Bias and Reporting Quality of Clinical Prognostic Models for Idiopathic Pulmonary Fibrosis: A Cross-Sectional Study

Jiaqi Di1, Xuanlin Li1, Jingjing Yang1, Luguang Li2, Xueqing Yu2

1Co-Construction Collaborative Innovation Center for Chinese Medicine and Respiratory Diseases by Henan & Education Ministry of P.R, Henan University of Chinese Medicine, Zhengzhou, 450046, People’s Republic of China; 2Department of Respiratory Diseases, The First Affiliated Hospital of Henan University of Chinese Medicine, Zhengzhou, Henan, 450000, People’s Republic of China

Correspondence: Xueqing Yu, Email yxqshi@163.com

Objective: This study aims to evaluate the risk of bias (ROB) and reporting quality of idiopathic pulmonary fibrosis (IPF) prediction models by assessing characteristics of these models.

Methods: The development and/or validation of IPF prognostic models were identified via an electronic search of PubMed, Embase, and Web of Science (from inception to 12 August, 2021). Two researchers independently assessed the risk of bias (ROB) and reporting quality of IPF prediction models based on the Prediction model Risk Of Bias Assessment Tool (PROBAST) and Transparent Reporting of a multivariable prognostic model for Individual Prognosis or Diagnosis (TRIPOD) checklist.

Results: Twenty prognostic model studies for IPF were included, including 7 (35%) model development and external validation studies, 8 (40%) development studies, and 5 (25%) external validation studies. According to PROBAST, all studies were appraised with high ROB, because of deficient reporting in the domains of participants (45.0%) and analysis (67.3%), and at least 55% studies were susceptible to 4 of 20 sources of bias. For the reporting quality, none of them completely adhered to the TRIPOD checklist, with the lowest mean reporting score for the methods and results domains (46.6% and 44.7%). For specific items, eight sub-items had a reporting rate ≥80% and adhered to the TRIPOD checklist, and nine sub-items had a very poor reporting rate, less than 30%.

Conclusion: Studies adhering to PROBAST and TRIPOD checklists are recommended in the future. The reproducibility and transparency can be improved when studies completely adhere to PROBAST and TRIPOD checklists.

Keywords: idiopathic pulmonary fibrosis, PROBAST, reporting quality, risk of bias, TRIPOD

Introduction

Idiopathic pulmonary fibrosis (IPF) is the most common type of interstitial lung disease and is characterized by dyspnea and progressive deterioration of lung function,1,2 with median survival of 2 to 3 years from time of diagnosis.3,4 The acute exacerbations and the associated complications often lead to hospitalization and death,5–8 resulting in significant economic and health-care burdens.5,9 At present, there are some challenges in the diagnosis and treatment of IPF, for instance, the complicated diagnosis process, limited and expensive interventions, and corresponding side effects.1,10 Therefore, to mitigate the risks and burdens of IPF and to improve the perceptions of best practices of care, more efficient prognosis predictions are needed.11

Multivariable prediction models in which multiple characteristics or pieces of information are applied can estimate an individual’s risk of a current condition in future.12 When a patient’s score passes a certain threshold, an alarm may be sent to the appropriate clinicians for further evaluation and intervention. In recent years, interest in using prediction models has increased and the models are more often recommended in clinical practice guidelines13,14 for individual prognosis and diagnosis. At present, there are more prognostic models for IPF.15,16 However, because there is a surplus of IPF prognostic models with widely variable quality, it is important to identify IPF prognostic models of high quality.
Risk of bias (ROB) is usually defined as the presence of a systematic error that may affect the study’s validity. The Prediction model Risk of Bias Assessment Tool (PROBAST)\textsuperscript{17} guideline has been developed to assess the ROB of model development and model validation, including updating of a prediction model.\textsuperscript{18} A full reporting is essential to evaluate the validation and applicability of a multivariable prediction model;\textsuperscript{19,20} therefore, a protocol was developed for a guided, Transparent reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) statement.\textsuperscript{21} Previous systematic reviews from other research teams showed that there are high ROB and suboptimal reporting quality in prediction models in oral health and preterm birth.\textsuperscript{22,23} To date, there has been limited data for evaluating the quality of existing IPF prognostic models; therefore, in this cross-sectional study, we assessed IPF prognostic models with PROBAST and TPRIPOD checklists, aiming to identify the ROB and reporting quality of these studies and highlight the strengths and limitations of the methodologies of IPF prognostic models.

**Materials and Methods**

**Study Design**

We conducted a critical cross-sectional appraisal on ROB and reporting quality of IPF prediction models using the PROBAST\textsuperscript{24} and TRIPOD checklists,\textsuperscript{21} respectively. Although this review was not a typical system review (SR), we did strictly adhere to the guidelines of conducting and reporting a SR.\textsuperscript{25} The details of reported PRISMA-Checklists are listed in Supplementary Table 1. The review protocol was registered on INPLASY.\textsuperscript{26}

**Inclusion and Exclusion Criteria**

Studies were included if they met the following criteria: 1) A prognostic multivariable prediction model of IPF describing development, validation, or both, and 2) They should predict events at the probability of future outcomes (prognosis) related to IPF.

Studies were excluded if they 1) examined independent prognostic factors and did not aim to develop a model; 2) combined IPF with other diseases; 3) were not original research (such as review, methodological articles, conference abstracts, protocols); and 4) performed comparison models.

**Literature Search**

A comprehensive literature search was conducted in PubMed, Embase, and Web of Science between inception to 12 August, 2021. The details of the search strategy can be found in Supplementary Table 2. All relevant articles in the reference list of all included articles were also retrieved.

**Study Selection**

We created a database in EndNote X9 software. After eliminating duplicates, we read the titles and abstracts for a preliminary screening. Then, we downloaded the full text and filtered it again until all relevant prediction models of IPF were confirmed. Two researchers selected the literature, and if there were discrepancies between them, it was addressed by the discussion with the third researcher.

**Data Extraction**

We focused our research on study design, outcome measurement, modeling methodology, and validation strategy. Therefore, one researcher extracted the above key information, including author, publication year, population characteristics, follow-up time, etc. Another researcher checked the extracted data, and if there existed different opinions between two researchers, they would refer to the original text and revise it.

**Application of Evaluation Tools**

We classified each study into model development with or without external validation in the same publication and external validation study of a previously developed model only. Two of us used the PROBAST tool to assess the ROB for each included study. Following four PROBAST domains (Participants, Predictors, Outcome, Analysis), we assessed 20
signaling questions for development models and 17 signaling questions for validation models within each domain with yes/probably yes, no/probably no, or no information. We rated domain-level and applicability assessments using “low risk of bias,” “high risk of bias,” and “unclear risk of bias” according to PROBAST suggestion.  

Two researchers assessed the reporting quality of the included models according to the TRIPOD statement. The checklist covered 37 sub-items in 6 domains, including title and abstract (items 1 and 2), introduction (item 3), methods (items 4 to 12), results (items 13 to 17), discussion (items 18 and 19), and other information (items 21 and 22). We rated items as “reported” if the relevant information was fully presented, “unreported” if all relevant information were lacking, and “not applicable” for inappropriate data, as reported by previous studies.

**Statistical Methods**

The reporting rates of PROBAST and TRIPOD were calculated in a descriptive manner using proportions (%). The PROBAST items for development studies only (items 4.5, 4.8, 4.9) and the TRIPOD “if done” item (item 5c), the validation items (items 10c, 10e, 13c, 17, 19a) were used as both the numerator and denominator when the overall adherence rate was calculated. In each evaluation, we completed two rounds of pilot evaluation, SPSS 25.0 software was used to calculate the Kappa value of the internal consistency coefficient, and the formal assessment was made when Kappa ≥0.8. Our pre-evaluation results (Kappa = 0.818, P ≤ 0.001) demonstrated a good agreement between our reviewers.

**Results**

**Literature Selection**

A total of 1670 records were collected. Of those, 46 eligible full-text articles were reviewed after removal of duplicates and irrelevant articles. We excluded 26 studies for reasons shown in Figure 1. Finally, 20 IPF clinical prognostic model studies met the inclusion criteria in this process.

**Overview of Characteristics for IPF Prognostic Model Studies**

Of the 20 prognostic model studies, 7 (35%) publications reported model development and external validation, 8 (40%) publications reported model development only, and 5 (25%) publications reported external validation with one updated. According to these studies, 20 prognostic models were identified but more than 50% models did not undergo any external validation. A summary of characteristics for the included studies is shown in Table 1.

**ROB of IPF Prediction Model Studies**

All models were at high ROB in development or validation. Eight models are rated as high ROB in applicability. Figures 2 and 3 show the proportion assessment for each PROBAST item and proportion of studies with potential bias using PROBAST, respectively. Overall, 55% of the studies were identified with at least four sources of bias (sub-items 1.1, 2.2, 4.2, and 4.8).

As for participants, predictors, outcomes, and analysis domains, there were 12, 12, 6, and 18 studies that had a high ROB, respectively (The “biased” domain, applicability identified in each study is provided in Supplementary Figure 1). Of the included studies, 55.0% resulted in a high risk of bias because of the inclusion of retrospective studies (sub-item 1.1). For predictors of definition (item 2.1) and assessment (item 2.2), one study proposed to exclude the predictor of diffusing capacity of lung for carbon monoxide (DLCO), and one study included a question as a predictor. In addition, 45.0% studies may not report any actions to blind assessment of the predictors. However, unreasonable predictor selection methods may limit the use of the model. Most sub-items in outcomes domains have low ROB, especially inconsistency in defining outcome (sub-item 3.4) and time interval between predictor and outcome assessment (sub-item 3.6). This is probably because most studies used outcomes that are easy to assess (eg, death, survival time). In addition, in the analysis domain, continuous variables were converted into dichotomous variables (sub-item 4.2) or there was lack of overfitting consideration (item 4.8) in most studies, which are the main reasons leading to the high ROB in analysis domain.
Bias Related to Applicability
There was high ROB about applicability, indicating that these studies were not well aligned with the search question. This is most commonly due to concerns about applicability in domain 1 (population) for included randomized clinical trials or specific groups. In addition, two out of 20 studies’ outcomes included exacerbation, but the definition of acute exacerbation may have varied between studies and may have been expanded to include events for which a trigger can be identified, such as infection.

Reporting Quality of IPF Prediction Model Studies
Of the six domains, the introduction part of TRIPOD with scores of 90.0% is the domain where IPF prediction model studies had the highest mean score reporting rate. The prediction model of IPF had the lowest mean score for the domains of methods and results (46.6% and 44.7%). The mean reporting rate of each domain according to the TRIPOD checklist is less than 65%. The details of reporting rates of each of the TRIPOD domains are shown in Figure 4.

For specific items, eight sub-items had a reporting rate ≥80% adherence to the TRIPOD checklist, and nine sub-items showed a low reporting rate with less than 30%. None of the included IPF prediction model studies reported any actions to blind assessment of the outcome/predictors (items 6b, 7b). In addition, with respect to validation models, few of them
<table>
<thead>
<tr>
<th>Author (Years)</th>
<th>Data Sources</th>
<th>Participant Sample Size (Events)</th>
<th>Time of Follow-Up</th>
<th>Predictors in Final Model</th>
<th>Modeling Method</th>
<th>Internal Validation Method</th>
<th>Outcomes</th>
<th>Name of the Model</th>
<th>Model Performance Measure</th>
<th>Overall Risk of Bias Using PROBAST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ley 2012&lt;sup&gt;15&lt;/sup&gt;</td>
<td>The patients enrolled in three hospitals from two countries (US and Italy)</td>
<td>D: 228(89); V: 330(186)</td>
<td>3 years</td>
<td>Gender, age, FVC%, DLCO%</td>
<td>Based on Fine–Gray models for survival</td>
<td>10-fold cross-validation</td>
<td>Survival time</td>
<td>GAP models (Calculator and Index)</td>
<td>C-index: 0.708 (95% CI [0.637–0.75]) (calculator); 0.693 95% CI [0.622–0.731] (index)</td>
<td>NP</td>
</tr>
<tr>
<td>Huang 2015&lt;sup&gt;16&lt;/sup&gt;</td>
<td>Patients from UCMC and UPMC</td>
<td>D: 45(NP); V1: 21(NP); V2: 75(NP)</td>
<td>D: 18.8 months; V1: 43.8 months; V2: 23.3 months</td>
<td>118 prognostic predictor genes within IL1R2, ERAF, CEACAM8, ARG1, FOXO3</td>
<td>Cox proportional hazard regression</td>
<td>10-fold cross validation</td>
<td>The progression of IPF (survival)</td>
<td>Prognostic index (PI) score</td>
<td>AUC: 0.96</td>
<td>Kaplan-Meier plot</td>
</tr>
<tr>
<td>Torrisi 2019&lt;sup&gt;17&lt;/sup&gt;</td>
<td>Patients from four international academic ILD centres (Italy, Germany, Netherland, US)</td>
<td>D: 476(NP); V: 461(NP)</td>
<td>28 months</td>
<td>Age, p%FVC, p%DLCO, diabetes mellitus, systemic hypertension, major depressive disorder, valvular heart disease, atrial arrhythmias</td>
<td>Cox proportional hazard regression</td>
<td>NP</td>
<td>Survival time</td>
<td>TORVAN models (full and sparse)</td>
<td>C-index: 0.71 (95% CI [67.8–74.2]) (full model); 0.725 95% CI [69.5–75.6] (sparse model)</td>
<td>Calibration plot</td>
</tr>
<tr>
<td>Nishikiori 2020&lt;sup&gt;18&lt;/sup&gt;</td>
<td>Patients from 3 regions of Japan and Korea</td>
<td>D: 326(NP); V1: 117(48); V2: 1262(415)</td>
<td>29 months</td>
<td>Gender, age, VC%, DLCO%</td>
<td>Cox proportional hazard regression</td>
<td>NP</td>
<td>Survival time</td>
<td>Modified GAP model (Index)</td>
<td>Harrell’s C-index: 0.67 (95% CI 0.601–0.739) (1 year); 0.698 (0.638–0.758) (2 year); 0.738 (0.782–0.795) (3 year)</td>
<td>NP</td>
</tr>
<tr>
<td>Li 2021&lt;sup&gt;19&lt;/sup&gt;</td>
<td>IPF patients from the Gene Expression Omnibus database</td>
<td>D: 176(121); V1: 57(NP); V2: 120(NP); V3: 119(NP)</td>
<td>5 years</td>
<td>Hypoxia-Immune-Related genes&lt;sup&gt;20&lt;/sup&gt;</td>
<td>Cox proportional hazard regression</td>
<td>Three-fold cross validation</td>
<td>The progression of IPF (survival)</td>
<td>Hypoxia-immune-related prediction model</td>
<td>AUC: 0.789 (1 year); 0.768 (2 year); 0.754 (3 year); 0.798 (4 year); 0.913 (5 year)</td>
<td>Kaplan–Meier plot</td>
</tr>
<tr>
<td>Lu 2021&lt;sup&gt;21&lt;/sup&gt;</td>
<td>The microarray expression matrix dataset of 75 IPF from GSE28042</td>
<td>D: 75(NP); V: 45(NP)</td>
<td>3 years</td>
<td>Inflammation-Related gene: S100A12, CCR7, TNFSF4</td>
<td>Cox proportional hazard regression</td>
<td>NP</td>
<td>The progression of IPF (survival)</td>
<td>Inflammation-Related Prognostic Model</td>
<td>AUC: 0.611 (1 year); 0.695 (2 year); 0.681 (3 year)</td>
<td>NP</td>
</tr>
</tbody>
</table>

(Continued)
<table>
<thead>
<tr>
<th>Author</th>
<th>Data Sources</th>
<th>Participant Sample Size (Events)</th>
<th>Time of Follow-Up</th>
<th>Predictors in Final Model</th>
<th>Modeling Method</th>
<th>Internal Validation Method</th>
<th>Outcomes</th>
<th>Name of the Model</th>
<th>Model Performance Measure</th>
<th>Overall Risk of Bias Using PROBAST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xia 2021</td>
<td>The gene expression data of BAL cells and clinical information for IPF patients came from the GEO database (Freiburg, Siena, and Leuven)</td>
<td>D: 176(NP); V: 64(NP)</td>
<td>3 years</td>
<td>Bronchoalveolar lavage cell-associated gene: TLR2, CCL2, HTRA1, SFN</td>
<td>Cox proportional hazard regression</td>
<td>NP</td>
<td>Survival time</td>
<td>NP</td>
<td>AUC: 0.773 (1 year); 0.772 (2 year); 0.752 (3 year)</td>
<td>High</td>
</tr>
<tr>
<td>King 2001</td>
<td>Patients enrolled into a Specialized Center of Research Study at the National Jewish Medical and Research Center</td>
<td>238 (155)</td>
<td>110 months</td>
<td>Age, smoking status, clubbing, profusion (radiographic abnormality), pulmonary hypertension, TLC, PaO2 at maximal exercise</td>
<td>Cox proportional hazard regression</td>
<td>NP</td>
<td>One-year mortality</td>
<td>CRP score (Complete and Abbreviated)</td>
<td>NP</td>
<td>NP</td>
</tr>
<tr>
<td>du Bois, 2011</td>
<td>Patients from two clinical trials of IFN-g1b and GIPF 007</td>
<td>1099 (152)</td>
<td>1 year</td>
<td>Age, history of respiratory hospitalization, %pFVC, 24-week in % pFVC</td>
<td>Methodology set forth by Wilson and coworkers</td>
<td>NP</td>
<td>One-year mortality</td>
<td>Mortality risk scoring system (Comprehensive model and Clinical model)</td>
<td>C-statistic: 0.77 (95% CI [0.72–0.81]) (Comprehensive model); 0.75 (95% CI [0.71–0.79]) (Clinical model)</td>
<td>High</td>
</tr>
<tr>
<td>Soares 2015</td>
<td>Patients from three reference centers for interstitial lung diseases (ILD) in São Paulo</td>
<td>120 (80)</td>
<td>37.5 months</td>
<td>Dyspnea, %DLCO, %FVC, FEV1/FVC</td>
<td>Cox proportional hazard regression</td>
<td>NP</td>
<td>Survival time</td>
<td>DDS index</td>
<td>C-statistic: 0.78 (95% CI [0.705–0.865])</td>
<td>NP</td>
</tr>
<tr>
<td>Ashley, 2016</td>
<td>Patients come from an observational cohort study of a multi-center</td>
<td>60 (35)</td>
<td>80 weeks</td>
<td>Biomarkers: ICOS, LGMN, FCN2, TRF3, VEGF sR2, Cathepsin S</td>
<td>Multivariable logistic and Cox regression models</td>
<td>Boot-strap</td>
<td>Progression-free survival†</td>
<td>Six-SOMAmer Index</td>
<td>AUC: 0.91</td>
<td>NP</td>
</tr>
<tr>
<td>Lee, 2018</td>
<td>Retrospective cohort study in Asan Medical Center</td>
<td>144 (106)</td>
<td>57.9 months (range, 13–131 months)</td>
<td>Age, desaturation, fibrosis score†, interval changes in fibrosis score</td>
<td>Cox proportional hazard regression</td>
<td>NP</td>
<td>Survival time</td>
<td>NP</td>
<td>C-index: 0.768 (95% CI [0.707–0.829])</td>
<td>NP</td>
</tr>
<tr>
<td>Study</td>
<td>Patients</td>
<td>Follow-up</td>
<td>Dyspnea, FVC%, ExSpO2, Oxygen use</td>
<td>Cox proportional hazard regression</td>
<td>Bootstrap</td>
<td>Survival time</td>
<td>DOS score</td>
<td>C-statistic</td>
<td>NP</td>
<td>High</td>
</tr>
<tr>
<td>-------</td>
<td>----------</td>
<td>-----------</td>
<td>----------------------------------</td>
<td>----------------------------------</td>
<td>-----------</td>
<td>---------------</td>
<td>-----------</td>
<td>-------------</td>
<td>-----</td>
<td>------</td>
</tr>
<tr>
<td>Fukuda 2020</td>
<td>Patients from 3 reference centers for ILD in São Paulo</td>
<td>43 months</td>
<td>Dyspnea, FVC%, ExSpO2 (oxygen desaturation during exercise)</td>
<td>Bootstrap</td>
<td>Survival time</td>
<td>DOS score</td>
<td>C-statistic: 0.7</td>
<td>NP</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>Tang 2020</td>
<td>Patients participated in INPULSIS-1 and 2</td>
<td>372 days</td>
<td>Age, decline in FVC to week 52, baseline % pFVC, supplement oxygen use</td>
<td>Laplace estimation method</td>
<td>Bootstrap</td>
<td>Exacerbation risk</td>
<td>Time-to Event (TTE) model</td>
<td>NP</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>Moor 2021</td>
<td>Patients come from a prospective cohort study in Netherlands</td>
<td>1 year</td>
<td>Surprise question, MRC score, %pDLCO</td>
<td>Multivariable logistic regression model</td>
<td>NP</td>
<td>One-year mortality</td>
<td>NP</td>
<td>C-statistic: 0.82 95% [0.73–0.91]</td>
<td>NP</td>
<td>High</td>
</tr>
</tbody>
</table>

**Model validation studies with or without updating**

<table>
<thead>
<tr>
<th>Study</th>
<th>Patients</th>
<th>Follow-up</th>
<th>RH, UCSD SOBQ, 6 MWD, FVC 24-week change</th>
<th>Cox proportional hazard regression</th>
<th>Bootstrap</th>
<th>Survival time</th>
<th>Longitudinal GAP model</th>
<th>C-statistic: 0.785 95% CI [0.78–0.79]</th>
<th>H-L test</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ley 2015</td>
<td>One study of interferon γ1b and two studies of pirfenidone in IPF</td>
<td>1.1 years (0.01–2.36 years)</td>
<td>RH, UCSD SOBQ, 6 MWD, FVC 24-week change</td>
<td>Cox proportional hazard regression</td>
<td>Bootstrap</td>
<td>Survival time</td>
<td>Longitudinal GAP model</td>
<td>C-statistic: 0.785 95% CI [0.78–0.79]</td>
<td>H-L test</td>
<td>High</td>
</tr>
<tr>
<td>Kim, 2015</td>
<td>Patients come from Seoul National University Hospital in Korean</td>
<td>4.64 years</td>
<td>Gender, age, FVC%, DLCO%</td>
<td>NA</td>
<td>NA</td>
<td>Survival time</td>
<td>GAP model</td>
<td>C statistic: 0.74 95% CI [0.35–1] (1 year), 0.71 95% [0.44–0.92] (2 year), 0.68 95% [0.46–0.87] (3 year); GAP index 0.72 95% CI [0.34–1] (1 year), 0.69 95% [0.42–0.91] (2 year), 0.66 95% CI [0.44–0.85] (3 year)</td>
<td>H-L test</td>
<td>High</td>
</tr>
</tbody>
</table>
Table 1 (Continued).

<table>
<thead>
<tr>
<th>Author</th>
<th>Data Sources</th>
<th>Participant Sample Size (Events)</th>
<th>Time of Follow-Up</th>
<th>Predictors in Final Model</th>
<th>Modeling Method</th>
<th>Internal Validation Method</th>
<th>Outcomes</th>
<th>Name of the Model</th>
<th>Model Performance Measure</th>
<th>Overall Risk of Bias Using PROBAST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lee 2016 †</td>
<td>Patients come from 54 university and teaching hospitals in Korean</td>
<td>1228 (NP)</td>
<td>19 ± 16 months</td>
<td>Gender, age, FVC%, DLCO%</td>
<td>NA</td>
<td>NA</td>
<td>Survival time</td>
<td>GAP model</td>
<td>C-statistic: GAP calculator 0.61 95% CI [0.559–0.653] (1 year), 0.61 95% [0.566–0.649] (2 year), 0.59 95% [0.549–0.627] (3 year); GAP index 0.59 95% CI [0.537–0.638] (1 year), 0.59 95% CI [0.544–0.631] (2 year), 0.57 95% [0.53–0.611] (3 year)</td>
<td>NP</td>
</tr>
<tr>
<td>Harari 2019 †</td>
<td>Patients treated with pirfenidone in 12 interstitial lung disease centers across Italy</td>
<td>68 (22)</td>
<td>2.4 years</td>
<td>Gender, age, FVC%, DLCO%</td>
<td>NA</td>
<td>NA</td>
<td>Survival time</td>
<td>GAP model</td>
<td>C-index: 0.74 95% CI [0.57–0.93] (GAP index); 0.77 95% CI [0.59–0.93] (GAP calculator)</td>
<td>H-L test</td>
</tr>
<tr>
<td>Abe 2020 †</td>
<td>Patients treated with nintedanib in the Chiba University Hospital</td>
<td>89 (18)</td>
<td>16.4 months</td>
<td>Gender, age, FVC%, DLCO%</td>
<td>NA</td>
<td>NA</td>
<td>Survival time</td>
<td>GAP model</td>
<td>NP</td>
<td>NP</td>
</tr>
</tbody>
</table>

Notes: †Progression-free survival as determined by the time until any of the following: death, acute exacerbation of IPF, lung transplant, or relative decrease in forced vital capacity (FVC, liters) of 10% or DLCO (mL/min/mmHg) of 15%; ††Fibrosis score: The fibrosis score was defined as the sum of the extent of honeycombing and reticular opacity; †††NALCN, IL1R2, S100A12, PROK2, CCL8, RAB15, MARCKSL1, TPCN1, HS3ST.  
Abbreviations: NA, not applicable; UCMC, University of Chicago Medical Center; UPMC, University of Pittsburgh Medical Center; NP, not provided; D, derivation; V, validation; BAL, bronchoalveolar lavage; RH, respiratory hospitalization in the preceding 24 weeks; UCSD SOBQ, University of California San Diego Shortness of Breath Questionnaire; ILD, interstitial lung disease; FVC, forced vital capacity; DLCO, diffusing capacity of carbon monoxide; VC, vital capacity; FEV1/FVC, forced expiratory volume in one second/forced vital capacity; AUC, area under curve; GAP, gender-age-physiology; H-L, Hosmer Lemeshow.
identified any difference or showed the comparison from the data between development and validation (items 12, 13c). The details of reporting for each item adherence to the TRIPOD checklist are shown in Figure 5.

**Discussion**

**Main Findings**

In this cross-sectional study of prognostic risk models related to the IPF, we identified and critically appraised 20 studies that were described 20 models. All models reported good to excellent predictive performance, but all of them had high ROB according to PROBAST, and none of them completely adhered to the TRIPOD checklist for reporting, which demonstrates deficiencies in applicability and reporting transparency. First, as for ROB, we identified the mean reporting rate of each domain, which was less than 70%. Similarly, using the TRIPOD standards, we found that the mean reporting
rate of each domain was less than 65%, especially the methods and results domains. The two tools suggest that there was a general lack of transparent reporting and identification of bias across the studies of IPF progression models.

The ROB of Included Studies According PROBAST
The main aim of prediction models is to support medical decision-making, and a high ROB implies that these models will probably perform worse in practice than in the studies reported by researchers.\textsuperscript{11} We identified the reporting rate in participants domain was 45.0% and 67.3% in analysis, which means researchers should pay attention to the ROB of IPF prediction models.

In the participants domain, most studies did not select the appropriate data source, and there were a large number of retrospective studies in our study. The Gender-Age-Physiology (GAP) system, which might help inform decision-making, has been externally validated five times with good discrimination, but the ROB for the external validation studies of the retrospective research data was high. While prospective cohort studies are recommended for model development, there are practical issues for using prospective cohort studies given that IPF is relatively rare. In addition, there are many well-designed prospective national IPF registries that could serve this purpose.

In the analysis domain, continuous predictors should not be dichotomized or categorized, as this will result in loss of information, which in turn may lead to imprecise risk estimates.\textsuperscript{47} In addition, we suggest that continuous predictors should appear as original data in future research on models. Calibration, which can assess the fit of models, is one of the essential features in the assessment of the usability of a predictive scale.\textsuperscript{48} However, few studies considered calibration in
our review. Although some studies evaluated this indicator, most of them used the statistical test of Hosmer-Lemeshow (H-L). However, the H-L test cannot retain the most information on possible miscalibration, so a calibration plot or other evaluation method is recommended. Lack of internal validation may lead to overfitting because quantifying the predictive performance of a model on the same data from which the model was developed tends to give optimistic estimates of performance, so that internal validation is recommended when there is no external validation.

The Reporting Quality of Included Studies According to TRIPOD

Full reporting of studies facilitates reproducibility of models, appraisal of model validity, and judgment of model generalizability to other clinical settings. We found there is room for improvement in the reporting quality of IPF prediction models.

In the methods domain, a large number of these prediction models skipped estimating sample size, mainly because of a lack of consensus in estimating sample size requirements for derivation and validation. However, a reasonable sample size is necessary, so until a canonical method of calculating sample size is available, we can use 10 or 15 times the flat number of events to initially calculate the sample size. It was brought to our attention that the GAP system has been externally validated several times but has rarely been updated. Models will be more generalizable when the case mix of the new population is within the case mix range of the development population. Therefore, in future studies, it is recommended to continuously update the GAP models according to the characteristics of the study population.

In the results domain, the reporting rates of participants’ engagement process and method of using the prediction model are lower than 30%. To make the results of prediction model research more complete and more transparent, we should consider not only the performance of the prediction model but also how to use the developed model for individuals. Therefore, we recommend that items such as how to use the model be reported clearly and directly.

Strengths and Limitations

To our knowledge, this is the first review to assess the ROB and reporting quality of multivariable prognostic models for IPF based on the PROBAST and TRIPOD checklists. We systematically searched the IPF prognostic models to assess the ROB and explore report quality of these studies and more importantly, to summarize the characteristics of the existing IPF prediction model studies to provide necessary references for future research. Before the formal assessment, we conducted a pilot experiment, requiring that the evaluator’s consistency coefficient was higher than 0.8 before the evaluation was carried out to further reduce the subjectivity of the researchers and improve the credibility of the results of our review.

Our study also has several limitations. First, we did not assess the relationship among ROB, reporting rates, and forecast accuracy. In addition, our study only considered the important factor of insufficient reporting, but other factors need to be further explored to promote the scientific implementation of the IPF prediction model in clinical practice.

Conclusion

According to the study, the rates of reporting adhering to PROBAST and TRIPOD checklists are low, especially in the domains of participants and analysis for PROBAST and the domains of methods and results for TRIPOD. Moreover, there is a general lack of identification of bias and transparent reporting, which can decrease the reproducibility rate of IPF models. However, the low rate of reporting, according to the two evaluation tools, does not mean low prediction accuracy, but the reproducibility and transparency can be improved when studies completely adhere to PROBAST and TRIPOD checklists. Therefore, studies adhering to PROBAST and TRIPOD checklists are recommended in the future.

Funding

This study was supported by National Natural Science Foundation of China (grant number 81973779, 82174307); Special Scientific Research Project of National Clinical Research Base of Traditional Chinese Medicine of China (grant number 2018JDZX006); Central Plains Thousand People Program (grant number 194200510018).
Disclosure
Jiaqi Di and Xuanlin Li are co-first authors. The authors declare that there are no conflicts of interest.

References