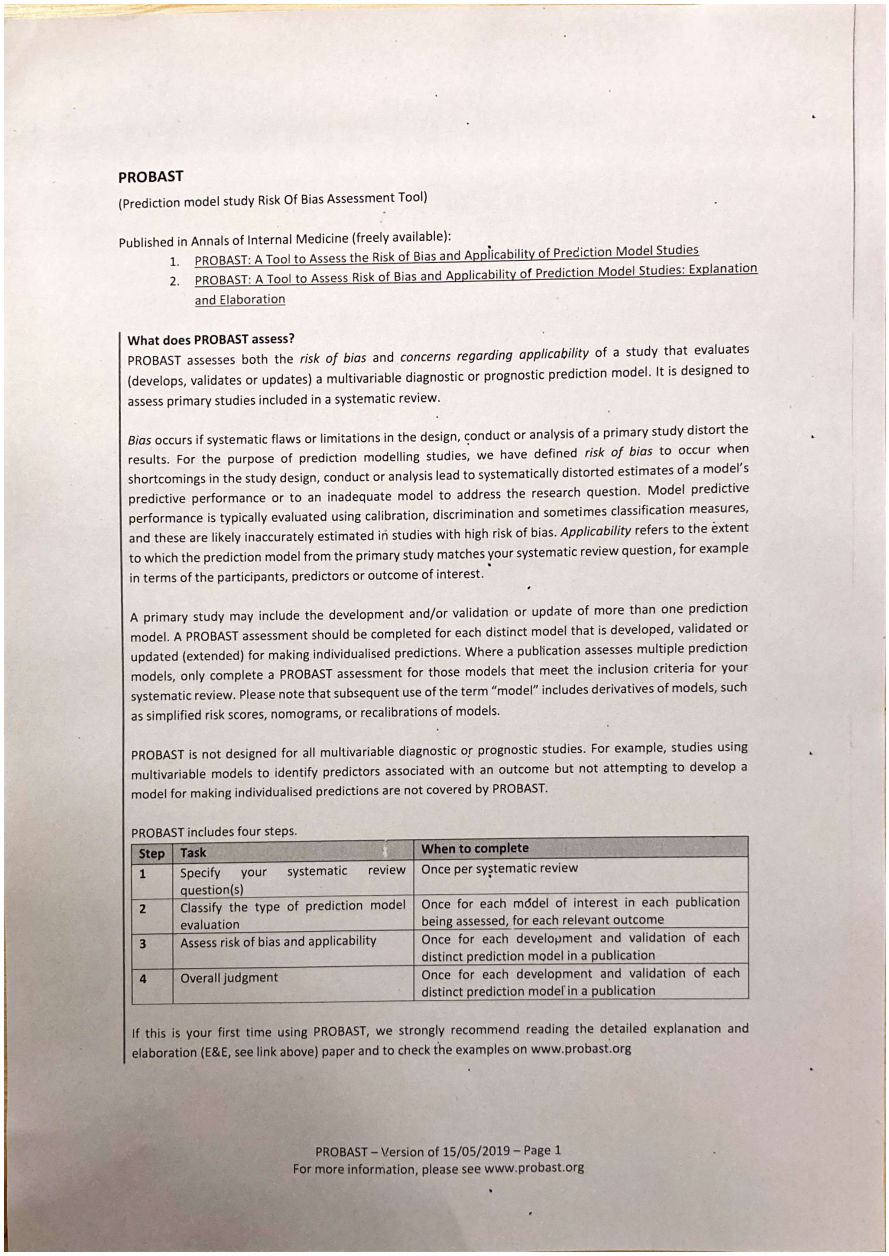
Appendix 3: PROBAST assessment of all included studies.

Zhang et al



PROBAST

(Prediction model study Risk Of Bias Assessment Tool)

Published in Annals of Internal Medicine (freely available):

1. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies
2. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration

What does PROBAST assess?

PROBAST assesses both the *risk of bias* and *concerns regarding applicability* of a study that evaluates (develops, validates or updates) a multivariable diagnostic or prognostic prediction model. It is designed to assess primary studies included in a systematic review.

Bias occurs if systematic flaws or limitations in the design, conduct or analysis of a primary study distort the results. For the purpose of prediction modelling studies, we have defined *risk of bias* to occur when shortcomings in the study design, conduct or analysis lead to systematically distorted estimates of a model's predictive performance or to an inadequate model to address the research question. Model predictive performance is typically evaluated using calibration, discrimination and sometimes classification measures, and these are likely inaccurately estimated in studies with high risk of bias. *Applicability* refers to the extent to which the prediction model from the primary study matches your systematic review question, for example in terms of the participants, predictors or outcome of interest.

A primary study may include the development and/or validation or update of more than one prediction model. A PROBAST assessment should be completed for each distinct model that is developed, validated or updated (extended) for making individualised predictions. Where a publication assesses multiple prediction models, only complete a PROBAST assessment for those models that meet the inclusion criteria for your systematic review. Please note that subsequent use of the term "model" includes derivatives of models, such as simplified risk scores, nomograms, or recalibrations of models.

PROBAST is not designed for all multivariable diagnostic or prognostic studies. For example, studies using multivariable models to identify predictors associated with an outcome but not attempting to develop a model for making individualised predictions are not covered by PROBAST.

PROBAST includes four steps.

Step	Task	When to complete
1	Specify your systematic review question(s)	Once per systematic review
2	Classify the type of prediction model evaluation	Once for each model of interest in each publication being assessed, for each relevant outcome
3	Assess risk of bias and applicability	Once for each development and validation of each distinct prediction model in a publication
4	Overall judgment	Once for each development and validation of each distinct prediction model in a publication

If this is your first time using PROBAST, we strongly recommend reading the detailed explanation and elaboration (E&E, see link above) paper and to check the examples on www.probast.org

Step 1: Specify your systematic review question

State your systematic review question to facilitate the assessment of the applicability of the evaluated models to your question. *The following table should be completed once per systematic review.*

Criteria	Specify your systematic review question
Intended use of model:	Prediction of MCIDs in PROMs for patients undergoing TKA/THA
***Participants** including selection criteria and setting:*	All patients undergoing total knee or hip arthroplasty
***Predictors** (used in prediction modelling), including types of predictors (e.g. history, clinical examination, biochemical markers, imaging tests), time of measurement, specific measurement issues (e.g., any requirements/ prohibitions for specialized equipment):*	All that are available to the Researchers prior to surgery
Outcome to be predicted:	Whether patients achieve a PROMs MCID or not (classification)

DOMAIN 1: Participants			
A. Risk of Bias			
Describe the sources of data and criteria for participant selection: Donat replacement registry i anslutning TKA.			
		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		Y	Y
1.2 Were all inclusions and exclusions of participants appropriate?		Y	Y
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	low	low
Rationale of bias rating: All TKA patients eligible for the review.			
B. Applicability			
Describe included participants, setting and dates:			
Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	low	low
Rationale of applicability rating: Fits the study purpose.			

DOMAIN 2: Predictors			
A. Risk of Bias			
List and describe predictors included in the final model, e.g. definition and timing of assessment: 18 Predictors before surgery			
		Dev	Val
2.1 Were predictors defined and assessed in a similar way for all participants?		Y	Y
2.2 Were predictor assessments made without knowledge of outcome data?		~~N~~	~~N~~
2.3 Are all predictors available at the time the model is intended to be used?		Y	Y
Risk of bias introduced by predictors or their assessment	**RISK:** *(low/ high/ unclear)*	low	low
Rationale of bias rating: "2.2: Blinded to the outcome"			
B. Applicability			
Concern that the definition, assessment or timing of predictors in the model do not match the review question	**CONCERN:** *(low/ high/ unclear)*	low	low
Rationale of applicability rating: Nothing to add.			

DOMAIN 3: Outcome			
A. Risk of Bias			
Describe the outcome, how it was defined and determined, and the time interval between predictor assessment and outcome determination: Anchor based MCID, two years after surgery			
		Dev	Val
3.1 Was the outcome determined appropriately?		Y	Y
3.2 Was a pre-specified or standard outcome definition used?		Y	Y
3.3 Were predictors excluded from the outcome definition?		N	N
3.4 Was the outcome defined and determined in a similar way for all participants?		Y	Y
3.5 Was the outcome determined without knowledge of predictor information?		N	N
3.6 Was the time interval between predictor assessment and outcome determination appropriate?		Y	Y
Risk of bias introduced by the outcome or its determination	**RISK:** *(low/ high/ unclear)*	Low	Low
Rationale of bias rating: Knowing outcome and including predictors is inherit to the study setting.			
B. Applicability			
At what time point was the outcome determined: 24 months after surgery			
If a composite outcome was used, describe the relative frequency/distribution of each contributing outcome:			
Concern that the outcome, its definition, timing or determination do not match the review question	**CONCERN:** *(low/ high/ unclear)*	Low	Low
Rationale of applicability rating: No concerns			

DOMAIN 4: Analysis			
Risk of Bias			
Describe numbers of participants, number of candidate predictors, outcome events and events per candidate predictor: 2840 ~18 Ev = 31,5			
Describe how the model was developed (for example in regards to modelling technique (e.g. survival or logistic modelling), predictor selection, and risk group definition): CV to optimize hyperparameters ; upsampling			
Describe whether and how the model was validated, either internally (e.g. bootstrapping, cross validation, random split sample) or externally (e.g. temporal validation, geographical validation, different setting, different type of participants): CV, upsampling			
Describe the performance measures of the model, e.g. (re)calibration, discrimination, (re)classification, net benefit, and whether they were adjusted for optimism: AUC, F1, Sen, Spec ; Yes, CV and test dataset			
Describe any participants who were excluded from the analysis: Bilateral TKA ; lost to follow up ; inflammatory arthritis			
Describe missing data on predictors and outcomes as well as methods used for missing data: Median value imputation			
		Dev	Val
4.1 Were there a reasonable number of participants with the outcome?		Y	Y
4.2 Were continuous and categorical predictors handled appropriately?		Y	Y
4.3 Were all enrolled participants included in the analysis?		Y	Y
4.4 Were participants with missing data handled appropriately?		X	X
4.5 Was selection of predictors based on univariable analysis avoided?		Y	
4.6 Were complexities in the data (e.g. censoring, competing risks, sampling of controls) accounted for appropriately?		PY	PY
4.7 Were relevant model performance measures evaluated appropriately?		Y	Y
4.8 Were model overfitting and optimism in model performance accounted for?		Y	
4.9 Do predictors and their assigned weights in the final model correspond to the results from multivariable analysis?		Y	
Risk of bias introduced by the analysis	**RISK:** *(low/ high/ unclear)*	low	low
Rationale of bias rating: Excellent analysis. Predictor selection based on literature.			

Step 4: Overall assessment

Use the following tables to reach overall judgements about risk of bias and concerns regarding applicability of the prediction model evaluation (development and/or validation) across all assessed domains.
Complete for each evaluation of a distinct model.

Reaching an overall judgement about risk of bias of the prediction model evaluation	
Low risk of bias	If all domains were rated low risk of bias. If a prediction model was developed without any external validation, and it was rated as low risk of bias for all domains, consider downgrading to **high risk of bias**. Such a model can only be considered as low risk of bias, if the development was based on a very large data set and included some form of internal validation.
High risk of bias	If at least one domain is judged to be at **high risk of bias**.
Unclear risk of bias	If an unclear risk of bias was noted in at least one domain and it was low risk for all other domains.

Reaching an overall judgement about applicability of the prediction model evaluation	
Low concerns regarding applicability	If low concerns regarding applicability for all domains, the prediction model evaluation is judged to have **low concerns regarding applicability**.
High concerns regarding applicability	If high concerns regarding applicability for at least one domain, the prediction model evaluation is judged to have **high concerns regarding applicability**.
Unclear concerns regarding applicability	If unclear concerns (but no "high concern") regarding applicability for at least one domain, the prediction model evaluation is judged to have **unclear concerns regarding applicability** overall.

Overall judgement about risk of bias and applicability of the prediction model evaluation		
Overall judgement of risk of bias	**RISK:** *(low/ high/ unclear)*	low
Summary of sources of potential bias:		
Overall judgement of applicability	**CONCERN:** *(low/ high/ unclear)*	low
Summary of applicability concerns:		

Step 2: Classify the type of prediction model evaluation

Use the following table to classify the evaluation as model development, model validation or model update, or combination. Different signalling questions apply for different types of prediction model evaluation. If the evaluation does not fit one of these classifications then PROBAST should not be used.

Classify the evaluation based on its aim			
Type of prediction study	**PROBAST boxes to complete**	**Tick as appropriate**	**Definition for type of prediction model study**
Development only	Development		Prediction model development without external validation. These studies may include internal validation methods, such as bootstrapping and cross-validation techniques.
Development and validation	Development and validation	X	Prediction model development combined with external validation in other participants in the same article.
Validation only	Validation		External validation of existing (previously developed) model in other participants.

This table should be completed once for each publication being assessed and for each relevant outcome in your review.

Publication reference	Zhang 21
Models of interest	All machine learning n odels
Outcome of interest	MCID

Step 3: Assess risk of bias and applicability

PROBAST is structured as four key domains. Each domain is judged for risk of bias (low, high or unclear) and includes signalling questions to help make judgements. Signalling questions are rated as yes (Y), probably yes (PY), probably no (PN), no (N) or no information (NI). All signalling questions are phrased so that "yes" indicates absence of bias. Any signalling question rated as "no" or "probably no" flags the potential for bias; you will need to use your judgement to determine whether the domain should be rated as "high", "low" or "unclear" risk of bias. The guidance document contains further instructions and examples on rating signalling questions and risk of bias for each domain.

The first three domains are also rated for concerns regarding applicability (low/ high/ unclear) to your review question defined above.

Complete all domains separately for each evaluation of a distinct model. Shaded boxes indicate where signalling questions do not apply and should not be answered.

Katakam et al

PROBAST

(Prediction model study Risk Of Bias Assessment Tool)

Published in Annals of Internal Medicine (freely available):

1. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies
2. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration

What does PROBAST assess?

PROBAST assesses both the *risk of bias* and *concerns regarding applicability* of a study that evaluates (develops, validates or updates) a multivariable diagnostic or prognostic prediction model. It is designed to assess primary studies included in a systematic review.

Bias occurs if systematic flaws or limitations in the design, conduct or analysis of a primary study distort the results. For the purpose of prediction modelling studies, we have defined *risk of bias* to occur when shortcomings in the study design, conduct or analysis lead to systematically distorted estimates of a model's predictive performance or to an inadequate model to address the research question. Model predictive performance is typically evaluated using calibration, discrimination and sometimes classification measures, and these are likely inaccurately estimated in studies with high risk of bias. *Applicability* refers to the extent to which the prediction model from the primary study matches your systematic review question, for example in terms of the participants, predictors or outcome of interest.

A primary study may include the development and/or validation or update of more than one prediction model. A PROBAST assessment should be completed for each distinct model that is developed, validated or updated (extended) for making individualised predictions. Where a publication assesses multiple prediction models, only complete a PROBAST assessment for those models that meet the inclusion criteria for your systematic review. Please note that subsequent use of the term "model" includes derivatives of models, such as simplified risk scores, nomograms, or recalibrations of models.

PROBAST is not designed for all multivariable diagnostic or prognostic studies. For example, studies using multivariable models to identify predictors associated with an outcome but not attempting to develop a model for making individualised predictions are not covered by PROBAST.

PROBAST includes four steps.

Step	Task	When to complete
1	Specify your systematic review question(s)	Once per systematic review
2	Classify the type of prediction model evaluation	Once for each model of interest in each publication being assessed, for each relevant outcome
3	Assess risk of bias and applicability	Once for each development and validation of each distinct prediction model in a publication
4	Overall judgment	Once for each development and validation of each distinct prediction model in a publication

If this is your first time using PROBAST, we strongly recommend reading the detailed explanation and elaboration (E&E, see link above) paper and to check the examples on www.probast.org

Step 1: Specify your systematic review question

State your systematic review question to facilitate the assessment of the applicability of the evaluated models to your question. *The following table should be completed once per systematic review.*

Criteria	Specify your systematic review question
Intended use of model:	Prediction of MCIDs in PROMs for patients undergoing TKA/THA
***Participants** including selection criteria and setting:*	All patients undergoing total knee or hip arthroplasty
***Predictors** (used in prediction modelling), including types of predictors (e.g. history, clinical examination, biochemical markers, imaging tests), time of measurement, specific measurement issues (e.g., any requirements/ prohibitions for specialized equipment):*	All that are available to the Researchers prior to surgery
Outcome to be predicted:	Whether patients achieve a PROMs MCID or not (classification)

Step 2: Classify the type of prediction model evaluation

Use the following table to classify the evaluation as model development, model validation or model update, or combination. Different signalling questions apply for different types of prediction model evaluation. If the evaluation does not fit one of these classifications then PROBAST should not be used.

Classify the evaluation based on its aim			
Type of prediction study	**PROBAST boxes to complete**	**Tick as appropriate**	**Definition for type of prediction model study**
Development only	Development		Prediction model development without external validation. These studies may include internal validation methods, such as bootstrapping and cross-validation techniques.
Development and validation	Development and validation	X	Prediction model development combined with external validation in other participants in the same article.
Validation only	Validation		External validation of existing (previously developed) model in other participants.

This table should be completed once for each publication being assessed and for each relevant outcome in your review.

Publication reference	Katakam 2021
Models of interest	All machine learning models
Outcome of interest	MCID

Step 3: Assess risk of bias and applicability

PROBAST is structured as four key domains. Each domain is judged for risk of bias (low, high or unclear) and includes signalling questions to help make judgements. Signalling questions are rated as yes (Y), probably yes (PY), probably no (PN), no (N) or no information (NI). All signalling questions are phrased so that "yes" indicates absence of bias. Any signalling question rated as "no" or "probably no" flags the potential for bias; you will need to use your judgement to determine whether the domain should be rated as "high", "low" or "unclear" risk of bias. The guidance document contains further instructions and examples on rating signalling questions and risk of bias for each domain.

The first three domains are also rated for concerns regarding applicability (low/ high/ unclear) to your review question defined above.

Complete all domains separately for each evaluation of a distinct model. Shaded boxes indicate where signalling questions do not apply and should not be answered.

DOMAIN 1: Participants			
A. Risk of Bias			
Describe the sources of data and criteria for participant selection: - 2 academic medical centers & 3 community medical centers - 2016 - 2018 - ≥ 18 yo - TKA			
		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		Y	Y
1.2 Were all inclusions and exclusions of participants appropriate?		Y	Y
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	low	low
Rationale of bias rating: Retrospective cual III study			
B. Applicability			
Describe included participants, setting and dates: see A.			
Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	low	low
Rationale of applicability rating: TKA patients			

DOMAIN 2: Predictors			
A. Risk of Bias			
List and describe predictors included in the final model, e.g. definition and timing of assessment:			
		Dev	Val
2.1 Were predictors defined and assessed in a similar way for all participants?		Y	Y
2.2 Were predictor assessments made without knowledge of outcome data?		N	N
2.3 Are all predictors available at the time the model is intended to be used?		Y	Y
Risk of bias introduced by predictors or their assessment	**RISK:** *(low/ high/ unclear)*	low	low
Rationale of bias rating: "2.2: Blinded to the outcome" - knowing the outcome inherit to this study setting (MCID calculation) & model training.			
B. Applicability			
Concern that the definition, assessment or timing of predictors in the model do not match the review question	**CONCERN:** *(low/ high/ unclear)*	low	low
Rationale of applicability rating: well defined pre-surgery data			

DOMAIN 3: Outcome

A. Risk of Bias

Describe the outcome, how it was defined and determined, and the time interval between predictor assessment and outcome determination:

- MCID ($\geq$ threshold)
- time: 1 year
- Distribution based

		Dev	Val
3.1 Was the outcome determined appropriately?		Y	Y
3.2 Was a pre-specified or standard outcome definition used?		Y	Y
3.3 Were predictors excluded from the outcome definition?		N	N
3.4 Was the outcome defined and determined in a similar way for all participants?		Y	Y
3.5 Was the outcome determined without knowledge of predictor information?		N	N
3.6 Was the time interval between predictor assessment and outcome determination appropriate?		Y	Y
Risk of bias introduced by the outcome or its determination	**RISK:** *(low/ high/ unclear)*	Low	low

Rationale of bias rating:

Knowing the outcome is relevant for calculating the MCID.

B. Applicability

At what time point was the outcome determined:

1 year

If a composite outcome was used, describe the relative frequency/distribution of each contributing outcome:

–

Concern that the outcome, its definition, timing or determination do not match the review question	**CONCERN:** *(low/ high/ unclear)*	Low	low

Rationale of applicability rating:

One year is a good time for assessment.

DOMAIN 4: Analysis			
Risk of Bias			
Describe numbers of participants, number of candidate predictors, outcome events and events per candidate predictor: 744 24 6,2			
Describe how the model was developed (for example in regards to modelling technique (e.g. survival or logistic modelling), predictor selection, and risk group definition): Hyperparameter tuning by CV. RF feature selection.			
Describe whether and how the model was validated, either internally (e.g. bootstrapping, cross validation, random split sample) or externally (e.g. temporal validation, geographical validation, different setting, different type of participants): CV			
Describe the performance measures of the model, e.g. (re)calibration, discrimination, (re)classification, net benefit, and whether they were adjusted for optimism: AUC, calibration, Brier Score			
Describe any participants who were excluded from the analysis: <18; no Numbers reported			
Describe missing data on predictors and outcomes as well as methods used for missing data: Multiple imputation < 30% missing			
		Dev	Val
4.1 Were there a reasonable number of participants with the outcome?		Y	Y
4.2 Were continuous and categorical predictors handled appropriately?		Y	Y
4.3 Were all enrolled participants included in the analysis?		Y	Y
4.4 Were participants with missing data handled appropriately?		Y	Y
4.5 Was selection of predictors based on univariable analysis avoided?		Y	
4.6 Were complexities in the data (e.g. censoring, competing risks, sampling of controls) accounted for appropriately?		Y	Y
4.7 Were relevant model performance measures evaluated appropriately?		Y	Y
4.8 Were model overfitting and optimism in model performance accounted for?		Y	
4.9 Do predictors and their assigned weights in the final model correspond to the results from multivariable analysis?		Y	
Risk of bias introduced by the analysis	**RISK:** *(low/ high/ unclear)*	low	low
Rationale of bias rating: Well performed analysis with all relevant metrics.			

EVP not a problem in machine learning.

Step 4: Overall assessment

Use the following tables to reach overall judgements about risk of bias and concerns regarding applicability of the prediction model evaluation (development and/or validation) across all assessed domains.
Complete for each evaluation of a distinct model.

Reaching an overall judgement about risk of bias of the prediction model evaluation	
Low risk of bias	If all domains were rated low risk of bias. If a prediction model was developed without any external validation, and it was rated as low risk of bias for all domains, consider downgrading to **high risk of bias**. Such a model can only be considered as low risk of bias, if the development was based on a very large data set and included some form of internal validation.
High risk of bias	If at least one domain is judged to be at **high risk of bias**.
Unclear risk of bias	If an unclear risk of bias was noted in at least one domain and it was low risk for all other domains.

Reaching an overall judgement about applicability of the prediction model evaluation	
Low concerns regarding applicability	If low concerns regarding applicability for all domains, the prediction model evaluation is judged to have **low concerns regarding applicability**.
High concerns regarding applicability	If high concerns regarding applicability for at least one domain, the prediction model evaluation is judged to have **high concerns regarding applicability**.
Unclear concerns regarding applicability	If unclear concerns (but no "high concern") regarding applicability for at least one domain, the prediction model evaluation is judged to have **unclear concerns regarding applicability** overall.

Overall judgement about risk of bias and applicability of the prediction model evaluation		
Overall judgement of risk of bias	**RISK:** *(low/ high/ unclear)*	Low
Summary of sources of potential bias:		
Overall judgement of applicability	**CONCERN:** *(low/ high/ unclear)*	Low
Summary of applicability concerns:		

PROBAST

(Prediction model study Risk Of Bias Assessment Tool)

Published in Annals of Internal Medicine (freely available):

1. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies
2. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration

What does PROBAST assess?

PROBAST assesses both the *risk of bias* and *concerns regarding applicability* of a study that evaluates (develops, validates or updates) a multivariable diagnostic or prognostic prediction model. It is designed to assess primary studies included in a systematic review.

Bias occurs if systematic flaws or limitations in the design, conduct or analysis of a primary study distort the results. For the purpose of prediction modelling studies, we have defined *risk of bias* to occur when shortcomings in the study design, conduct or analysis lead to systematically distorted estimates of a model's predictive performance or to an inadequate model to address the research question. Model predictive performance is typically evaluated using calibration, discrimination and sometimes classification measures, and these are likely inaccurately estimated in studies with high risk of bias. *Applicability* refers to the extent to which the prediction model from the primary study matches your systematic review question, for example in terms of the participants, predictors or outcome of interest.

A primary study may include the development and/or validation or update of more than one prediction model. A PROBAST assessment should be completed for each distinct model that is developed, validated or updated (extended) for making individualised predictions. Where a publication assesses multiple prediction models, only complete a PROBAST assessment for those models that meet the inclusion criteria for your systematic review. Please note that subsequent use of the term "model" includes derivatives of models, such as simplified risk scores, nomograms, or recalibrations of models.

PROBAST is not designed for all multivariable diagnostic or prognostic studies. For example, studies using multivariable models to identify predictors associated with an outcome but not attempting to develop a model for making individualised predictions are not covered by PROBAST.

PROBAST includes four steps.

Step	Task	When to complete
1	Specify your systematic review question(s)	Once per systematic review
2	Classify the type of prediction model evaluation	Once for each model of interest in each publication being assessed, for each relevant outcome
3	Assess risk of bias and applicability	Once for each development and validation of each distinct prediction model in a publication
4	Overall judgment	Once for each development and validation of each distinct prediction model in a publication

If this is your first time using PROBAST, we strongly recommend reading the detailed explanation and elaboration (E&E, see link above) paper and to check the examples on www.probast.org

Step 1: Specify your systematic review question

State your systematic review question to facilitate the assessment of the applicability of the evaluated models to your question. *The following table should be completed once per systematic review.*

Criteria	Specify your systematic review question
Intended use of model:	Prediction of MCIDs in PROMs for patients undergoing TKA/THA
***Participants** including selection criteria and setting:*	All patients undergoing total knee or hip arthroplasty
***Predictors** (used in prediction modelling), including types of predictors (e.g. history, clinical examination, biochemical markers, imaging tests), time of measurement, specific measurement issues (e.g., any requirements/ prohibitions for specialized equipment):*	All that are available to the Researchers prior to surgery
Outcome to be predicted:	Whether patients achieve a PROMs MCID or not (classification)

Step 2: Classify the type of prediction model evaluation

Use the following table to classify the evaluation as model development, model validation or model update, or combination. Different signalling questions apply for different types of prediction model evaluation. If the evaluation does not fit one of these classifications then PROBAST should not be used.

Classify the evaluation based on its aim			
Type of prediction study	**PROBAST boxes to complete**	**Tick as appropriate**	**Definition for type of prediction model study**
Development only	Development		Prediction model development without external validation. These studies may include internal validation methods, such as bootstrapping and cross-validation techniques.
Development and validation	Development and validation	X	Prediction model development combined with external validation in other participants in the same article.
Validation only	Validation		External validation of existing (previously developed) model in other participants.

This table should be completed once for each publication being assessed and for each relevant outcome in your review.

Publication reference	Fontana et al 2019
Models of interest	All machine learning models
Outcome of interest	MCID

Step 3: Assess risk of bias and applicability

PROBAST is structured as four key domains. Each domain is judged for risk of bias (low, high or unclear) and includes signalling questions to help make judgements. Signalling questions are rated as yes (Y), probably yes (PY), probably no (PN), no (N) or no information (NI). All signalling questions are phrased so that "yes" indicates absence of bias. Any signalling question rated as "no" or "probably no" flags the potential for bias; you will need to use your judgement to determine whether the domain should be rated as "high", "low" or "unclear" risk of bias. The guidance document contains further instructions and examples on rating signalling questions and risk of bias for each domain.

The first three domains are also rated for concerns regarding applicability (low/ high/ unclear) to your review question defined above.

Complete all domains separately for each evaluation of a distinct model. Shaded boxes indicate where signalling questions do not apply and should not be answered.

DOMAIN 1: Participants			
A. Risk of Bias			
Describe the sources of data and criteria for participant selection: - Single high volume institution - TJA patients who completed 2y follow up			
		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		PY	PY
1.2 Were all inclusions and exclusions of participants appropriate?		Y	Y
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	Low	Low
Rationale of bias rating: Selection was reasonable.			
B. Applicability			
Describe included participants, setting and dates: See A. ; 2007 - 2012			
Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	Low	low
Rationale of applicability rating:			

DOMAIN 2: Predictors			
A. Risk of Bias			
List and describe predictors included in the final model, e.g. definition and timing of assessment: - Before surgery - ≥ 66 predictors before surgery depending on model			
		Dev	Val
2.1 Were predictors defined and assessed in a similar way for all participants?		Y	Y
2.2 Were predictor assessments made without knowledge of outcome data?		Y	Y
2.3 Are all predictors available at the time the model is intended to be used?		Y	Y
Risk of bias introduced by predictors or their assessment	**RISK:** *(low/ high/ unclear)*	low	low
Rationale of bias rating: - Same predictor determination for all outcomes - All predictors are available pre-surgery			
B. Applicability			
Concern that the definition, assessment or timing of predictors in the model do not match the review question	**CONCERN:** *(low/ high/ unclear)*	low	low
Rationale of applicability rating: predictors available before surgery and therefore fit the study purpose.			

DOMAIN 3: Outcome			
A. Risk of Bias			
Describe the outcome, how it was defined and determined, and the time interval between predictor assessment and outcome determination: MCID, anchor - and distribution based.			
		Dev	Val
3.1 Was the outcome determined appropriately?		Y	Y
3.2 Was a pre-specified or standard outcome definition used?		Y	Y
3.3 Were predictors excluded from the outcome definition?		N	N
3.4 Was the outcome defined and determined in a similar way for all participants?		Y	Y
3.5 Was the outcome determined without knowledge of predictor information?		N	N
3.6 Was the time interval between predictor assessment and outcome determination appropriate?		PY	PY
Risk of bias introduced by the outcome or its determination	**RISK:** *(low/ high/ unclear)*	low	low
Rationale of bias rating: MCID is by nature determined with the outcome PROM value.			
B. Applicability			
At what time point was the outcome determined: Pre- and post surgery.			
If a composite outcome was used, describe the relative frequency/distribution of each contributing outcome: -			
Concern that the outcome, its definition, timing or determination do not match the review question	**CONCERN:** *(low/ high/ unclear)*	low	low
Rationale of applicability rating: Outcome defined as suggested in the studies inclusion criteria.			

DOMAIN 4: Analysis			
Risk of Bias			
Describe numbers of participants, number of candidate predictors, outcome events and events per candidate predictor: - 6480 TKA & 7239 THA - EVP > 100 - 17 predictors (candidate)			
Describe how the model was developed (for example in regards to modelling technique (e.g. survival or logistic modelling), predictor selection, and risk group definition): - ML feature selection			
Describe whether and how the model was validated, either internally (e.g. bootstrapping, cross validation, random split sample) or externally (e.g. temporal validation, geographical validation, different setting, different type of participants): - 5 fold CV on training dataset (80% of total data)			
Describe the performance measures of the model, e.g. (re)calibration, discrimination, (re)classification, net benefit, and whether they were adjusted for optimism: - AUC - calibration			
Describe any participants who were excluded from the analysis: - Not completed follow-up			
Describe missing data on predictors and outcomes as well as methods used for missing data: - Imputation to the mean (non-categorical values). - Separate category for missing in categorical variables.			
		Dev	Val
4.1 Were there a reasonable number of participants with the outcome?		Y	Y
4.2 Were continuous and categorical predictors handled appropriately?		Y	Y
4.3 Were all enrolled participants included in the analysis?		Y	Y
4.4 Were participants with missing data handled appropriately?		PY	PY
4.5 Was selection of predictors based on univariable analysis avoided?		Y	
4.6 Were complexities in the data (e.g. censoring, competing risks, sampling of controls) accounted for appropriately?		Y	Y
4.7 Were relevant model performance measures evaluated appropriately?		Y	Y
4.8 Were model overfitting and optimism in model performance accounted for?		Y	
4.9 Do predictors and their assigned weights in the final model correspond to the results from multivariable analysis?		PY	
Risk of bias introduced by the analysis	**RISK:** *(low/ high/ unclear)*	low	low
Rationale of bias rating: EVP > 100; Both pre- and post surgery PROMs reported.			

Step 4: Overall assessment

Use the following tables to reach overall judgements about risk of bias and concerns regarding applicability of the prediction model evaluation (development and/or validation) across all assessed domains.
Complete for each evaluation of a distinct model.

Reaching an overall judgement about risk of bias of the prediction model evaluation	
Low risk of bias	If all domains were rated low risk of bias. If a prediction model was developed without any external validation, and it was rated as low risk of bias for all domains, consider downgrading to **high risk of bias**. Such a model can only be considered as low risk of bias, if the development was based on a very large data set and included some form of internal validation.
High risk of bias	If at least one domain is judged to be at **high risk of bias**.
Unclear risk of bias	If an unclear risk of bias was noted in at least one domain and it was low risk for all other domains.

Reaching an overall judgement about applicability of the prediction model evaluation	
Low concerns regarding applicability	If low concerns regarding applicability for all domains, the prediction model evaluation is judged to have **low concerns regarding applicability**.
High concerns regarding applicability	If high concerns regarding applicability for at least one domain, the prediction model evaluation is judged to have **high concerns regarding applicability**.
Unclear concerns regarding applicability	If unclear concerns (but no "high concern") regarding applicability for at least one domain, the prediction model evaluation is judged to have **unclear concerns regarding applicability** overall.

Overall judgement about risk of bias and applicability of the prediction model evaluation		
Overall judgement of risk of bias	**RISK:** *(low/ high/ unclear)*	low
Summary of sources of potential bias: –		
Overall judgement of applicability	**CONCERN:** *(low/ high/ unclear)*	low
Summary of applicability concerns: –		

PROBAST

(Prediction model study Risk Of Bias Assessment Tool)

Published in Annals of Internal Medicine (freely available):

1. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies
2. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration

What does PROBAST assess?

PROBAST assesses both the *risk of bias* and *concerns regarding applicability* of a study that evaluates (develops, validates or updates) a multivariable diagnostic or prognostic prediction model. It is designed to assess primary studies included in a systematic review.

Bias occurs if systematic flaws or limitations in the design, conduct or analysis of a primary study distort the results. For the purpose of prediction modelling studies, we have defined *risk of bias* to occur when shortcomings in the study design, conduct or analysis lead to systematically distorted estimates of a model's predictive performance or to an inadequate model to address the research question. Model predictive performance is typically evaluated using calibration, discrimination and sometimes classification measures, and these are likely inaccurately estimated in studies with high risk of bias. *Applicability* refers to the extent to which the prediction model from the primary study matches your systematic review question, for example in terms of the participants, predictors or outcome of interest.

A primary study may include the development and/or validation or update of more than one prediction model. A PROBAST assessment should be completed for each distinct model that is developed, validated or updated (extended) for making individualised predictions. Where a publication assesses multiple prediction models, only complete a PROBAST assessment for those models that meet the inclusion criteria for your systematic review. Please note that subsequent use of the term "model" includes derivatives of models, such as simplified risk scores, nomograms, or recalibrations of models.

PROBAST is not designed for all multivariable diagnostic or prognostic studies. For example, studies using multivariable models to identify predictors associated with an outcome but not attempting to develop a model for making individualised predictions are not covered by PROBAST.

PROBAST includes four steps.

Step	Task	When to complete
1	Specify your systematic review question(s)	Once per systematic review
2	Classify the type of prediction model evaluation	Once for each model of interest in each publication being assessed, for each relevant outcome
3	Assess risk of bias and applicability	Once for each development and validation of each distinct prediction model in a publication
4	Overall judgment	Once for each development and validation of each distinct prediction model in a publication

If this is your first time using PROBAST, we strongly recommend reading the detailed explanation and elaboration (E&E, see link above) paper and to check the examples on www.probast.org

Step 1: Specify your systematic review question

State your systematic review question to facilitate the assessment of the applicability of the evaluated models to your question. *The following table should be completed once per systematic review.*

Criteria	Specify your systematic review question
Intended use of model:	Prediction of MCIDs in PROMs for patients undergoing TKA/THA
***Participants** including selection criteria and setting:*	All patients undergoing total knee or hip arthroplasty
***Predictors** (used in prediction modelling), including types of predictors (e.g. history, clinical examination, biochemical markers, imaging tests), time of measurement, specific measurement issues (e.g., any requirements/ prohibitions for specialized equipment):*	All that are available to the Researchers prior to surgery
Outcome to be predicted:	Whether patients achieve a PROMs MCID or not (classification)

Step 2: Classify the type of prediction model evaluation

Use the following table to classify the evaluation as model development, model validation or model update, or combination. Different signalling questions apply for different types of prediction model evaluation. If the evaluation does not fit one of these classifications then PROBAST should not be used.

Classify the evaluation based on its aim			
Type of prediction study	**PROBAST boxes to complete**	**Tick as appropriate**	**Definition for type of prediction model study**
Development only	Development		Prediction model development without external validation. These studies may include internal validation methods, such as bootstrapping and cross-validation techniques.
Development and validation	Development and validation	X	Prediction model development combined with external validation in other participants in the same article.
Validation only	Validation		External validation of existing (previously developed) model in other participants.

This table should be completed once for each publication being assessed and for each relevant outcome in your review.

Publication reference	Habe et al
Models of interest	All machine learning models
Outcome of interest	MCID

Step 3: Assess risk of bias and applicability

PROBAST is structured as four key domains. Each domain is judged for risk of bias (low, high or unclear) and includes signalling questions to help make judgements. Signalling questions are rated as yes (Y), probably yes (PY), probably no (PN), no (N) or no information (NI). All signalling questions are phrased so that "yes" indicates absence of bias. Any signalling question rated as "no" or "probably no" flags the potential for bias; you will need to use your judgement to determine whether the domain should be rated as "high", "low" or "unclear" risk of bias. The guidance document contains further instructions and examples on rating signalling questions and risk of bias for each domain.

The first three domains are also rated for concerns regarding applicability (low/ high/ unclear) to your review question defined above.

Complete all domains separately for each evaluation of a distinct model. Shaded boxes indicate where signalling questions do not apply and should not be answered.

DOMAIN 1: Participants			
A. Risk of Bias			
Describe the sources of data and criteria for participant selection: - NHS data on TKA / THA - All patients included [except those with missing values]			
		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		Y	Y
1.2 Were all inclusions and exclusions of participants appropriate?		PY	PY
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	low	low
Rationale of bias rating: Participants fit the purpose. No exclusions criteria.			
B. Applicability			
Describe included participants, setting and dates: See A. 1.4.15 – 31.7.17			
Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	low	
Rationale of applicability rating:			

DOMAIN 2: Predictors			
A. Risk of Bias			
List and describe predictors included in the final model, e.g. definition and timing of assessment: 81 variables available before Activity			
		Dev	Val
2.1 Were predictors defined and assessed in a similar way for all participants?		Y	Y
2.2 Were predictor assessments made without knowledge of outcome data?		N	N
2.3 Are all predictors available at the time the model is intended to be used?		Y	Y
Risk of bias introduced by predictors or their assessment	**RISK:** *(low/ high/ unclear)*	low	low
Rationale of bias rating: 2.2. "Blinded to the outcome"			
B. Applicability			
Concern that the definition, assessment or timing of predictors in the model do not match the review question	**CONCERN:** *(low/ high/ unclear)*	low	low
Rationale of applicability rating: No concerns			

DOMAIN 3: Outcome

A. Risk of Bias

Describe the outcome, how it was defined and determined, and the time interval between predictor assessment and outcome determination:

Anchor and distribution based MCID

Time interval: 1 year

		Dev	Val
3.1 Was the outcome determined appropriately?		Y	Y
3.2 Was a pre-specified or standard outcome definition used?		Y	Y
3.3 Were predictors excluded from the outcome definition?		N	N
3.4 Was the outcome defined and determined in a similar way for all participants?		Y	Y
3.5 Was the outcome determined without knowledge of predictor information?		N	N
3.6 Was the time interval between predictor assessment and outcome determination appropriate?		Y	Y
Risk of bias introduced by the outcome or its determination	**RISK:** *(low/ high/ unclear)*	low	low

Rationale of bias rating:

3.3. and 3.5. are inbuilt to the problem and therefore no ROB

B. Applicability

At what time point was the outcome determined:

Retrospectively

If a composite outcome was used, describe the relative frequency/distribution of each contributing outcome:

–

Concern that the outcome, its definition, timing or determination do not match the review question	**CONCERN:** *(low/ high/ unclear)*	low	low

Rationale of applicability rating:

DOMAIN 4: Analysis

Risk of Bias

Describe numbers of participants, number of candidate predictors, outcome events and events per candidate predictor:

30524 (THA); 34,110 (TKA); 81 CPs; EVP > 20

Describe how the model was developed (for example in regards to modelling technique (e.g. survival or logistic modelling), predictor selection, and risk group definition):

ML models

Describe whether and how the model was validated, either internally (e.g. bootstrapping, cross validation, random split sample) or externally (e.g. temporal validation, geographical validation, different setting, different type of participants):

CV and test data; unsure why

Describe the performance measures of the model, e.g. (re)calibration, discrimination, (re)classification, net benefit, and whether they were adjusted for optimism:

C-statistic (main); No calibration

Describe any participants who were excluded from the analysis:

Missing values

Describe missing data on predictors and outcomes as well as methods used for missing data:

Record of all MVs

		Dev	Val
4.1 Were there a reasonable number of participants with the outcome?		Y	Y
4.2 Were continuous and categorical predictors handled appropriately?		PY	PY
4.3 Were all enrolled participants included in the analysis?		N	N
4.4 Were participants with missing data handled appropriately?		N	N
4.5 Was selection of predictors based on univariable analysis avoided?		Y	
4.6 Were complexities in the data (e.g. censoring, competing risks, sampling of controls) accounted for appropriately?		Y	Y
4.7 Were relevant model performance measures evaluated appropriately?		PN	PN
4.8 Were model overfitting and optimism in model performance accounted for?		Y	
4.9 Do predictors and their assigned weights in the final model correspond to the results from multivariable analysis?		Y	
Risk of bias introduced by the analysis	**RISK:** *(low/ high/ unclear)*	high	high

Rationale of bias rating:

~~4.2~~ 4.3, 4.4, 4.7

Step 4: Overall assessment

Use the following tables to reach overall judgements about risk of bias and concerns regarding applicability of the prediction model evaluation (development and/or validation) across all assessed domains.
Complete for each evaluation of a distinct model.

Reaching an overall judgement about risk of bias of the prediction model evaluation	
Low risk of bias	If all domains were rated low risk of bias. If a prediction model was developed without any external validation, and it was rated as low risk of bias for all domains, consider downgrading to **high risk of bias**. Such a model can only be considered as low risk of bias, if the development was based on a very large data set and included some form of internal validation.
High risk of bias	If at least one domain is judged to be at **high risk of bias**.
Unclear risk of bias	If an unclear risk of bias was noted in at least one domain and it was low risk for all other domains.

Reaching an overall judgement about applicability of the prediction model evaluation	
Low concerns regarding applicability	If low concerns regarding applicability for all domains, the prediction model evaluation is judged to have **low concerns regarding applicability**.
High concerns regarding applicability	If high concerns regarding applicability for at least one domain, the prediction model evaluation is judged to have **high concerns regarding applicability**.
Unclear concerns regarding applicability	If unclear concerns (but no "high concern") regarding applicability for at least one domain, the prediction model evaluation is judged to have **unclear concerns regarding applicability** overall.

Overall judgement about risk of bias and applicability of the prediction model evaluation		
Overall judgement of risk of bias	**RISK:** *(low/ high/ unclear)*	HIGH
Summary of sources of potential bias:		
Overall judgement of applicability	**CONCERN:** *(low/ high/ unclear)*	low
Summary of applicability concerns:		

PROBAST

(Prediction model study Risk Of Bias Assessment Tool)

Published in Annals of Internal Medicine (freely available):

1. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies
2. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration

What does PROBAST assess?

PROBAST assesses both the *risk of bias* and *concerns regarding applicability* of a study that evaluates (develops, validates or updates) a multivariable diagnostic or prognostic prediction model. It is designed to assess primary studies included in a systematic review.

Bias occurs if systematic flaws or limitations in the design, conduct or analysis of a primary study distort the results. For the purpose of prediction modelling studies, we have defined *risk of bias* to occur when shortcomings in the study design, conduct or analysis lead to systematically distorted estimates of a model's predictive performance or to an inadequate model to address the research question. Model predictive performance is typically evaluated using calibration, discrimination and sometimes classification measures, and these are likely inaccurately estimated in studies with high risk of bias. *Applicability* refers to the extent to which the prediction model from the primary study matches your systematic review question, for example in terms of the participants, predictors or outcome of interest.

A primary study may include the development and/or validation or update of more than one prediction model. A PROBAST assessment should be completed for each distinct model that is developed, validated or updated (extended) for making individualised predictions. Where a publication assesses multiple prediction models, only complete a PROBAST assessment for those models that meet the inclusion criteria for your systematic review. Please note that subsequent use of the term "model" includes derivatives of models, such as simplified risk scores, nomograms, or recalibrations of models.

PROBAST is not designed for all multivariable diagnostic or prognostic studies. For example, studies using multivariable models to identify predictors associated with an outcome but not attempting to develop a model for making individualised predictions are not covered by PROBAST.

PROBAST includes four steps.

Step	Task	When to complete
1	Specify your systematic review question(s)	Once per systematic review
2	Classify the type of prediction model evaluation	Once for each model of interest in each publication being assessed, for each relevant outcome
3	Assess risk of bias and applicability	Once for each development and validation of each distinct prediction model in a publication
4	Overall judgment	Once for each development and validation of each distinct prediction model in a publication

If this is your first time using PROBAST, we strongly recommend reading the detailed explanation and elaboration (E&E, see link above) paper and to check the examples on www.probast.org

Step 1: Specify your systematic review question

State your systematic review question to facilitate the assessment of the applicability of the evaluated models to your question. *The following table should be completed once per systematic review.*

Criteria	Specify your systematic review question
Intended use of model:	Prediction of MCIDs in PROMs for patients undergoing TKA/THA
***Participants** including selection criteria and setting:*	All patients undergoing total knee or hip arthroplasty
***Predictors** (used in prediction modelling), including types of predictors (e.g. history, clinical examination, biochemical markers, imaging tests), time of measurement, specific measurement issues (e.g., any requirements/ prohibitions for specialized equipment):*	All that are available to the Researchers prior to surgery
Outcome to be predicted:	Whether patients achieve a PROMs MCID or not (classification)

Step 2: Classify the type of prediction model evaluation

Use the following table to classify the evaluation as model development, model validation or model update, or combination. Different signalling questions apply for different types of prediction model evaluation. If the evaluation does not fit one of these classifications then PROBAST should not be used.

Classify the evaluation based on its aim			
Type of prediction study	**PROBAST boxes to complete**	**Tick as appropriate**	**Definition for type of prediction model study**
Development only	Development	X	Prediction model development without external validation. These studies may include internal validation methods, such as bootstrapping and cross-validation techniques.
Development and validation	Development and validation		Prediction model development combined with external validation in other participants in the same article.
Validation only	Validation		External validation of existing (previously developed) model in other participants.

This table should be completed once for each publication being assessed and for each relevant outcome in your review.

Publication reference	Harris et al
Models of interest	All machine learning models
Outcome of interest	MCID

Step 3: Assess risk of bias and applicability

PROBAST is structured as four key domains. Each domain is judged for risk of bias (low, high or unclear) and includes signalling questions to help make judgements. Signalling questions are rated as yes (Y), probably yes (PY), probably no (PN), no (N) or no information (NI). All signalling questions are phrased so that "yes" indicates absence of bias. Any signalling question rated as "no" or "probably no" flags the potential for bias; you will need to use your judgement to determine whether the domain should be rated as "high", "low" or "unclear" risk of bias. The guidance document contains further instructions and examples on rating signalling questions and risk of bias for each domain.

The first three domains are also rated for concerns regarding applicability (low/ high/ unclear) to your review question defined above.

Complete all domains separately for each evaluation of a distinct model. Shaded boxes indicate where signalling questions do not apply and should not be answered.

DOMAIN 1: Participants			
A. Risk of Bias			
Describe the sources of data and criteria for participant selection: - 3 medical centers - Reported follow up - TKA, no revision, bilateral TKA or unicompartmental KA			
		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		Y	
1.2 Were all inclusions and exclusions of participants appropriate?		Y	
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	low	
Rationale of bias rating: No "N"			
B. Applicability			
Describe included participants, setting and dates: - 587 participants - see A			
Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	Low	
Rationale of applicability rating: Patients undergoing TKA and reported 1 year Follow up			

DOMAIN 2: Predictors			
A. Risk of Bias			
List and describe predictors included in the final model, e.g. definition and timing of assessment: 6-106 predictors assessed before surgery			
		Dev	Val
2.1 Were predictors defined and assessed in a similar way for all participants?		Y	
2.2 Were predictor assessments made without knowledge of outcome data?		N	
2.3 Are all predictors available at the time the model is intended to be used?		Y	
Risk of bias introduced by predictors or their assessment	**RISK:** *(low/ high/ unclear)*	low	
Rationale of bias rating: 2.2 "Blinded to the outcome"			
B. Applicability			
Concern that the definition, assessment or timing of predictors in the model do not match the review question	**CONCERN:** *(low/ high/ unclear)*	low	
Rationale of applicability rating: Timing of predictors before surgery - relevant predictors included			

DOMAIN 3: Outcome			
A. Risk of Bias			
Describe the outcome, how it was defined and determined, and the time interval between predictor assessment and outcome determination: MCID anchor - based 1 year			
		Dev	Val
3.1 Was the outcome determined appropriately?		Y	
3.2 Was a pre-specified or standard outcome definition used?		Y	
3.3 Were predictors excluded from the outcome definition?		N	
3.4 Was the outcome defined and determined in a similar way for all participants?		X	
3.5 Was the outcome determined without knowledge of predictor information?		N	
3.6 Was the time interval between predictor assessment and outcome determination appropriate?		Y	
Risk of bias introduced by the outcome or its determination	**RISK:** *(low/ high/ unclear)*	low	
Rationale of bias rating: 3.3. and 3.5. due to the nature of the prediction task.			
B. Applicability			
At what time point was the outcome determined: 1 year post surgery *If a composite outcome was used, describe the relative frequency/distribution of each contributing outcome:* -			
Concern that the outcome, its definition, timing or determination do not match the review question	**CONCERN:** *(low/ high/ unclear)*	low	
Rationale of applicability rating: Outcome assessed at reasonable time horizon			

DOMAIN 4: Analysis			
Risk of Bias			
Describe numbers of participants, number of candidate predictors, outcome events and events per candidate predictor: 587 ; 6-106 ; EVP < 10			
Describe how the model was developed (for example in regards to modelling technique (e.g. survival or logistic modelling), predictor selection, and risk group definition): Predictor selection based on MC methods			
Describe whether and how the model was validated, either internally (e.g. bootstrapping, cross validation, random split sample) or externally (e.g. temporal validation, geographical validation, different setting, different type of participants): CV and bootstrapping			
Describe the performance measures of the model, e.g. (re)calibration, discrimination, (re)classification, net benefit, and whether they were adjusted for optimism: AUC , Brier Score			
Describe any participants who were excluded from the analysis: see Domain 1			
Describe missing data on predictors and outcomes as well as methods used for missing data: No missing data management reported			
		Dev	Val
4.1 Were there a reasonable number of participants with the outcome?		P N	
4.2 Were continuous and categorical predictors handled appropriately?		P Y	
4.3 Were all enrolled participants included in the analysis?		Y	
4.4 Were participants with missing data handled appropriately?		P N	
4.5 Was selection of predictors based on univariable analysis avoided?		Y	
4.6 Were complexities in the data (e.g. censoring, competing risks, sampling of controls) accounted for appropriately?		P Y	
4.7 Were relevant model performance measures evaluated appropriately?		Y	
4.8 Were model overfitting and optimism in model performance accounted for?		P Y	
4.9 Do predictors and their assigned weights in the final model correspond to the results from multivariable analysis?		Y	
Risk of bias introduced by the analysis	**RISK:** *(low/ high/ unclear)*	~~unclear~~ high	
Rationale of bias rating: - Did not report missing value handling - 4.1. accounted for with CV			

Step 4: Overall assessment

Use the following tables to reach overall judgements about risk of bias and concerns regarding applicability of the prediction model evaluation (development and/or validation) across all assessed domains.
Complete for each evaluation of a distinct model.

Reaching an overall judgement about risk of bias of the prediction model evaluation	
Low risk of bias	If all domains were rated low risk of bias. If a prediction model was developed without any external validation, and it was rated as low risk of bias for all domains, consider downgrading to **high risk of bias**. Such a model can only be considered as low risk of bias, if the development was based on a very large data set and included some form of internal validation.
High risk of bias	If at least one domain is judged to be at **high risk of bias**.
Unclear risk of bias	If an unclear risk of bias was noted in at least one domain and it was low risk for all other domains.

Reaching an overall judgement about applicability of the prediction model evaluation	
Low concerns regarding applicability	If low concerns regarding applicability for all domains, the prediction model evaluation is judged to have **low concerns regarding applicability**.
High concerns regarding applicability	If high concerns regarding applicability for at least one domain, the prediction model evaluation is judged to have **high concerns regarding applicability**.
Unclear concerns regarding applicability	If unclear concerns (but no "high concern") regarding applicability for at least one domain, the prediction model evaluation is judged to have **unclear concerns regarding applicability** overall.

Overall judgement about risk of bias and applicability of the prediction model evaluation		
Overall judgement of risk of bias	**RISK:** *(low/ high/ unclear)*	~~Unclear~~ High
Summary of sources of potential bias:		
Overall judgement of applicability	**CONCERN:** *(low/ high/ unclear)*	low
Summary of applicability concerns:		

Kunze et al

PROBAST

(Prediction model study Risk Of Bias Assessment Tool)

Published in Annals of Internal Medicine (freely available):

1. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies
2. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration

What does PROBAST assess?

PROBAST assesses both the *risk of bias* and *concerns regarding applicability* of a study that evaluates (develops, validates or updates) a multivariable diagnostic or prognostic prediction model. It is designed to assess primary studies included in a systematic review.

Bias occurs if systematic flaws or limitations in the design, conduct or analysis of a primary study distort the results. For the purpose of prediction modelling studies, we have defined *risk of bias* to occur when shortcomings in the study design, conduct or analysis lead to systematically distorted estimates of a model's predictive performance or to an inadequate model to address the research question. Model predictive performance is typically evaluated using calibration, discrimination and sometimes classification measures, and these are likely inaccurately estimated in studies with high risk of bias. *Applicability* refers to the extent to which the prediction model from the primary study matches your systematic review question, for example in terms of the participants, predictors or outcome of interest.

A primary study may include the development and/or validation or update of more than one prediction model. A PROBAST assessment should be completed for each distinct model that is developed, validated or updated (extended) for making individualised predictions. Where a publication assesses multiple prediction models, only complete a PROBAST assessment for those models that meet the inclusion criteria for your systematic review. Please note that subsequent use of the term "model" includes derivatives of models, such as simplified risk scores, nomograms, or recalibrations of models.

PROBAST is not designed for all multivariable diagnostic or prognostic studies. For example, studies using multivariable models to identify predictors associated with an outcome but not attempting to develop a model for making individualised predictions are not covered by PROBAST.

PROBAST includes four steps.

Step	Task	When to complete
1	Specify your systematic review question(s)	Once per systematic review
2	Classify the type of prediction model evaluation	Once for each model of interest in each publication being assessed, for each relevant outcome
3	Assess risk of bias and applicability	Once for each development and validation of each distinct prediction model in a publication
4	Overall judgment	Once for each development and validation of each distinct prediction model in a publication

If this is your first time using PROBAST, we strongly recommend reading the detailed explanation and elaboration (E&E, see link above) paper and to check the examples on www.probast.org

Step 1: Specify your systematic review question

State your systematic review question to facilitate the assessment of the applicability of the evaluated models to your question. *The following table should be completed once per systematic review.*

Criteria	Specify your systematic review question
Intended use of model:	Prediction of MCIDs in PROMs for patients undergoing TKA/THA
***Participants** including selection criteria and setting:*	All patients undergoing total knee or hip arthroplasty
***Predictors** (used in prediction modelling), including types of predictors (e.g. history, clinical examination, biochemical markers, imaging tests), time of measurement, specific measurement issues (e.g., any requirements/ prohibitions for specialized equipment):*	All that are available to the Researchers prior to surgery
Outcome to be predicted:	Whether patients achieve a PROMs MCID or not (classification)

Step 2: Classify the type of prediction model evaluation

Use the following table to classify the evaluation as model development, model validation or model update, or combination. Different signalling questions apply for different types of prediction model evaluation. If the evaluation does not fit one of these classifications then PROBAST should not be used.

Classify the evaluation based on its aim			
Type of prediction study	**PROBAST boxes to complete**	**Tick as appropriate**	**Definition for type of prediction model study**
Development only	Development		Prediction model development without external validation. These studies may include internal validation methods, such as bootstrapping and cross-validation techniques.
Development and validation	Development and validation		Prediction model development combined with external validation in other participants in the same article.
Validation only	Validation		External validation of existing (previously developed) model in other participants.

This table should be completed once for each publication being assessed and for each relevant outcome in your review.

Publication reference	Kunze et al 2020
Models of interest	All machine learning models
Outcome of interest	MCID

Step 3: Assess risk of bias and applicability

PROBAST is structured as four key domains. Each domain is judged for risk of bias (low, high or unclear) and includes signalling questions to help make judgements. Signalling questions are rated as yes (Y), probably yes (PY), probably no (PN), no (N) or no information (NI). All signalling questions are phrased so that "yes" indicates absence of bias. Any signalling question rated as "no" or "probably no" flags the potential for bias; you will need to use your judgement to determine whether the domain should be rated as "high", "low" or "unclear" risk of bias. The guidance document contains further instructions and examples on rating signalling questions and risk of bias for each domain.

The first three domains are also rated for concerns regarding applicability (low/ high/ unclear) to your review question defined above.

Complete all domains separately for each evaluation of a distinct model. Shaded boxes indicate where signalling questions do not apply and should not be answered.

DOMAIN 1: Participants			
A. Risk of Bias			
Describe the sources of data and criteria for participant selection: - Institution of the authors - THA - End of stage osteoarthritis - exclusion based on etiology, cost to follow up			
		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		Y	Y
1.2 Were all inclusions and exclusions of participants appropriate?		Y	Y
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	low	low
Rationale of bias rating: No ROB in 1.1 and 1.2.			
B. Applicability			
Describe included participants, setting and dates: see A. 2014 - 2016; 2 years between MCID calculation			
Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	low	low
Rationale of applicability rating: No concerns			

DOMAIN 2: Predictors			
A. Risk of Bias			
List and describe predictors included in the final model, e.g. definition and timing of assessment: 8 (see page 3 of the study). Assessment before surgery.			
		Dev	Val
2.1 Were predictors defined and assessed in a similar way for all participants?		Y	Y
2.2 Were predictor assessments made without knowledge of outcome data?		N	N
2.3 Are all predictors available at the time the model is intended to be used?		Y	Y
Risk of bias introduced by predictors or their assessment	**RISK:** *(low/ high/ unclear)*	low	low
Rationale of bias rating: 2.2. "blinded to the outcome".			
B. Applicability			
Concern that the definition, assessment or timing of predictors in the model do not match the review question	**CONCERN:** *(low/ high/ unclear)*	low	low
Rationale of applicability rating: No concern.			

DOMAIN 3: Outcome			
A. Risk of Bias			
Describe the outcome, how it was defined and determined, and the time interval between predictor assessment and outcome determination: MCID Distribution based			
		Dev	Val
3.1 Was the outcome determined appropriately?		Y	Y
3.2 Was a pre-specified or standard outcome definition used?		Y	Y
3.3 Were predictors excluded from the outcome definition?		N	N
3.4 Was the outcome defined and determined in a similar way for all participants?		Y	Y
3.5 Was the outcome determined without knowledge of predictor information?		N	N
3.6 Was the time interval between predictor assessment and outcome determination appropriate?		Y	Y
Risk of bias introduced by the outcome or its determination	**RISK:** *(low/ high/ unclear)*	low	low
Rationale of bias rating: 3.3 and 3.5 due to the nature of the prediction problem.			
B. Applicability			
At what time point was the outcome determined: Before surgery			
If a composite outcome was used, describe the relative frequency/distribution of each contributing outcome: –			
Concern that the outcome, its definition, timing or determination do not match the review question	**CONCERN:** *(low/ high/ unclear)*	low	low
Rationale of applicability rating: No concern			

DOMAIN 4: Analysis

Risk of Bias

Describe numbers of participants, number of candidate predictors, outcome events and events per candidate predictor:

616 ; 9 ; EVP > 10.

Describe how the model was developed (for example in regards to modelling technique (e.g. survival or logistic modelling), predictor selection, and risk group definition):

Recursive feature selection.

Describe whether and how the model was validated, either internally (e.g. bootstrapping, cross validation, random split sample) or externally (e.g. temporal validation, geographical validation, different setting, different type of participants):

CV ; test dataset

Describe the performance measures of the model, e.g. (re)calibration, discrimination, (re)classification, net benefit, and whether they were adjusted for optimism:

AUC , calibration , brier score

Describe any participants who were excluded from the analysis:

see Domain 1

Describe missing data on predictors and outcomes as well as methods used for missing data:

> 30% missing data per variable → exclusion. Otherwise multiple imputation.

		Dev	Val
4.1 Were there a reasonable number of participants with the outcome?		Y	Y
4.2 Were continuous and categorical predictors handled appropriately?		Y	Y
4.3 Were all enrolled participants included in the analysis?		Y	Y
4.4 Were participants with missing data handled appropriately?		PY	Y
4.5 Was selection of predictors based on univariable analysis avoided?		Y	
4.6 Were complexities in the data (e.g. censoring, competing risks, sampling of controls) accounted for appropriately?		Y	Y
4.7 Were relevant model performance measures evaluated appropriately?		Y	Y
4.8 Were model overfitting and optimism in model performance accounted for?		Y	
4.9 Do predictors and their assigned weights in the final model correspond to the results from multivariable analysis?		Y	
Risk of bias introduced by the analysis	**RISK:** *(low/ high/ unclear)*	low	low

Rationale of bias rating:

accounted for overfitting ~~and therefore~~ due to CV and test data.

Step 4: Overall assessment

Use the following tables to reach overall judgements about risk of bias and concerns regarding applicability of the prediction model evaluation (development and/or validation) across all assessed domains.
Complete for each evaluation of a distinct model.

Reaching an overall judgement about risk of bias of the prediction model evaluation	
Low risk of bias	If all domains were rated low risk of bias. If a prediction model was developed without any external validation, and it was rated as low risk of bias for all domains, consider downgrading to **high risk of bias**. Such a model can only be considered as low risk of bias, if the development was based on a very large data set and included some form of internal validation.
High risk of bias	If at least one domain is judged to be at **high risk of bias**.
Unclear risk of bias	If an unclear risk of bias was noted in at least one domain and it was low risk for all other domains.

Reaching an overall judgement about applicability of the prediction model evaluation	
Low concerns regarding applicability	If low concerns regarding applicability for all domains, the prediction model evaluation is judged to have **low concerns regarding applicability**.
High concerns regarding applicability	If high concerns regarding applicability for at least one domain, the prediction model evaluation is judged to have **high concerns regarding applicability**.
Unclear concerns regarding applicability	If unclear concerns (but no "high concern") regarding applicability for at least one domain, the prediction model evaluation is judged to have **unclear concerns regarding applicability** overall.

Overall judgement about risk of bias and applicability of the prediction model evaluation		
Overall judgement of risk of bias	**RISK:** *(low/ high/ unclear)*	low
Summary of sources of potential bias:		
Overall judgement of applicability	**CONCERN:** *(low/ high/ unclear)*	low
Summary of applicability concerns:		