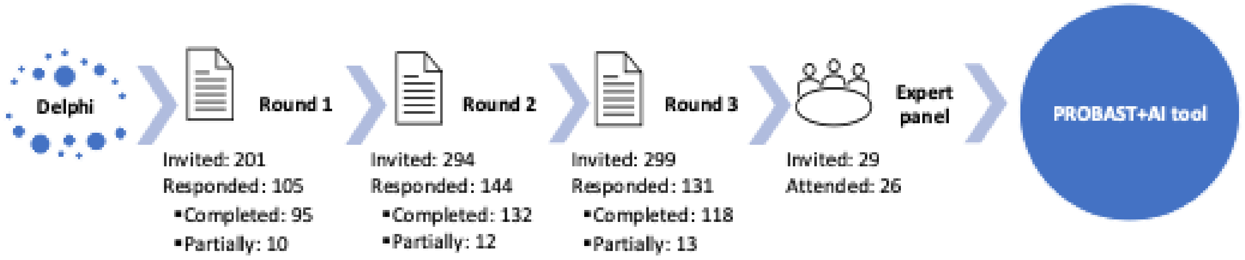
Supplemental materials

PROBAST+AI: An updated quality, risk of bias and applicability assessment tool for prediction models using regression or machine learning methods

Table of Contents

Supplementary Figure 1. Delphi and participant flow



Supplementary Table 1. Characteristics of participants that responded to the survey rounds

Characteristics	**Round 1 (n=105)** **N (%)**	**Round 2 (n=144)** **N (%)**	**Round 3 (n=131)** **N (%)**
Sex			
Female	25 (24)	46 (32)	38 (29)
Male	68 (65)	93 (64)	85 (65)
Non-binary	0 (0)	0 (0)	1 (1)
Prefer not to say	2 (1)	4 (3)	4 (3)
Missing	10 (10)	1 (1)	3 (2)
*Field of research/work**			
Statistics and data science	84 (72)	102 (71)	91 (70)
AI/ML	61 (53)	74 (51)	74 (57)
Healthcare professional	41 (35)	53 (37)	51 (39)
Systematic reviews	45 (39)	62 (45)	53 (41)
Epidemiology	37 (32)	51 (35)	50 (38)
Prediction	67 (58)	105 (73)	94 (72)
Health policy	15 (13)	15 (10)	9 (7)
Ethics	7 (6)	7 (5)	7 (5)
Other	11 (10)	7 (5)	12 (9)

*more than one field could be noted

AI: artificial intelligence; ML: machine learning

Supplementary Table 2. Flow of items through Delphi rounds

Number of items per domain in each of the Delphi rounds

Domain	**PROBAST-2019**	**Round 1**	**Round 2**	**Round 3**	**PROBAST+AI**
Development					
Participants and Data Sources	2	4	4	3	3
Predictors	3	4	4	4	4
Outcome	6	6	6	4	4
Analysis	9	13	7	7	5
Evaluation					
Participants and Data Sources	2	4	4	3	3
Predictors	3	4	4	4	4
Outcome	6	6	6	4	4
Analysis	6	13	8	8	7

Supplementary Table 5: Comparison between PROBAST and PROBAST+AI

PROBAST	**PROBAST+AI Development**	**PROBAST+AI Evaluation**	**Changes**
Participant selection	**Participants and data sources**	**Participants and data sources**	Name of domaine changed
Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?	Were appropriate data sources used?	Were appropriate data sources used?	Item split up in two items
	Was an appropriate study design used?	Was an appropriate study design used?	Item split up in two items
Were all inclusions and exclusions of participants appropriate?	Did the in- and exclusions of study participants result in a representative dataset?	Did the in- and exclusions of study participants result in a representative dataset?	Wording changed
Predictors	**Predictors**	**Predictors**	-
Were predictors defined and assessed in a similar way for all participants?	Were predictors defined and assessed in a similar way for all participants?	Were predictors defined and assessed in a similar way for all participants?	-
	Was any pre-processing of predictors similar for all participants?	Was any pre-processing of predictors similar for all participants?	New item
Were predictor assessments made without knowledge of outcome data?	Were predictor assessments made without knowledge of outcome data?	Were predictor assessments made without knowledge of outcome data?	-
Are all predictors available at the time the model is intended to be used?	Were the predictors included in the model available at the time the model was intended to be used?	Were the predictors included in the model available at the time the model was intended to be used?	Wording changed
Outcome	**Outcome**	**Outcome**	-
Was the outcome determined appropriately?	Were outcomes defined and assessed appropriately?	Were outcomes defined and assessed appropriately?	Three items combined in one.
Was a pre-specified or standard outcome definition used?			Three items combined in one.
Were predictors excluded from the outcome definition?			Three items combined in one.
Was the outcome defined and determined in a similar way for all participants?	Were outcomes defined and assessed in a similar way for all participants?	Were outcomes defined and assessed in a similar way for all participants?	Wording changed

PROBAST	**PROBAST+AI Development**	**PROBAST+AI Evaluation**	**Changes**
Was the outcome determined without knowledge of predictor information?	Were outcome assessments made without use or knowledge of predictor data?	Were outcome assessments made without use or knowledge of predictor data?	Wording changed
Was the time interval between predictor assessment and outcome determination appropriate?	Was the time interval between predictor assessment and outcome assessment appropriate?	Was the time interval between predictor assessment and outcome assessment appropriate?	Wording changed
Analysis	**Analysis**	**Analysis**	-
		Was model evaluation based on only apparent performance avoided?	New item
Were there a reasonable number of participants with the outcome?	Was there evidence that the sample size was reasonable?	Was there evidence that the sample size was reasonable?	Wording changed
Were continuous and categorical predictors handled appropriately?	Were continuous and categorical predictors handled appropriately?		Only applicable to development
Were all enrolled participants included in the analysis?			Deleted
Were participants with missing data handled appropriately?	Were participants with missing or censored data handled appropriately in the analysis?	Were participants with missing or censored data handled appropriately in the analysis?	Wording changed
	If methods to address class imbalance were used, was the model or the model predictions recalibrated?	If methods to address class imbalance were used, was the evaluation done in a dataset without imbalance correction?	New item
		If data splitting was done to create training and test datasets, was there evidence that data leakage was avoided?	New item
		If resampling methods were used to evaluate model performance, were all model development steps replicated in the resampling process?	New item
Was selection of predictors based on univariable analysis avoided? (development only)			Deleted

PROBAST	**PROBAST+AI Development**	**PROBAST+AI Evaluation**	**Changes**
Were complexities in the data (e.g. censoring, competing risks, sampling of controls) accounted for appropriately?			Deleted
Were relevant model performance measures evaluated appropriately?		Was the predictive performance of the model evaluated appropriately, e.g., calibration, discrimination, and net benefit?	Wording changed
Was model overfitting, underfitting and optimism in model performance accounted for? (development only)	Were methods used to address potential model overfitting?		Wording changed
Do predictors and their assigned weights in the final model correspond to the results from multivariable analysis? (development only)			Deleted