

Supplementary Table 4. PROBAST+AI Explanation and Elaboration Light (PROBAST+AI E&E Light).

Cite as: Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ* 2025;388:e082505 doi:10.1136/bmj-2024-082505

MODEL DEVELOPMENT

DOMAIN 1: Participants and data sources
<p>1.1 Were appropriate data sources used?</p> <ul style="list-style-type: none"> • <i>This item addresses whether data origin (provenance) is traceable.</i> • <i>Data sources from open data repositories with insufficient details about how data were collected and on measurement procedures should raise concern.</i> • <i>In data sources wherein sufficient details on participant sampling and measurements are lacking, important issues around fairness may be hidden from view.</i>
<p>1.2 Was an appropriate study design used?</p> <ul style="list-style-type: none"> • <i>For prognostic model development studies, prospective longitudinal cohort designs are generally preferred, as methods on design, conduct and analysis tend to be predefined and consistently applied for all participants. Data from (one or more arms of) randomized treatment trials can also be used; the randomized treatments may need to be addressed to account for any treatment effects. Randomized treatment trials usually have more restricted in/exclusion criteria.</i> • <i>For both prognostic and diagnostic model studies, selective sampling may be used, such as case-cohort or nested case-control designs, in which participants are sampled on their outcome status from a data source (e.g., cohort) of known size. Such selective sampling designs are often used for efficiency reasons (participant and predictor data only needs to be collected for the sampled cases and non-cases). Without appropriate adjustments, prediction models derived from selective sampling designs can provide incorrect (miscalibrated) outcome probabilities, because the baseline risks/hazards and thus the absolute outcome probabilities are incorrect. A correction where a reweighting is applied to the sampled cases and non-cases in the analysis by the inverse sampling fraction, to adjust the results towards the original outcome-class frequency may overcome this (see domain 4, item 4). If the sampling fractions from the original data source or cohort are unknown, such reweighting is impossible, and the baseline risks/hazards and thus outcome probabilities are incorrect.</i> • <i>Participant data sources from existing datasets or routine care or administrative registries, tend to have more data quality issues, such as missing data and difficulties with selecting the right participants for analyses, because the data collection had not been specifically designed for the development and validation of prediction models (no prespecified study design or protocol is used).</i>
<p>1.3 Did the in- and exclusions of study participants result in a representative data set?</p> <ul style="list-style-type: none"> • <i>This item relates to the intended use of the prediction model: for whom the model was developed and whether this matches with the in-/excluded participants. It addresses any potential for selective in-/exclusions or enrolment of participants, that yielded an unrepresentative data set of the population of intended use.</i> • <i>For example, in a diagnostic model development study, excluding participants that were difficult to diagnose (e.g., excluding all obese participants in a study on the diagnostic value of ultrasound) would generally be inappropriate. Or in a prognostic model study, including participants already known to have the outcome at the time of the predictor measurements would generally be inappropriate (e.g., in a study on healthy individuals to predict the development of COVID-19, participants having asymptomatic COVID-19 could be erroneously included if the absence of COVID-19 was based on self-reporting).</i>

<ul style="list-style-type: none"> • <i>This item also addresses fairness, when individuals from marginalised subgroups, e.g., defined by age, sex/gender, race/ethnicity, geographic location or specific health conditions/comorbidities, are not included or explicitly excluded.</i> • <i>Sometimes individuals from marginalized subgroups are explicitly oversampled, often for considerations of fairness. However, this may distort the original predictor distributions in the targeted population, may lead to incorrect predictor-outcome associations or miscalibration of the model predictions, and may thus affect the validity of the estimated outcome probabilities.</i> • <i>Sometimes in- and exclusions are invisible or unknown even to the investigators themselves, particularly when existing data sources or open data repositories were used. The assessor then needs to make an assumption whether the analysed data set was representative or not (the 'No Information' option may be used).</i> • <i>This item does not address participant exclusion due to missing data or loss to follow-up after enrolment (such as partial or differential outcome verification) - that is covered under domain 4.</i>
Applicability Assessment
<p>Concern that the (data of the) included participants do not match the review question or the assessor's intended use of the prediction model</p> <ul style="list-style-type: none"> • <i>Applicability assessment is entirely dependent on the review question or the assessor's intended use of the developed prediction model, defined under step 1. It is important to distinguish between selective in/exclusion imposed on a dataset by implicit or explicit restrictions in the participant recruitment (which is addressed by the quality assessment items of this domain) versus a dataset with participants that just limit its applicability to the review question or assessor's intended use of the model.</i> • <i>For example, consider a review that aims to identify all developed and validated models to diagnose bacterial meningitis in children. Studies that included only children have low applicability concerns, whereas studies in which adults and children were combined receive higher applicability concerns.</i> • <i>Applicability of prediction model studies based on randomized treatment trials need careful applicability consideration, due to, e.g., typically stricter in- and exclusion criteria, fewer predictors, shorter follow-up times.</i>

DOMAIN 2: Predictors
<p>2.1 Were predictors defined and assessed in a similar way for all participants?</p> <ul style="list-style-type: none"> • <i>Predictor definitions, thresholds for categorization, and assessment methods should ideally be the same across all participants. Differences in this between participants, may result in inaccurate predictor-outcome associations and may thus affect the validity of the estimated outcome probabilities.</i> • <i>This potential is higher for predictors requiring subjective judgement or specific measurement skills, such as predictors obtained from imaging, electrophysiology or pathology tests. It may then rather reflect the predictive ability of the measurer than that of the predictors. Where special skills or training is required, specifying who assessed the predictor (e.g., the level of experience of the person doing the measurements) may be important.</i> • <i>This item also addresses fairness, when different predictor definitions or assessments are used across individuals from marginalised subgroups, e.g., defined by age, sex/gender, race/ethnicity, geographic location or specific health conditions/comorbidities.</i>
<p>2.2 Was any pre-processing of predictors similar for all participants?</p> <ul style="list-style-type: none"> • <i>Pre-processing is a typical preparatory step before further analysis. Although predictor pre-processing applies to all types of predictors (e.g., combining different predictors to a single summary predictor, standardization of predictor values), it often applies to predictors that are retrieved from non-structured data, such as from medical images (e.g., from radiology, electrophysiology, pathology, biopsy). This pre-processing should be the same across all participants. Differences in pre-processing of the same predictors, e.g., across different study centres or marginalised subgroups, may result in inaccurate predictor-outcome associations and may thus affect the validity of the estimated outcome probabilities.</i> • <i>Sometimes pre-processing steps are invisible or unknown, even to the investigators themselves, particularly when existing data sources or open data repositories were used. In such situations, the</i>

<p>assessor needs to make an assumption whether pre-processing was indeed similar for all participants or fill in 'No Information'.</p> <ul style="list-style-type: none"> Similar to item 2.1, this item also addresses fairness when pre-processing procedures of the same predictors differ across individuals from marginalised subgroups defined, e.g., by age, sex/gender, race/ethnicity, geographic location or specific health conditions/comorbidities.
<p>2.3 Were predictor assessments made without knowledge of outcome data?</p> <ul style="list-style-type: none"> Predictor assessments are ideally made without knowledge of (also called blinded or masked for) the outcome status. This is particularly important for predictors requiring subjective judgement. Lack of blinding increases risk for incorporating outcome information into predictor assessments, which likely increases the predictor-outcome associations and may thus affect the validity of the estimated outcome. Blinding predictor assessors to outcome information may in particular occur in studies that retrospectively assess predictors (e.g., when using or reinterpreting stored imaging tests or frozen tissue samples to measure novel biomarkers) where the outcome is already known or assessed. Or in cross-sectional (diagnostic model) studies where predictors and outcomes are both assessed in a relatively short time period. Studies often do not explicitly report information on blinding predictor assessments to outcome data. This item may then be answered as 'No Information', though the overall domain can still be rated as high quality certainly when predictors were measured clearly preceding the outcome assessment.
<p>2.4 Were the predictors included in the model available at the time the model was intended to be used?</p> <ul style="list-style-type: none"> For a model to be used in practice it is necessary that all included predictors can potentially be assessed at the moment the model is intended to be used (i.e., at the moment of prediction). This may seem straightforward, but sometimes prediction models include predictor information that cannot be known at the time the model is to be used. For example, a prognostic model that is intended to be used preoperatively to predict post-operative nausea and vomiting within 72 hours, should not include predictors that become available intra- or post-operatively (such as intra-operatively applied medication). Or a diagnostic model to be used in primary care after history taking and physical examination, should not include predictors that become available after subsequent laboratory or imaging tests. This item does not address whether all possible predictors (of the targeted outcome) were indeed present in the dataset nor whether all possible predictors were eventually included in the finalized prediction model. It addresses whether the predictors that were finally included in the developed model, can indeed be obtained at the moment the model is intended to be used. This item does not address missing predictors due to missing data - that is covered under domain 4.
<p>Applicability Assessment</p> <p>Concern that the definition, pre-processing, assessment, or timing of assessment of the predictors in the model do not match the review question or the assessor's intended use</p> <ul style="list-style-type: none"> Applicability assessment is directed by the review question or the assessor's intended use of the prediction model, defined under step 1. Concerns on applicability in this domain addresses inconsistencies between definition, assessment and timing of predictors given the review question or assessor's intended use. For example, predictors in the model should be defined and measured using methods potentially applicable to the targeted setting or context of the reviewer or assessor. Studies that used specialized measurement techniques for predictors may yield less applicable models for the targeted setting. Similar to domain 1, sometimes, there is a subtle distinction between quality versus applicability assessment. Consider a diagnostic model including the D-dimer (a biomarker in blood) measurement as a predictor for the presence of pulmonary embolism. D-dimer can be assessed by a rapid point-of-care test providing a dichotomous (high or low) test result or by a quantitative ELISA laboratory test. A review might be focused only on the diagnostic models that used a quantitative ELISA D-dimer measurement, thus, studies that used a dichotomous rapid point-of-care D-dimer test should raise applicability concerns (and may be excluded). However, it would not be an applicability concern if the

aim of the review or assessor is to include all developed models for a specific targeted population, outcome or setting, regardless of the definition and measurement methods of its predictors.

DOMAIN 3: Outcome

3.1 Were outcomes defined and assessed appropriately?

- *This item addresses potential errors in outcome measurements (misclassification/measurement error) due to the use of non-standard outcome definitions or inferior outcome assessment methods, which may lead to incorrect predictor-outcome associations or miscalibration in the large and may thus affect the validity of the estimated outcome probabilities.*
- *This potential for error is lower when prespecified or standard outcome definitions and assessment methods (e.g., supported by medical guidelines, published studies, or study protocols), and may be higher for outcomes requiring subjective judgement or specific measurement skills (alike for predictors, see domain 2). For example, presence/absence of diabetes mellitus type 2 can be assessed preferably by fasting glucose levels in blood, but also by less accurate oral glucose tolerance tests or even by self-reporting.*
- *Practices that may be flagged as concerning are, e.g., the use of investigator-chosen thresholds on a continuous outcome scale to define an outcome as present versus absent, or when multiple data-driven thresholds are tested to select the largest outcome prevalence (in diagnostic model studies) or incidence (in prognostic model studies), or methods to obtain the largest predictor-outcome associations.*
- *Composite outcomes require special attention, certainly when investigators may have in-/excluded particular component-outcomes, e.g., to obtain the largest outcome prevalence or incidence.*
- *When open data repositories, routine care or administrative healthcare or disease registries are used, scoring this item requires special attention, as outcome definitions and measurement methods may be different (less optimal thus yielding more measurement error) than one would apply in a predesigned prediction model study.*
- *Many outcomes in biomedical research have consensus-based outcome definitions, thresholds, and measurement methods (see, e.g., Core Outcome Measures in Effectiveness Trials; www.cometinitiative.org).*

3.2 Were outcomes defined and assessed in a similar way for all participants?

- *Outcome definitions, thresholds, measurement methods (including for example the number of follow-up visits for outcome assessment) should be the same across all participants. Differences between participants, e.g., across different centres or marginalised subgroups, may result in inaccurate outcome prevalence/incidence, baseline risks/hazards and predictor-outcome associations. As applies to predictors (domain 2), this potential is higher for outcomes requiring subjective judgement or specific measurement skills.*
- *This potential for error is also higher when data were not obtained from predesigned studies but from open data repositories or routine care or administrative data registries, where typically different outcome definitions and assessment methods may be applied across participants.*
- *For diagnostic (model) studies this phenomenon is well known, where investigators sometimes explicitly do not or cannot apply the same outcome definition or assessment method in each participant. Often, a reference standard is only applied to participants who have positive results on a particular predictor (index test). This may lead to either partial outcome verification (outcome data are missing for those with, e.g., negative result on a particular index test) or differential outcome verification (participants not referred to the preferred reference standard are assessed by an alternative reference standard of usually lower accuracy). Methods to account for partial and differential verification have been described elsewhere.*
- *This item also addresses fairness when definition and assessment of the same outcome differs across individuals of marginalised subgroups defined, e.g., by different ages, sexes/genders, races/ethnicities, geographic locations or health conditions/comorbidities.*

3.3 Were outcome assessments made without use or knowledge of predictor data?

- *Outcome assessments are ideally made without use or knowledge of (blinded or masked for) the predictor values (i.e., predictor information should not be leaked into the outcome assessments). This is again particularly important for outcomes requiring subjective judgement. One should carefully judge whether or not predictor information was available to the outcome assessors.*

<ul style="list-style-type: none"> • <i>If predictor information is known by the outcome assessor or forms part of the outcome definition or assessment method, predictor information may be partly or is completely, respectively, incorporated in the outcome assessment, which can lead to spurious predictor-outcome associations and inflated estimates of predictive performance of a prediction model.</i> • <i>Ensuring that outcome assessors are masked for the information of the predictors under study can be prevented in most cases in predesigned (prospective) diagnostic and prognostic model studies. It can be difficult to prevent in studies that retrospectively assess outcomes, and notably in studies using routine care or real-world data sets. In daily practice, masked or independent outcome assessment is atypical.</i> • <i>Sometimes a masked outcome assessment is not possible. For example, when outcome assessment is done by a clinical expert, expert consensus or an outcome-adjudication panel where all participant information, including that of the predictors under study, may be used in the outcome assessment. This can occur in both diagnostic and prognostic model studies and is typically used for outcomes that are difficult to assess by a single measurement test or method.</i>
<p>3.4 Was the time interval between predictor assessment and outcome assessment appropriate?</p> <ul style="list-style-type: none"> • <i>This item addresses whether the time interval between predictor and outcome assessment is too short or too long, which requires subject matter knowledge.</i> • <i>In diagnostic model studies, assessment of predictors and outcome should happen ideally at the ‘same’ (cross-sectional) time point. Sometimes this is not feasible, and the targeted outcome may diminish over time and thus become undetected (even though it was originally present at the moment of predictor assessment). This may typically occur if the delay is too long, and for acute or self-limiting diseases. Sometimes patient follow-up over time is used as the reference standard, such that a time interval between diagnostic predictor and outcome assessment is pre-designed. This is generally less problematic for chronic diseases than for acute and self-limiting diseases. Sometimes tissue samples for predictor and outcome assessment are taken (and stored) at the same time point.</i> • <i>In prognostic studies, the time interval between predictor and outcome assessment can also be too short for the outcome to have been developed or expressed since the predictor assessment, or too long and may have been missed at the time of the outcome assessment, for instance in case of self-limiting diseases.</i>
<p>Applicability Assessment</p>
<p>Concern that the outcome, its definition, assessment, or timing of assessment do not match the review question or the assessor’s intended use</p> <ul style="list-style-type: none"> • <i>Applicability assessment is directed by the review question or the assessor’s intended use of the prediction model, defined under step 1.</i> • <i>Concerns on applicability in this domain addresses inconsistencies between definition, method and timing of outcome assessment given the review question or assessor’s intended use. Outcomes should be defined and assessed at the proper time intervals, using methods potentially applicable to the targeted context of the review or assessor. For example, a review may specifically aim to determine a short- or long-term prognosis, making the judgement of the proper time interval between predictor and outcome assessment relevant for applicability judgements. Or a study might have used a composite outcome that consisted of components different from the components of the outcome definition of the review question.</i> • <i>Similar to domain 1 and 2, a subtle distinction between quality versus applicability assessment may arise. Consider a review of prognostic models for predicting the occurrence of major depressive disorders. The presence of depressive disorder may be assessed using the comprehensive depression section of the Composite International Diagnostic Interview (CIDI). However, studies may also have used the Patient Health Questionnaire-9 (PHQ-9), a reduced version of this PHQ (called the PHQ-2), the mental health section of the SF-36 questionnaire, or even used self-reported presence of major depression. One might then focus the review on only those studies that used the CIDI approach, defining all other studies as not applicable. Alternatively, there is no applicability concern, if the explicit aim of the review or assessor is to include all developed models for major depression, regardless of the outcome assessment method.</i>

DOMAIN 4: Analysis

4.1 Was there evidence that the sample size was reasonable?

- *In general: the larger the sample size, the better. Larger samples yield more precisely estimated predictor-outcome associations, decreased risk of overfitting, and lead to more stably estimated predictions.*
- *The adequacy of sample size should be considered in light of model complexity (higher number of parameters to be estimated generally requires larger sample size), the anticipated strength of predictor-outcome associations (weaker predictor-outcome associations generally need a larger sample size) and for dichotomous outcomes also the fraction of events (event fraction further away from 0.5 generally requires a larger sample size). Minimal sample size criteria for prediction models with a continuous, dichotomous, multinomial or survival outcome are currently available (1-3).*
- *Regularization and shrinkage methods can be used to lower the risk of overfitting and to reduce model complexity. It should be noted that applying these methods is not a substitute for adequate sample size and may result in unstable predictions and miscalibration when the prediction model is applied to new patients (4-6).*
- *We recommend reviewers to be careful with applying rules of thumb to judge adequate sample size (e.g., the 1 parameter per 10 events rule). However, in the absence of minimal sample size criteria used by developers or other information to judge prediction stability and overfitting (e.g., prediction stability plots, learning curves, precisely estimated calibration slopes), reviewers may still wish to judge the adequacy of the used sample size. Reviewers may therefore wish to apply more liberal sample size criteria for their judgement or in the absence of information on adequacy of the sample size provided by the original developers mark this item as 'unclear'.*
- *We recognize the increased complexity of judging the adequacy of sample size for models other than generalized linear models and survival models. Some AI-based prediction models, for which it is often difficult to evaluate model complexity, and the number of independent parameters estimated, may require huge sample sizes to generate stable predictions and to minimize the risk of overfitting, especially when hyperparameters are not carefully tuned. In the absence of information on adequacy of the sample size provided by the original developers mark this item as 'unclear'.*

4.2 Were continuous and categorical predictors handled appropriately?

- *Categorizing (e.g., dichotomizing) continuous predictors or deleting or collapsing categories of categorical predictors (e.g., deleting categories with less- or uninterpretable findings) before entering them in the model development may lead to an avoidable loss of information (7).*
- *Data-driven categorization of variables, especially those that maximize the apparent predictive performance of the model should be avoided. These may create inflated and spurious predictor-outcome associations and may lead to overfitted prediction models (8).*
- *Continuous predictors for regression models are ideally modelled in a flexible manner to allow for non-linear associations, e.g., using spline functions or fractional polynomials.*
- *In tree-based learning (e.g., Random Forest), continuous data may be essentially treated as categorical. However, as these approaches may average over many different categorizations (e.g., in bootstrap samples), such approaches must not be judged in the same way as described above and may thus not be judged as indications of low quality.*
- *This item also addresses fairness when handling predictors differs across individuals of marginalised subgroups defined, e.g., by different ages, sexes/genders, races/ethnicities, geographic locations or health conditions/comorbidities.*

4.3 Were participants with missing or censored data handled appropriately in the analysis?

- *In general, selective exclusions of eligible participants in a study, dataset or analysis should be avoided. For instance, participants should generally not be excluded because of missing values on predictors or outcomes.*
- *In general, the higher the percentage of missing or censored data, the higher the risks of a remaining selective analysed subset. However, it should be noted that even small amounts of missing or censored data may have a profound impact on the prediction model (in particular when missing data are 'missing*

not at random' (MNAR)). It is therefore not feasible to provide maximum percentages of missing or censored data that can be considered acceptable. It may be hard to judge the validity of prediction models developed on data where a small number of participants are left out due to missing values. For appropriate handling of missing data in prediction model development we refer to the literature.

- Multiple imputation approaches are often considered preferred approaches to deal with missing predictor data. In some specific cases, handling of alternative missing data handling approaches, e.g., using missing indicator methods and tree-based learning approaches that use surrogate splitting are sometimes appropriate in a prediction modelling setting under strong assumptions (9).
- Approaches to adequately deal with censored data in the development of prediction models, such as competing risk modelling, have been described in the literature. Especially prediction models with outcomes other than all-cause mortality, and with long or varying times of follow-up and subsequent large amounts of (informative) censoring, may require the use of competing risk approaches.
- It is sometimes not clear from reports whether study participants with any missing or censored data were excluded from analyses (i.e., complete-case or available-case analysis). Although one may score this item then formally with 'No Information', one may still judge this aspect as 'low quality' since most data analysis techniques automatically exclude records with any missing value on any of the predictors or outcomes.
- This item also addresses fairness when handling of missing or censored data differs across individuals of marginalised subgroups defined, e.g., by different ages, sexes/genders, races/ethnicities, geographic locations or health conditions/comorbidities.

4.4 If methods to address class imbalance were used, was the model or the model predictions recalibrated?

- Approaches to correct for imbalance in the outcome status (i.e., an event fraction away from 0.5) is increasingly common in prediction research, particularly in AI-based prediction. These class imbalance corrections, such as undersampling of the majority class (i.e., random selection from the group with the highest prevalence), oversampling of the minority class (i.e., duplicating from the group with the lowest prevalence), may do more harm than good if the interest is in risk prediction. This also holds for more advanced approaches, such as SMOTE.
- The issues with class imbalance corrections are similar to those in case-control designs, discussed under domain 1 item 3. In particular if the interest is in providing well calibrated risk predictions it is important to adjust the model predictions (e.g., by reweighting cases or re-estimating the intercept). However, unlike case-control designs, class imbalance corrections are not applied for statistical efficiency reasons, but typically used with the aim to improve prediction model classification. Current literature suggests the benefits of such approaches for risk prediction models are typically small or even harmful to model performance.

4.5 Were methods used to address potential model overfitting?

- An overfitted predicted model can be expected to have suboptimal performance when applied to new individuals and should therefore generally be avoided.
- Approaches to avoid overfitting include (but are not limited to): ensure the data is large enough relative to model complexity (see also item 4.1), avoid data driven predictor selection without penalization (e.g., univariable screening, forward selection) as this will increase the chances of introducing the 'winner's curse', and by applying regularization / shrinkage methods.
- Regularization and shrinkage methods are of particular importance for complex models. For more advanced AI-based prediction models, tuning the hyperparameters involved in the regularization is of critical importance. As pointed out in 4.1., regularization is not a substitute for adequate sample size. Even carefully regularized models (e.g., models with appropriate shrinkage approaches) may become very unstable in small datasets, and thus may perform badly when applied to new patients.
- We recognize that the adequacy of used approaches to address overfitting may be complex to judge, especially for highly complex AI-based models.
- In general, studies with a very large sample size relative to the model complexity may have no benefit from approaches to address overfitting as the risk of overfitting is generally small.

MODEL EVALUATION

DOMAIN 1: Participants and data sources
<p>1.1 Were appropriate data sources used?</p> <ul style="list-style-type: none"> • <i>This item addresses whether data origin (provenance) is traceable.</i> • <i>Data sources from open data repositories with insufficient details about how data were collected, and measurement procedures should raise concern.</i> • <i>Data sources wherein sufficient details on participant sampling and measurements are lacking, important issues around fairness may be hidden from view.</i>
<p>1.2 Was an appropriate study design used?</p> <ul style="list-style-type: none"> • <i>For prognostic model evaluation/validation studies, prospective longitudinal cohort designs are generally preferred, as methods on design, conduct and analysis tend to be predefined and consistently applied for all participants. Data from (one or more arms of) randomized treatment trials can also be used; the randomized treatments may need to be addressed to account for any treatment effects. Randomized treatment trials usually have more restricted in/exclusion criteria.</i> • <i>For both prognostic and diagnostic model studies, selective sampling, such as case-cohort or nested case-control designs, in which participants are sampled on their outcome status from a data source (e.g., cohort) of known size. Such selective sampling designs are often used for efficiency reasons (participant and predictor data only needs to be collected for the sampled cases and non-cases). Without appropriate adjustments, prediction models derived from selective sampling designs can provide incorrect (miscalibrated) outcome probabilities, because the baseline risks/hazards and thus the absolute outcome probabilities are incorrect. A correction where a reweighting is applied to the sampled cases and non-cases in the analysis by the inverse sampling fraction, to adjust the results towards the original outcome-class frequency may overcome this (see domain 4, item 4). If the sampling fractions from the original data source or cohort are unknown, such reweighting is impossible, and the baseline risks/hazards and thus outcome probabilities are incorrect.</i> • <i>Participant data sources from existing datasets or routine care or administrative registries, tend to have more data quality issues, such as missing data and difficulties with selecting the right participants for analyses, because participant data collection has typically not been designed for the development and validation of prediction models (no prespecified study design or protocol is used).</i>
<p>1.3 Did the in- and exclusions of study participants result in a representative data set?</p> <ul style="list-style-type: none"> • <i>This item relates to the intended use of the prediction model: for whom the model was validated/evaluated for and whether this matches with the in-/excluded participants. It addresses any potential for selective in-/exclusions or enrolment of participants, that yielded an unrepresentative data set of the population of intended use.</i> • <i>For example, in a diagnostic model evaluation/validation study, excluding participants that were difficult to diagnose (e.g., excluding all obese participants in a study on the diagnostic value of ultrasound) would generally be inappropriate. Or in a prognostic model study, including participants already known to have the outcome at the time of the predictor measurements would generally be inappropriate (e.g., in a study on healthy individuals to predict the development of Covid-19, participants having asymptomatic Covid-19 could be erroneously included if the absence of Covid-19 was based on self-reporting).</i> • <i>This item also addresses fairness, when individuals from marginalised subgroups, e.g., defined by age, sex/gender, race/ethnicity, geographic location or specific health conditions/comorbidities, are not included or explicitly excluded.</i> • <i>Sometimes individuals from marginalized subgroups are explicitly oversampled, often for considerations of fairness. However, this may distort the original predictor distributions in the targeted population, may lead to incorrect predictor-outcome associations or miscalibration of the model predictions, and may thus affect the validity of the estimated outcome probabilities.</i>

- Sometimes in- and exclusions are invisible or unknown even to the investigators themselves, particularly when existing data sources or open data repositories were used. The assessor then needs to make an assumption whether the analysed data set was representative or not (the 'No Information' option may be used).
- This item does not address participant exclusion due to missing data or loss to follow-up after enrolment - that is covered under domain 4.

Applicability Assessment

Concern that the included (data of the) participants and setting do not match the review question or the assessor's intended use

- Applicability assessment is entirely dependent on the review question or the assessor's intended use of the evaluated/validated prediction model, defined under step 1. It is important to distinguish between selective in/exclusion imposed on a dataset by implicit or explicit restrictions in the participant recruitment (which is addressed by the quality assessment items of this domain) versus a dataset with participants that just limit its applicability to the review question or assessor's intended use of the model.
- For example, consider a review that aims to identify all developed and validated models to diagnose bacterial meningitis in children. Studies that included only children have low applicability concerns, whereas studies in which adults and children were combined receive higher applicability concerns.
- Applicability of prediction model studies based on randomized treatment trials need careful applicability consideration, due to, e.g., typically stricter in- and exclusion criteria, fewer predictors, shorter follow-up times.
- External validation studies do not have the purpose to compare or find similar performance as found in the development study. They just aim to quantify the performance of a previously developed model in other participant data than from which it was developed.
- We emphasize that primary studies sometimes evaluate a previously developed model in participant data that were intentionally different from the participants used in the development study. For example, cardiovascular prognostic prediction models developed from data of a healthy general population have been intentionally evaluated in patients diagnosed with type 2 diabetes mellitus (that were not represented in the original development study (training data set)). Or a model to diagnose deep venous thrombosis that was developed in an emergency secondary care setting was explicitly evaluated in a primary care setting. In both situations, model evaluation was explicitly chosen in a totally different population or setting than the development study, and thus heterogeneity in model performance between the development study and the evaluation study is expected.

DOMAIN 2: Predictors

2.1 Were predictors defined and assessed in a similar way for all participants?

- Predictor definitions, thresholds for categorization, and assessment methods should be the same across all participants. Differences between participants may result in inaccurate predictor-outcome associations and may thus affect the validity of the estimated outcome probabilities.
- This potential is higher for predictors requiring subjective judgement or specific measurement skills, such as predictors obtained from imaging, electrophysiology or pathology tests. It may then rather reflect the predictive ability of the measurer than that of the predictors. Where special skills or training is required, specifying who assessed the predictor (e.g., the level of experience of the person doing the measurements) may be important.
- This item also addresses fairness, when different predictor definitions or assessments are used across individuals from marginalised subgroups, e.g., defined by age, sex/gender, race/ethnicity, geographic location or specific health conditions/comorbidities.

2.2 Was any pre-processing of predictors similar for all participants?

- Pre-processing is a typical preparatory step before further analysis. Although predictor pre-processing applies to all types of predictors (e.g., combining different predictors to a single summary predictor,

<p>standardization of predictor values), it often applies to predictors that are retrieved from non-structured data, such as from medical images (e.g., from radiology, electrophysiology, pathology, biopsy). This pre-processing should be the same across all participants. Differences in pre-processing of the same predictors, e.g., across different study centres or marginalised subgroups, may result in inaccurate predictor-outcome associations and may thus affect the validity of the estimated outcome probabilities.</p> <ul style="list-style-type: none"> • Sometimes pre-processing steps are invisible or unknown, even to the investigators themselves, particularly when existing data sources or open data repositories were used. In such situations, the assessor needs to make an assumption whether pre-processing was indeed similar for all participants or fill in 'No Information'.
<p>2.3 Were predictor assessments made without knowledge of outcome data?</p> <ul style="list-style-type: none"> • Predictor assessments are ideally made without knowledge of (also called blinded or masked for) the outcome status. This is particularly important for predictors requiring subjective judgement. Lack of blinding increases risk for incorporating outcome information into predictor assessments, which likely increases the predictor-outcome associations and may thus affect the validity of the estimated outcome. • Blinding predictor assessors to outcome information may in particular occur in studies that retrospectively assess predictors (e.g., when using or reinterpreting stored imaging tests or frozen tissue samples to measure novel biomarkers) where the outcome is already known or assessed. Or in cross-sectional (diagnostic model) studies where predictors and outcomes are both assessed in a relatively short time period. • Studies often do not explicitly report information on blinding predictor assessments to outcome data. This item may then be answered as 'No Information', though the overall domain can still be rated as high quality certainly when predictors were measured clearly preceding the outcome assessment.
<p>2.4 Were the predictors included in the model available at the time the model was intended to be used?</p> <ul style="list-style-type: none"> • For a model to be used in practice it is necessary that all included predictors can potentially be assessed at the moment the model is intended to be used (i.e., at the moment of prediction). This may seem straightforward, but sometimes prediction models may include predictor information that cannot be known at the time the model is to be used. • For example, a prognostic model that is intended to be used preoperatively to predict post-operative nausea and vomiting within 72 hours, should not include predictors that become available intra- or post-operatively (such as intra-operatively applied medication). Or a diagnostic model to be used in primary care after history taking and physical examination, should not include predictors that become available after subsequent laboratory or imaging tests. • This item addresses whether the predictors in the model that is being evaluated, can indeed be obtained at the moment the model is intended to be used. It does not address whether all possible predictors (of the targeted outcome) were indeed present in the dataset nor whether all possible predictors were eventually included in the finalized prediction model. • This item does also not address missing predictors due to missing data - that is covered under domain 4.
<p>Applicability Assessment</p>
<p>Concern that the definition, pre-processing, assessment, or timing of assessment of the predictors in the model do not match the review question or the assessor's intended use</p> <ul style="list-style-type: none"> • Applicability assessment is directed by the review question or the assessor's intended use of the prediction model, defined under step 1. • Concerns on applicability in this domain addresses inconsistencies between definition, assessment and timing of predictors given the review question or assessor's intended use. For example, predictors in the model should be defined and measured using methods potentially applicable to the targeted setting or context of the reviewer or assessor. Studies that used specialized measurement techniques for predictors may yield less applicable models for the targeted setting. • Similar to domain 1, sometimes, there is a subtle distinction between quality versus applicability assessment. Consider a diagnostic model including the D-dimer (a biomarker in blood) measurement as a predictor for the presence of pulmonary embolism. D-dimer can be assessed by a rapid point-of-care test providing a dichotomous (high or low) test result or by a quantitative ELISA laboratory test. A review might be focused only on the diagnostic models that used a quantitative ELISA D-dimer

measurement, thus, studies that used a dichotomous rapid point-of-care D-dimer test should raise applicability concerns (and may be excluded). However, it would not be an applicability concern if the aim of the review or assessor is to include all evaluated/validated models for a specific targeted population, outcome or setting, regardless of the definition and measurement methods of its predictors.

- Similarly, as in domain 1, in reviews that aim to review and meta-analyse the performance of a specific model, heterogeneity in the observed performance of that model between the original development study and the various validation studies is expected due to differences in the definition and measurement of predictors. This does not necessarily mean a concern for applicability of such studies.
- Moreover, sometimes researchers intentionally apply different definitions or measurement methods—for example, explicitly changing a laboratory measurement to a point-of-care measurement for certain blood values. This might not be an applicability problem if the explicit aim of the systematic review is to include all validations of a particular model regardless of the definition and measurement methods of its predictors.

DOMAIN 3: Outcome

3.1 Were outcomes defined and assessed appropriately?

- This item addresses potential errors in outcome measurements (misclassification/measurement error) due to the use of non-standard outcome definitions or inferior outcome assessment methods, which may lead to incorrect predictor-outcome associations or miscalibration in the large and may thus affect the validity of the estimated outcome probabilities.
- This potential for error is lower when prespecified or standard outcome definitions and assessment methods (e.g., supported by medical guidelines, published studies, or study protocols), and may be higher for outcomes requiring subjective judgement or specific measurement skills (alike for predictors, see domain 2). For example, presence/absence of diabetes mellitus type 2 can be assessed preferably by fasting glucose levels in blood, but also by less accurate oral glucose tolerance tests or even by self-reporting.
- Practices that may be flagged as concerning are, e.g., the use of investigator-chosen thresholds on a continuous outcome scale to define an outcome as present versus absent, or when multiple data-driven thresholds are tested to select the largest outcome prevalence (in diagnostic model studies) or incidence (in prognostic model studies), or methods to obtain the largest predictor-outcome associations.
- Composite outcomes require special attention, certainly when investigators may have in-/excluded particular component-outcomes, e.g., to obtain the largest outcome prevalence or incidence.
- When open data repositories, routine care or administrative data registries are used, scoring this item requires special attention, as outcome definitions and measurement methods may be different (less optimal thus yielding more measurement error) than one would apply in a predesigned prediction model study.
- Many outcomes in biomedical research have consensus-based outcome definitions, thresholds, and measurement methods (see, e.g., Core Outcome Measures in Effectiveness Trials; www.cometinitiative.org).

3.2 Were outcomes defined and assessed in a similar way for all participants?

- Outcome definitions, thresholds, measurement methods (including for example the number of follow-up visits for outcome assessment) should be the same across all participants. Differences between participants, e.g., across different centres or marginalised subgroups, may result in inaccurate outcome prevalence/incidence, baseline risks/hazards and predictor-outcome associations. As applies to predictors (domain 2), this potential is higher for outcomes requiring subjective judgement or specific measurement skills.
- This potential for error is also higher when data were not obtained from predesigned studies but from open data repositories or routine care or administrative data registries, where typically different outcome definitions and assessment methods may be applied across participants.
- For diagnostic (model) studies this phenomenon is well known, where investigators sometimes explicitly do not or cannot apply the same outcome definition or assessment method in each participant. Often, a reference standard is only applied to participants who have positive results on a particular predictor (index test). This may lead to either partial outcome verification (outcome data are missing for those with, e.g., negative result on a particular index test) or differential outcome verification (participants

<p><i>not referred to the preferred reference standard are assessed by an alternative reference standard of usually lower accuracy). Methods to account for partial and differential verification have been described elsewhere.</i></p> <ul style="list-style-type: none"> <i>• This item also addresses fairness when definition and assessment of the same outcome differs across individuals of marginalised subgroups defined, e.g., by different ages, sexes/genders, races/ethnicities, geographic locations or health conditions/comorbidities.</i>
<p>3.3 Were outcome assessments made without use of predictor data?</p> <ul style="list-style-type: none"> <i>• Outcome assessments are ideally made without use or knowledge of (blinded or masked for) the predictor values (i.e., predictor information should not be leaked into the outcome assessments). This is again particularly important for outcomes requiring subjective judgement. One should carefully judge whether or not predictor information was available to the outcome assessors.</i> <i>• If predictor information is known or forms part of the outcome definition or assessment method, predictor information may be (partly) incorporated in the outcome assessment, which can lead to spurious predictor-outcome associations and inflated estimates of predictive performance of a prediction model.</i> <i>• Ensuring that outcome assessors are masked for the information of the predictors under study can be prevented in most cases in predesigned (prospective) diagnostic and prognostic model studies. It can be difficult to prevent in studies that retrospectively assess outcomes, and notably in studies using routine care or real-world data sets. In daily practice, masked or independent outcome assessment is atypical.</i> <i>• Sometimes a masked outcome assessment is not possible. For example, when outcome assessment is done by a clinical expert, expert consensus or an outcome-adjudication panel where all participant information, including that of the predictors under study, may be used in the outcome assessment. This can occur in both diagnostic and prognostic model studies and is typically used for outcomes that are difficult to assess by a single measurement test or method.</i>
<p>3.4 Was the time interval between predictor assessment and outcome assessment appropriate?</p> <ul style="list-style-type: none"> <i>• This item addresses whether the time interval between predictor and outcome assessment is too short or too long, which requires subject matter knowledge.</i> <i>• In diagnostic model studies, assessment of predictors and outcome should happen ideally at the ‘same’ (cross-sectional) time point. Sometimes this is not feasible, and the targeted outcome may diminish over time and thus become undetected (even though it was originally present at the moment of predictor assessment). This may typically occur if the delay is too long, and for acute or self-limiting diseases. Sometimes patient follow-up over time is used as the reference standard, such that a time interval between diagnostic predictor and outcome assessment is pre-designed. This is generally less problematic for chronic diseases than for acute and self-limiting diseases. Sometimes tissue samples for predictor and outcome assessment are taken (and stored) at the same time point.</i> <i>• In prognostic studies, the time interval between predictor and outcome assessment can also be too short for the outcome to have been developed or expressed since the predictor assessment, or too long and may have been missed at the time of the outcome assessment, for instance in case of self-limiting diseases.</i>
<p>Applicability Assessment</p>
<p>Concern that the outcome, its definition, assessment, or timing of assessment do not match the review question or the assessor’s intended use</p> <ul style="list-style-type: none"> <i>• Applicability assessment is directed by the review question or the assessor’s intended use of the prediction model, defined under step 1.</i> <i>• Concerns on applicability in this domain addresses inconsistencies between definition, method and timing of outcome assessment given the review question or assessor’s intended use. Outcomes should be defined and assessed at the proper time intervals, using methods potentially applicable to the targeted context of the review or assessor. For example, a review may specifically aim to determine a short- or long-term prognosis, making the judgement of the proper time interval between predictor and outcome assessment relevant for applicability judgements. Or a study might have used a composite outcome that consisted of components different from the components of the outcome definition of the review question.</i> <i>• Similar to domain 1 and 2, a subtle distinction between risk of bias versus applicability assessment may arise. Consider a review of prognostic models for predicting the occurrence of major depressive</i>

disorders. The presence of depressive disorder may be assessed using the comprehensive depression section of the Composite International Diagnostic Interview (CIDI). However, studies may also have used the Patient Health Questionnaire-9 (PHQ-9), a reduced version of this PHQ (called the PHQ-2), the mental health section of the SF-36 questionnaire, or even used self-reported presence of major depression. One might then focus the review on only those studies that used the CIDI approach, defining all other studies as not applicable. Alternatively, there is no applicability concern, if the explicit aim of the review or assessor is to include all developed models for major depression, regardless of the outcome assessment method. Sometimes in model validation studies investigators intentionally apply different outcome definitions or measurement methods than in the development study. This might not be a problem if the systematic review explicitly aimed to include all validations of the model regardless of outcome definition and measurement method.

- As discussed for domains 1 and 2, in reviews that aim to meta-analyse the average performance of a specific model across multiple validation studies, heterogeneity in performance among validation studies is expected due to differences in definition and measurement of the outcome.

The Analysis domain explicitly distinguishes between three types of model evaluation:

- Apparent performance (A): Model performance estimated using exactly the same data as that used for model development.
- Internal validation or evaluation (I): Model performance estimated from the development data set but after using internal validation, e.g., resampling, techniques.
- External validation (E): Model performance estimated in external participant data that was not used for development or internal validation of the model.

DOMAIN 4: Analysis

4.1 Was model evaluation based on only apparent performance avoided?

- If exactly the same data for model development and the evaluation of model performance is used (i.e., apparent performance evaluation, see Box 1 and Step 2), the estimated model performance estimates tend to be too optimistic (i.e., model performance is estimated too high, sometimes severely).
- Generally, evaluation of model performance based only on apparent performance should be avoided
- If this item is scored 'No', then I and E are irrelevant for the upcoming items and do not need to be scored because they were not done.
- If this item is scored 'No' but the sample size for the model development is large (relative to the complexity in the model, the risk of bias in the estimated performance measures may still be judged as low.

4.2 Was there evidence that the sample size was reasonable?

- The estimated performance of a prediction model is more likely optimistic when both model development and performance assessment are based on exactly the same data (apparent performance). This optimism is larger with smaller effective sample sizes (see also domain 4 item 1 under Model Development).
- Guidance for minimal sample size for (the three different types of) model evaluation is currently available. However, similar to model development, in the absence of information about the adequacy of sample size provided by the developers, we recognize it can be hard to judge the adequacy for reviewers.
- Small sample size for evaluation means the prediction model performance will be estimated with too much imprecision, in some cases leading to results that are uninformative and interpretations that are inconclusive about the performance of the model
- We recommend reviewers to be careful with applying rules of thumb to judge adequate sample size of evaluation (e.g., some have recommended at least 100 participants with the predicted event in prediction modelling studies with a binary outcome). The minimal required sample size for model evaluation may depend on the expected performance and even the distribution of the predictions (e.g., estimated outcome probabilities)
- When model performance is evaluated in particular (e.g., marginalized) subgroups, the effective sample size may be too low to be conclusive about that subgroup.

4.3 Were participants with missing or censored data handled appropriately in the analysis?

- *Generally, selective exclusions of eligible participants in a study, dataset or analysis should be avoided. For instance, participants should generally not be excluded because of missing values on predictors or outcomes (e.g., due to censoring).*
- *In general, the higher the percentage of missing or censored data, the higher the risks of a remaining selective analysed subset. However, it should be noted that even small amounts of missing or censored data may have a profound impact on the evaluation of the prediction model performance (in particular when missing data have the form of MNAR). It is therefore not feasible to provide percentages of missing or censored data that can be considered acceptable. It may thus be hard to judge the validity of prediction models evaluated on data where a small number of participants are left out due to missing values.*
- *For appropriate handling of missing data in prediction model evaluation we refer to the literature.*
- *Multiple imputation approaches are often considered preferred approaches to deal with missing predictor data. However, in some cases an alternative approach to handling missing data may be preferred when evaluating the performance of a prediction model (9). For instance, if the intended use for the prediction model is in a setting where predictor data may be missing (e.g., when using data from routine healthcare records), an evaluation of the performance of that model while using the method of handling missing data as intended in practice, may provide better reflection of its actual performance.*
- *When the performance of a particular model is evaluated in external data, sometimes a predictor included in the model is systematically missing (i.e., has not been recorded in the dataset at hand). Investigators sometimes then simply omit the predictor from the model (i.e., set the coefficient to zero). This is usually a bad idea: predictor-outcome associations of the remaining predictors would generally change if the model would have been fitted without that omitted predictor.*
- *Approaches to adequately deal with censored data in the evaluation of prediction models have been described in the literature.*
- *Especially prediction models with outcomes other than all-cause mortality, and with long or varying times of follow-up and subsequent large amounts of (informative) censoring, may require the use of competing risk approaches.*
- *It is sometimes not clear from reports whether study participants with any missing or censored data were excluded from analyses (i.e., complete-case or available-case analysis). Although one may score this item then formally with 'No Information', one may still judge this aspect as 'low quality' since most data analysis techniques automatically exclude records with any missing value on any of the predictors or outcomes.*

This item also addresses fairness when handling of missing or censored data differs across individuals of marginalised subgroups defined, e.g., by different ages, sexes/genders, races/ethnicities, geographic locations or health conditions/comorbidities.

4.4 If methods to address class imbalance were used, was the evaluation done in a dataset without imbalance correction?

- *Approaches to correct for imbalance in the outcome status (i.e., an event fraction away from 0.5) are increasingly common in prediction research, particularly in AI-based classification. These class imbalance corrections, such as undersampling of the majority class (i.e., random selection from the group with the highest prevalence), oversampling of the minority class (i.e., duplicating from the group with the lowest prevalence), should not be applied when evaluating the performance of a model.*
- *If imbalance corrections were applied in the development of the model, evaluation of the model should be done in data where this imbalance correction is avoided.*

4.5 If data splitting was done to create training and test datasets, was there evidence that data leakage was avoided?

- *Data leakage typically refers to the situation where data that are used for the evaluation of the prediction model overlaps with the training data such that it typically will lead to an overestimation of the performance of the prediction model.*
- *Given that data leakage can lead to performance overestimation this can be overcome when the appropriate internal or external validation techniques are used.*

<ul style="list-style-type: none"> • <i>For instance, data leakage can occur if the prediction model's parameters are again tuned on the data used to evaluate the model (rather than just using the tuning parameters based on the development data).</i> • <i>This item is not applicable to apparent performance and external validation.</i>
<p>4.6 If resampling methods were used to evaluate model performance, were all model development steps replicated in the resampling process?</p> <ul style="list-style-type: none"> • <i>A developed model may be too fitted or adapted to the dataset at hand, yielding optimistic model performance measures. This risk is higher when the development data set is relatively small in view of, e.g., the complexities in the data, the number of parameters (e.g., predictor-outcome associations) that need to be estimated, and the number of model selection steps being applied.</i> • <i>Investigators should preferably apply some form of internal validation, such as resampling based on bootstrapping or cross validation, to quantify the possible extent of model performance optimism. Subsequently, they should adjust or correct the estimated model performance measures (e.g., c-index) for this optimism. Any resampling procedure for the evaluation of predictive should include all model development steps including imputation, variable selection, hyperparameter tuning and model building. If not, the extent of optimism in the model performance measures tends to be underestimated and model performance overestimated.</i> • <i>This item is not applicable to apparent performance and external validation.</i>
<p>4.7 Was the predictive performance of the model evaluated appropriately, e.g., calibration, discrimination, and net benefit?</p> <ul style="list-style-type: none"> • <i>PROBAST+AI addresses models providing predictions on individual level (i.e., estimating outcome probabilities given a specific combination of predictor values). Accordingly, a model performance evaluation includes assessing the agreement of observed and estimated outcome values over the entire range of predicted outcome values (calibration plots), assessing if the model predictions differ by observed outcome values (discrimination indices such as c-index), and ideally some measure of correct decision making based on the model (such as the net benefit measure).</i> • <i>Calibration (ideally a calibration plot and not only a statistical test on goodness-of-fit) and discrimination are important model performance parameters. The omission of any information on each of these two groups of measures would signal this item as 'No' or 'Probably No'.</i> • <i>Just providing a model's apparent (instead of internally or externally evaluated) calibration is not very informative, as a model is typically well calibrated on exactly the same data set as from which it was developed (or miscalibrated if developed under strict regulation or when using strong shrinkage).</i> • <i>Sometimes model probability thresholds are introduced after which classification measures such as sensitivity or specificity are estimated, typically for diagnostic models. Use of probability thresholds may lead to important loss of information, as the entire range of predicted probabilities is ignored. Also, threshold choices are often data-driven (rather than prespecified), to maximize the classification measures, resulting in higher risks of bias.</i> • <i>This item also addresses fairness by judging whether equity of model performance (e.g., subgroup calibrations) was assessed across marginalised subgroups defined, e.g., by different ages, sexes/genders, races/ethnicities, geographic locations or health conditions/comorbidities.</i>

References

1. Riley RD, Snell KI, Ensor J, Burke DL, Harrell FE, Jr., Moons KG, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat Med*. 2019;38(7):1276-96.
2. Riley RD, Snell KIE, Ensor J, Burke DL, Harrell FE, Jr., Moons KGM, et al. Minimum sample size for developing a multivariable prediction model: Part I - Continuous outcomes. *Stat Med*. 2019;38(7):1262-75.
3. Riley RD, Ensor J, Snell KIE, Harrell FE, Jr., Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. *Bmj*. 2020;368:m441.
4. Šinkovec H, Heinze G, Blagus R, Geroldinger A. To tune or not to tune, a case study of ridge logistic regression in small or sparse datasets. *BMC Medical Research Methodology*. 2021;21(1):199.
5. Riley RD, Snell KIE, Martin GP, Whittle R, Archer L, Sperrin M, et al. Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small. *J Clin Epidemiol*. 2021;132:88-96.
6. Van Calster B, van Smeden M, De Cock B, Steyerberg EW. Regression shrinkage methods for clinical prediction models do not guarantee improved performance: Simulation study. *Statistical Methods in Medical Research*. 2020;29(11):3166-78.
7. Collins GS, Ogundimu EO, Cook JA, Manach YL, Altman DG. Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. *Stat Med*. 2016;35(23):4124-35.
8. Sauerbrei W, Perperoglou A, Schmid M, Abrahamowicz M, Becher H, Binder H, et al. State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues. *Diagnostic and Prognostic Research*. 2020;4(1):3.
9. Sisk R, Sperrin M, Peek N, van Smeden M, Martin GP. Imputation and missing indicators for handling missing data in the development and deployment of clinical prediction models: A simulation study. *Statistical Methods in Medical Research*. 2023;32(8):1461-77.