applying artificial intelligence (AI) methods, typically powered by advances in machine learning

To ensure that a prediction model study is valuable to users, authors should prepare a transparent, complete, and accurate account of why the research was done, what they did, and what they found

An update of the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) statement aims to harmonise the landscape of prediction model studies using AI methods and to provide guidance regardless of whether regression models or machine learning methods have been used

The TRIPOD+AI statement consists of a 27 item checklist, an expanded checklist that details reporting recommendations for each item, and a TRIPOD+AI for Abstracts checklist containing 13 items

TRIPOD+AI aims to assist authors in the complete reporting of their study and help peer reviewers, editors, policymakers, end users, and patients understand the data, methods, findings and conclusions of AI driven research

Adherence to the TRIPOD+AI reporting recommendations could encourage the improved use of research time, effort, and money

Prediction models are used across different healthcare settings. They are used to estimate an outcome value or risk. Most models estimate the probability of the presence of a particular health condition (diagnostic) or whether a particular outcome will occur in the future (prognostic).[1] Their primary use is to support clinical decision making, such as whether to refer patients for further testing, monitor disease deterioration or treatment effects, or initiate treatment or lifestyle changes. Examples of well known prediction models include EuroSCORE II (cardiac surgery),[2] the Gail model (breast cancer),[3] the Framingham risk score (cardiovascular disease),[4] IMPACT (traumatic brain injury),[5] and FRAX (osteoporotic and hip fractures).[6]

Prediction models are abundant in the biomedical literature, with thousands of models published annually (and increasing), and have been developed for many outcomes and health conditions.[7 8] At least 731 diagnostic and prognostic prediction model studies on covid-19 were published during the first 12 months of the pandemic.[9] Despite this interest in developing prediction models, there have been longstanding

TRIPOD 2015 Adherence assessment form

	i.	It is described how predictions for individuals (in the validation set) were obtained from the model being validated *E.g. Using the original reported model coefficients with or without the intercept, and/or using updated or refitted model coefficients, or using a nomogram, spreadsheet or web calculator.*	Not applicable	Y / N	Y / N	=V10ci
	10d	**Specify all measures used to assess model performance and, if relevant, to compare multiple models.[2]** *These should be described in the methods section of the paper (item 16 addresses the reporting of the results for model performance).*	**Score 1 if elements 10di and 10dii are scored as "Y"[2]**	**Score 1 if elements 10di and 10dii are scored as "Y"[2]**	**Score 1 if all elements are scored as "Y"[2]**	**Score 1 if elements 10di and 10dii are scored as "Y"[2]**
	i	Measures for model discrimination are described *E.g. C-index / area under the ROC curve*		Y / N	Y / N	Y / N	=Y if D10di=Y AND V10di=Y
	ii	Measures for model calibration are described *E.g. calibration plot, calibration slope or intercept, calibration table, Hosmer Lemeshow test, O/E ratio.*		Y / N	Y / N	Y / N	=Y if D10dii=Y AND V10dii=Y
	iii	Other performance measures are described *E.g. R^2, Brier score, predictive values, sensitivity, specificity, AUC difference, decision curve analysis, net reclassification improvement, integrated discrimination improvement, AIC*		Y / N	Y / N	Y / N	=Y if D10diii=Y AND V10diii=Y
	10e	**Describe any model updating (e.g., recalibration) arising from the validation, if done.**	**Not applicable**	**Score 1 if element is scored as "Y"; score *Not applicable* if element is scored as "NA"**	**Score 1 if element is scored as "Y"; score *Not applicable* if element is scored as "NA"**	**Score 1 if element is scored as "Y"; score *Not applicable* if element is scored as "NA"**
	i	A description of model-updating is given *E.g. Intercept recalibration, regression coefficient recalibration, refitting the whole model, adding a new predictor* *If updating was done, it should be clear which updating method was applied to score Yes.* *If it is not explicitly mentioned that updating was applied in the study, score this item as 'Not applicable'.*	Not applicable	Y / N / NA	Y / N / NA	=V10ei

NTCP Model selection in TETRIS

A set of fundamental discrimination + calibration metrics should be used. For example:

Discrimination

Use gold standard together with metrics robust with respect to imbalanced datasets

→ AUC-ROC

→ Balanced accuracy

→ Matthews correlation coefficient

Calibration

Visual + bias analysis + model's confidence

→ Calibration plot, with intercept + slope

→ Spiegelhalter's Z-score

→ BCE (log-Loss) or ECE

No penalization for number of predictors (AIC, BIC)

THANK YOU

TETRIS - Risk assessment *Tools* for severe side *Effects* after *breasT Radiotherapy*: radiation safety through biological extended models and *dIgital twinS*

EU Grant Agreement n. 101166699

This project has received funding under Euratom Research and Training Programme (EURATOM) Grant ID 101166699

tetris

Follow us on social media