Precautions to Consider in the Analysis of Prognostic and Predictive Indices

Loïc Chartier[1], Aurélien Belot[1], Isabelle Chaillol[1], Mad-Hélénie Elsensohn[1], Cédric Portugues[1], Marguerite Fournier[1], Clémentine Joubert[1], Elodie Gat[1], Cécile Pizot[1], Patrick Fogarty[1], Tesla Muraïri[1], Romain Ould Ammar[1], Jérôme Paget[1], Fanny Cherblanc[2], Romain Ricci[3], Laetitia Vercellino[4], Salim Kanoun[5], Anne-Ségolène Cottereau[6], Catherine Thieblemont[7], and Olivier Casasnovas[8]

[1]Biostatistics Department, LYSARC, Hôpital Lyon Sud, Pierre-Bénite, France; [2]Medical Department, LYSARC, Hôpital Lyon Sud, Pierre-Bénite, France; [3]Imaging Department, LYSARC, Hôpital Henri-Mondor, Créteil, France; [4]Department of Nuclear Medicine, Hôpital Saint-Louis, AP-HP, INSERM UMR S942, Université Paris Cité, Paris, France; [5]Department of Hematology, Cancer Research Center of Toulouse, Team 9, INSERM Unité Mixte de Recherche 1037, Toulouse, France; [6]Department of Nuclear Medicine, Hôpital Cochin, AP-HP, Université Paris Cité, Paris, France; [7]Assistance Publique-Hôpitaux de Paris, Université de Paris, and Hemato-Oncologie, Hôpital Saint-Louis, Paris, France; and [8]Department of Hematology and INSERM 1231, CHU Dijon Bourgogne, Dijon, France

Learning Objectives: On successful completion of this activity, participants should be able to describe (1) the difference between prognostic indices and predictive indices; (2) the differences between prognostic scores (advantages vs. disadvantages) to determine the best; (3) how to set up a clinical trial to evaluate the predictive values of a biomarker.

Financial Disclosure: Dr. Casasnovas is a consultant for Roche, Takeda, Gilead/Kite, MSD, and BMS and has received research grants from Takeda and Gilead/Kite. Dr. Thieblemont is a board member for Gilead/Kite, Novartis, BMS, and Incyte. Dr. Vercellino is a consultant for Gilead/Kite and Genmab and has received hospitality or transportation fees from MSD France, Siemens Healthcare, and Sanofi Aventis. The authors of this article have indicated no other relevant relationships that could be perceived as a real or apparent conflict of interest.

CME Credit: SNMMI is accredited by the Accreditation Council for Continuing Medical Education (ACCME) to sponsor continuing education for physicians. SNMMI designates each *JNM* continuing education article for a maximum of 2.0 AMA PRA Category 1 Credits. Physicians should claim only credit commensurate with the extent of their participation in the activity. For CE credit, SAM, and other credit types, participants can access this activity through the SNMMI website (http://www.snmmilearningcenter.org) through November 2027.

Understanding the differences between prognostic and predictive indices is imperative for medical research advances. We have developed a new prognostic measure that will identify the strengths, limitations, and potential applications in clinical practice.

Key Words: comparison; imaging; prognostic

J Nucl Med 2024; 65:1672–1678
DOI: 10.2967/jnumed.123.267021

The goal of this article is to explain how to compare prognostic indices and how to demonstrate the predictive impact of a measure through clinical trials. Prognostic indices in oncology have increased in sophistication and utility over the last decade. This evolution can be attributed to advancements in technology (such as radiomics from PET scans as well as development in data science using artificial intelligence techniques). This progress has been facilitated by improvements in data availability and quality (including data warehouses) associated with a better understanding of complex models (with statistical tools), as well as the development of personalized medicine. However, few were defined as clinically useful (*1*). A variety of methodologic problems could explain this loss of promising prognostic indices. Some guidelines are suggested in this article to provide relevant information to researchers about patient characteristics, statistical methods, and study designs.

Received May 15, 2024; revision accepted Sep. 10, 2024.

For correspondence or reprints, contact Loïc Chartier (loic.chartier@lysarc.org).

DIFFERENCE BETWEEN PROGNOSTIC AND PREDICTIVE MEASURES

Before comparing prognostic indices, it is crucial to understand the difference between prognostic and predictive measures (*2*). The confusion between prognostic and predictive variables is often due to their overlapping concepts, but they have distinct meanings and implications in medical research and clinical practice.

Prognostic measures provide information about the outcome of a disease (e.g., progression-free survival [PFS] or overall survival [OS]) regardless of treatment. They can be used to determine patient risk and, therefore, may guide therapy choices. From a statistical point of view, the biomarker is an explanatory variable, and the question is whether this biomarker has a mathematic relationship with the outcome. Vercellino et al. (*3*) showed that total metabolic tumor volume at baseline was prognostic of survival outcomes in diffuse large B-cell lymphoma (DLBCL) patients receiving either lenalidomide maintenance or placebo. Patients with a high total metabolic tumor volume had worse PFS and OS than patients with a low total metabolic tumor volume, whatever the treatment group (lenalidomide or placebo arm).

Regarding predictive measures, they determine how a patient will respond to a specific treatment. This indicator helps clinicians identify patients who are most likely to benefit from a particular therapy or those who are unlikely to respond, avoiding

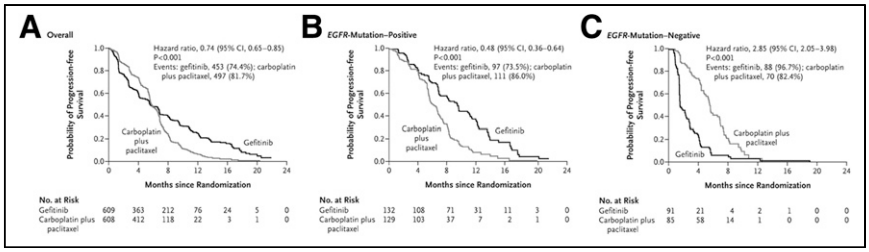

FIGURE 1. (A) PFS for overall population of adenocarcinoma lung patients suggesting 2 subpopulations. (B) PFS for patients with positive EGFR mutation showing that gefitinib is greater than carboplatin-paclitaxel. (C) PFS for patients with negative EGFR mutation showing that carboplatin plus paclitaxel is greater than gefitinib. (Reprinted with permission of (*4*).)

unnecessary treatment. From a statistical point of view, testing for an interaction between the biomarker and the treatment group can determine whether the biomarker is predictive. For illustration, Mok et al. (*4*) observed in a trial enrolling lung adenocarcinoma patients that PFS was greater with carboplatin-paclitaxel in the first 6 mo and greater with gefitinib in the following 16 mo, suggesting 2 subpopulations (Fig. 1A). The interaction between treatment and epidermal growth factor receptor (EGFR) mutation was significant ($P < 0.001$), meaning that in the mutation-positive subgroup, PFS was significantly longer in patients receiving gefitinib than in those receiving carboplatin-paclitaxel (Fig. 1B), whereas in the mutation-negative subgroup, PFS was significantly shorter in those receiving gefitinib than in those receiving carboplatin-paclitaxel (Fig. 1C).

Note that some measures may have both prognostic and predictive properties and, therefore, can exhibit multifaceted relationships with disease outcomes and treatment responses. Ballman (*2*) illustrated prognostic and predictive properties for a biomarker. On the one hand, we observed predictive properties with a different treatment effect according to the status of the biomarker, with a larger treatment effect observed for biomarker-positive patients. On the other hand, we observed prognostic properties with longer survival in biomarker-positive patients than in biomarker-negative patients, independently of treatment group.

EVALUATION OF PROGNOSTIC INDEX PERFORMANCE

The purpose of a prognostic index is to provide a quantitative estimate of the outcome (e.g., PFS or OS) for an individual patient based on various prognostic factors. Prognostic indices play a critical role in clinical practice by providing clinicians, patients, and researchers with information on the likely outcome of a disease. By combining information from multiple prognostic factors, these indices help optimize patient care, improve clinical decision-making, and enhance the overall management of disease.

When a prognostic index is being created, evaluation of its performance is crucial. Several methods, such as discrimination and calibration, can be used to evaluate the performance of a prognostic index.

Discrimination refers to the ability of the prognostic index to distinguish between patients with different outcomes. Sensitivity and specificity are 2 important metrics and can be summarized over a range of cut points for a continuous predictor using the receiver operating characteristic and the area under the curve (AUC).

Sensitivity measures the ability of the prognostic index to correctly identify patients with the outcome of interest (e.g., disease progression or death). A high sensitivity indicates that the prognostic index has a low rate of false negativity, meaning it rarely misses patients who have the outcome. Specificity measures the ability of the prognostic index to correctly identify patients without the outcome of interest. A high specificity indicates that the prognostic index has a low rate of false positivity, meaning it rarely misclassifies patients without the outcome. The AUC is a valuable metric for quantifying the discriminatory power and overall performance of binary classification models such as prognostic index. The AUC ranges from 0.5 (no discriminatory power) to 1 (excellent discrimination).

The concordance index (C-index) (*5, 6*) is a measure of predictive accuracy commonly used in survival analysis. A higher C-index (>0.6) indicates better discriminative ability, suggesting that the prognostic index is more effective at distinguishing between patients who experience the event and those who do not. However, the value of the C-index depends on the specific context of the study (in which the prognostic index is built) and the distribution of the outcome in the population. For example, the C-index was calculated for the combined total metabolic tumor volume and performance status (assessed using Eastern Cooperative Oncology Group score (*7*)) in different populations (clinical trials [GOYA and PETAL] and real-life data) and for different outcomes (PFS and OS). The C-index ranged from 0.6045 in the GOYA trial to 0.655 in the PETAL trial for PFS and from 0.6237 to 0.666 for OS.

The prognostic index should be not only discriminative but also accurate in estimating the likelihood of an event that is well calibrated. Calibration refers to the agreement between the predicted probabilities of an event (from the prognostic index) and the observed frequencies of that event in a dataset. Calibration plots can also be used to visually compare the predicted probabilities from the prognostic index against the observed frequencies of the outcome. The plot usually consists of bins or groups of patients with similar predicted probabilities, and for each group, the average predicted probability is plotted against the observed frequency of the outcome within that group. A prognostic index close to 45° line shows perfect calibration. Jelicic et al. (*8*) presented calibration curves for the 13 prognostic indices in DLBCL patients and concluded that prognostic indices with the highest C-index (National Comprehensive Cancer Network–International Prognostic Index [IPI] and DLBCL Prognostic Index) also had a good calibration (Fig. 2).

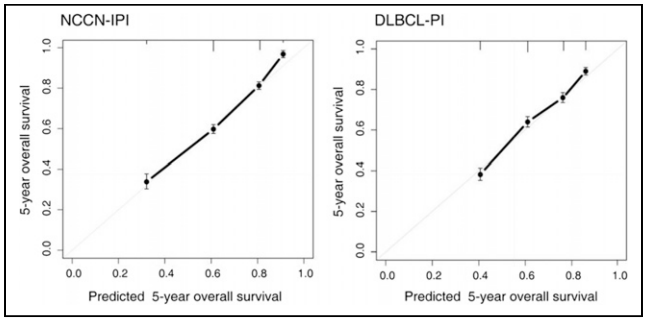

FIGURE 2. Calibration curves for National Comprehensive Cancer Network–IPI (NCCN-IPI) and DLBCL Prognostic Index (DLBCL-PI) in DLBCL patients concerning OS. (Reprinted with permission of (*8*).)

An overall performance measure is the Brier score (*9*), with lower values (closest to 0) indicating better performance. It represents the prediction error. A well-calibrated prognostic index is crucial for informed decision-making and risk assessment.

In summary, a valuable prognostic index should demonstrate good performance with good discrimination (AUC, C-index) and a good calibration (plot, Brier score) before being compared with other prognostic indices.

HOW TO DEFINE LOW- AND HIGH-RISK PATIENTS

Defining high- and low-risk patients depends on the context of the disease or outcome of interest. Threshold can be established from the prognostic indices to define different risk groups.

For example, the IPI is a widely used prognostic tool in DLBCL (*10*) that helps stratify patients into risk groups based on several clinical factors (age [>60 y vs. ≤60 y], Ann Arbor stage [stage III/IV vs. stage I/II], performance status [Eastern Cooperative Oncology Group score ≥ 2 vs. < 2], serum lactate dehydrogenase level [elevated vs. normal], and number of extranodal sites involved [≥2 vs. <2]). The IPI score varies from 0 (no factor met) to 5 (all factors met), and patients are classified from low-risk (IPI score of 0–1) to high-risk (IPI score of 4–5). This score is easy for clinicians to calculate and use to determinate the appropriate treatment approach.

Clinicians can also establish cutoffs from a continuous prognostic index to classify patients into different risk groups. For example, a continuous prognostic index was developed for patients with mantle cell lymphoma (*11*). The Mantle Cell Lymphoma–IPI score defines 3 risk groups: high-risk patients with a score of 6.2 or higher, intermediate-risk patients with a score of 5.7 to less than 6.2, and low-risk patients with a score of less than 5.7. This score is more complex for clinicians to use, requiring a calculator to determine the risk group.

Some prognostic models provide a probability for reaching specific outcomes rather than defining patients into predefined risk groups. For example, the International Metabolic Prognostic Index (IMPI) was developed for DLBCL patients (*12*). The IMPI provides a continuous risk score that estimates the probability of 3-y PFS based on factors such as metabolic total volume, age, and Ann Arbor stage. Instead of categorizing patients into risk groups such as low, intermediate, or high, the IMPI assigns a numeric score to each patient, and this score is associated with a probability of PFS. The calculation of this probability cannot be performed manually by the clinician; a tool was therefore developed for this purpose. The clinician specifies the value of the patient for each parameter, and the tool calculates the probability. Thus, a 30-y-old patient with a metabolic total volume of 250 cm^3 and an Ann Arbor stage of III has a 3-y PFS of 80% whereas a 60-y-old patient with a metabolic total volume of 500 cm^3 and an Ann Arbor stage of IV has a 3-y PFS of 69%. This approach can offer a more personalized risk assessment and can improve clinical decision-making compared with predefined risk groups. However, it is essential to validate individual probability models rigorously and ensure that they demonstrate a robust predictive performance across diverse patient populations and clinical settings.

HOW TO COMPARE VALUES OF DIFFERENT PROGNOSTIC INDICES

Most prognostic indices use predefined risk groups. High-risk patients may require alternative treatment strategies such as chimeric antigen receptor T-cell therapy or bispecific antibody in lymphoma, whereas low-risk patients may benefit from less intensive therapies, thus avoiding unnecessary toxicities. Therefore, it is crucial to define the purpose of the prognostic indices and to define the target population for which the prognostic index is intended. Describing the proportion of patients in each risk group provides information on the clinical relevance and applicability of prognostic indices. This descriptive analysis allows the identification of similarities and differences in risk stratification between different prognostic indices.

To get a comprehensive picture of the prognostic indices, it is essential to describe the outcome for each prognostic index. For time-to-event outcome, several measures should be described, such as median survival time, survival probabilities at specified time points (e.g., 1-y PFS and 3-y OS), and hazard ratio (HR) between risk groups. Kaplan–Meier curves of time-to-event outcome can be presented by risk group for each prognostic index. Jelicic et al. (*8*) described the distribution of patients according to 3 risk-group models as well as the 3- and 5-y OS. Kaplan–Meier curves were also generated for the 13 prognostic indices.

If the prognostic indices are continuous, it could be interesting to check the correlation between them. Several statistical methods can be used to assess the strength and direction of the relationship between continuous prognostic indices. Common correlation coefficients include the Pearson correlation coefficient and the Spearman rank correlation coefficient. Whatever the statistical method, a correlation coefficient close to 1 indicates a strong positive correlation, meaning that the prognostic indices tend to move together, whereas a correlation coefficient of 0.3–0.7 indicates a moderate positive correlation. A weak positive correlation is observed with a correlation coefficient of 0–0.3.

HOW TO COMPARE MODELS ON WHICH PROGNOSTIC INDICES ARE BASED

Comparing prognostic indices involves a systematic process to evaluate their performance, reliability, and clinical utility. Several performance measures can be used to compare prognostic indices (*13*) and are summarized in Table 1.

First, the likelihood ratio test can be used to compare the fit of 2 models when one is nested within the other. This statistical method is also often used in multivariable regression analysis (stepwise regression) to reduce model complexity and improve interpretability. It is particularly useful when one has fitted a more complex model and wants to assess whether a simpler model fits the data as well as the complex one. For example, the prognostic index from the PRIMA trial (*14*) (including β_2-microglobulin and bone marrow involvement) is a simplified scoring system in de novo follicular lymphoma and is nested within the Follicular Lymphoma International Prognostic Index score (including β_2-microglobulin, bone marrow involvement, hemoglobin, age, and longest diameter of largest involved node). The likelihood ratio test can be used to compare PRIMA Prognostic Index scores with Follicular Lymphoma International Prognostic Index scores and to assess whether the former fits the data as well as the latter. The question is whether the 3 parameters (hemoglobin, age, and longest diameter of largest involved node) not included in the PRIMA Prognostic Index score significantly improve the fit to the data.

The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) can also be used to compare models (not necessarily nested) and determine which best balances goodness

TABLE 1
Interpretation of Performance Measures

Measure	Interpretation	Warning
Likelihood ratio test (for nested models only)	Significant result: complex model gives better results than simpler model; no significant result: simpler model fits data as well as complex one	Depends on specific context of study and incidence of outcome in population
AIC and BIC (for nested and nonnested models)	Lower AIC and BIC indicates better trade-off on model fit and complexity; prognostic index with lowest AIC and BIC is best	
Calibration (Brier score)	Brier score closer to 0 means excellent calibration; prognostic index with lowest Brier score is best	
Discrimination (AUC and C-index [survival analysis])	Higher AUC indicates better discriminative ability; higher C-index indicates better discriminative ability	
Clinical risk reclassification (categoric and category-free Net Reclassification Index)	Not recommended	
Net benefit	Positive net benefit suggests that using prognostic index leads to more appropriate treatment recommendations than treating all patients or treating none	Calculate net benefit according to different weights to proportion of unnecessary intervention in order to evaluate performance and clinical impact across spectrum of decision scenarios

of fit and model complexity. For nonnested models, the likelihood ratio test cannot be used and only the AIC and BIC can be interpreted to define which model is best. A lower AIC indicates a better trade-off between model fit and complexity. It is crucial to have the same sample between prognostic indices to assess the AIC and BIC. For example, the IMPI (*12*) and IPI scores are nonnested models because metabolic tumor volume is included in the IMPI score and not in the IPI score, and lactate dehydrogenase, performance status (Eastern Cooperative Oncology Group score), and extranodal involvement are included in the IPI score but not in the IMPI score. Age and Ann Arbor stage are common between both scores but are considered differently. AIC was used in this study, and IMPI was defined as better than IPI in the prediction of outcomes.

As shown previously, the performance of each prognostic index can be assessed with different statistical metrics such as C-index, calibration plots, Brier score, sensitivity, specificity, positive predictive value, and negative predictive value. Jelicic et al. (*8*) compared 13 prognostic indices in DLBCL. On the basis of statistical metrics, the 3 best models were National Comprehensive Cancer Network–IPI, DLBCL Prognostic Index, and modified National Comprehensive Cancer Network–IPI, with a lower AIC and BIC indicating a better model fit and a higher AUC/C-index indicating better model discrimination.

To assess the prediction increment when a new biomarker is incorporated, several statistical methods can be used. The improvement in the AUC with the new prognostic index compared with the old one can be difficult to demonstrate. The Net Reclassification Index (*15*) is another way to measure the improvement in risk prediction performance when a new prognostic index is compared with the old one. It evaluates whether the new prognostic index correctly classifies patients into higher or lower risk categories compared with the old prognostic index. Even if the net reclassification improvement is easy to calculate, several limitations were identified such as the dependence on arbitrary risk categories and sensitivity to the choice of cutoffs for the category-based Net Reclassification Index, the potential for misinterpretation of the Net Reclassification Index (in terms of discrimination or calibration, whereas it assesses reclassification improvement only), and the use of the category-free Net Reclassification Index, which may lose the clinical intuition (*16–18*).

HOW TO EVALUATE CLINICAL IMPACT BETWEEN PROGNOSTIC INDICES

All these statistical measures give information about model improvement between prognostic indices but do not assess clinical impact. To assess clinical utility, it is necessary to calculate the net benefit (*19, 20*), which assesses the clinical relevance and usefulness of a prognostic model by quantifying the balance between benefits and harms. Net benefit is based on the proportion of justified interventions (true positives, or correctly identified outcomes) minus the proportion of unnecessary interventions (false positives, or incorrectly identified outcomes). This statistic assigns a weight to the proportion of unnecessary interventions. A highly effective intervention with few side effects suggests the use of a low weight. Different clinicians might prefer to use different weights to evaluate the performance and clinical impact of the prognostic index across a spectrum of decision scenarios. A positive net benefit suggests that using the prognostic index leads to more appropriate treatment recommendations than treating all patients or treating none.

After calculating performance measures (calibration, discrimination, and prediction error), Geloven et al. (*21*) generated a decision curve showing the net benefit for predicting breast cancer recurrence within 5 y. Among the 115 patients with an event predicted by the model, 34 were correctly identified as high-risk and

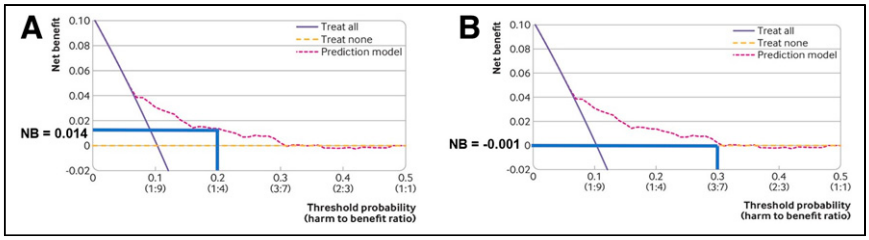

FIGURE 3. Decision curve showing net benefit according to 2 different thresholds: weight of 20% for proportion of unnecessary intervention (A) and weight of 30% for proportion of unnecessary intervention (B). NB = net benefit. (Reprinted with permission of (*21*).)

were recommended for chemotherapy (benefit), and 81 were incorrectly identified as high-risk and were overtreated (harm). If we consider a weight of 0.2 (20%) (Fig. 3A) in the proportion of unnecessary intervention, the net benefit is 0.014 [34 − (81 × 0.2)], indicating that there will be 14 net benefiting patients when the prediction model is applied to 1,000 patients. If we consider a weight of 0.3 (30%) (Fig. 3B), indicating that we have more side effects with the chemotherapy, the net benefit is 0 [34 − (81 × 0.3)], indicating that no net benefit is expected with the prediction model.

HOW TO VALIDATE A PROGNOSTIC INDEX

Prognostic indices have been increasingly being developed and published over the last decade in oncology, but few were externally validated.

Internal validation is used by splitting the dataset into a training subset (used for model development) and a validation subset (used for model evaluation). The training set is used to select relevant prognostic factors and build the prognostic index, whereas the validation set is used to assess its performance (discrimination, calibration, and prediction error). However, some risk of overfitting exists with internal validation. To limit this risk, several statistical approaches can be used, such as bootstrapping, cross-validation, or penalized regression methods (e.g., LASSO [least absolute shrinkage and selection operator] regression). However, these approaches provide information on the reproducibility of the prognostic index and not on the generalizability. Only external validation can provide evidence of the generalizability to various patient populations (*22*).

Validating a prognostic index (*22, 23*) involves assessing its performance, reliability, and generalizability using independent datasets (from other clinical trials or real-life data). External validation is crucial to assess the generalizability of the prognostic index compared with other prognostic indices across different patient cohorts and clinical settings. Real-world data validation can demonstrate the prognostic index's performance outside the controlled environment of clinical trials. Thieblemont et al. (*7*) used different cohorts (GOYA and PETAL trials as well as real-life data) to validate the combination of total metabolic tumor volume with Eastern Cooperative Oncology Group developed initially with data from the REMARC study (*3*).

A 22-item checklist was provided in the TRIPOD statement (*24*) to improve the reporting of studies developing, validating, or updating a prediction model for either diagnostic or prognostic purposes.

HOW TO IDENTIFY A PREDICTIVE MEASURE

A predictive measure is an indicator of how likely a patient is to respond to a specific treatment. A biomarker will be defined as predictive if the interaction between it and the treatment group is significant. This result can be observed in exploratory analyses. Mok et al. (*4*) observed a significant interaction ($P < 0.001$) between treatment and EGFR mutation, meaning that PFS was significantly longer among patients receiving gefitinib in the mutation-positive subgroup. To formally demonstrate the predictive value of this biomarker, it is crucial to perform a dedicated clinical trial.

Before initiating a clinical trial, it is important to have a clear hypothesis regarding the potential predictive value of the biomarker. This hypothesis should be based on preclinical evidence, exploratory analyses, or a biologic rationale suggesting that the biomarker may be associated with treatment response, disease progression, or clinical outcomes. The future clinical trial should incorporate the biomarker assessment as an integral component of the study protocol. In our example, the hypothesis was based on exploratory analyses showing that gefitinib is better than carboplatin-paclitaxel in patients selected by EGFR mutation. To demonstrate the predictive value of EGFR mutation, 2 clinical trials (*25, 26*) including patients with only EGFR mutation were performed. Significant results were observed in favor of gefitinib, confirming the need to use gefitinib in patients selected by EGFR.

Another approach was used to demonstrate treatment effect in the overall population and in a specific subgroup. The predictive biomarker is present in the study to manage the possibility that the treatment effect might be observed only in the subgroup. In this case, subgroup analysis should be prespecified in the study protocol to minimize the risk of bias and data-driven results. Defining coprimary endpoints requires careful consideration and should adhere to certain principles to ensure scientific rigor and interpretability of study results. Statistical methods for handling multiple comparisons should be clearly outlined to avoid inflated type I error rates.

Cappuzzo et al. (*27*) performed a placebo-controlled phase 3 study to assess use of erlotinib as maintenance therapy in patients with nonprogressive non–small cell lung cancer after first-line platinum-doublet chemotherapy. Coprimary endpoints were PFS in all patients, irrespective of EGFR status, and PFS in patients with tumors that overexpress EGFR. Interpretation of each coprimary endpoint included adjustments for multiplicity (3% for PFS in all patients and 2% for PFS in the subgroup with EGFR) to control an overall 2-sided 5% type I error rate. Median PFS was significantly longer with erlotinib than with placebo in the overall population (HR, 0.71; 95% CI, 0.62–0.82; $P < 0.0001$) and in patients with EGFR-positive immunohistochemistry (HR, 0.69; 95% CI, 0.58–0.82; $P < 0.0001$). Despite the significant results, the predictive value of EGFR immunohistochemistry status was not demonstrated in this clinical trial, mainly because of absence of treatment effect assessment in patients with EGFR-negative immunohistochemistry (Table 2). The predictive value of EGFR immunohistochemistry status would be demonstrated if significant results were observed in the subgroup and not in the overall population. This example shows limitations in demonstrating the predictive value of the biomarker.

The study design illustrated in Figure 4 could be used to demonstrate a significant interaction (defined as a primary endpoint) between the biomarker result and the treatment. The use of this study design has several advantages such as demonstration of a

TABLE 2
Interpretation of Results to Know Whether Biomarker Is Predictive

Treatment benefit			
Overall population	Patients EGFR-positive	Patients EGFR-negative	Interpretation
Yes	Yes	Not done (in study)	Treatment benefit not demonstrated in patients with negative biomarker because of low proportion of patients with negative biomarker OR very high treatment benefit observed in patients with positive biomarker; loss of predictive marker
No	Yes	No	Treatment benefit proven in EGFR-positive subgroup; predictive biomarker demonstrated
Yes	Yes	Yes	Treatment benefit demonstrated in overall population as well as in each subgroup; predictive biomarker not demonstrated

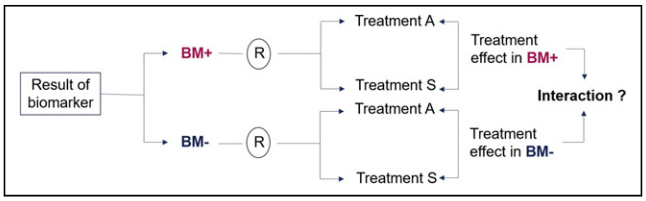

FIGURE 4. Study design with interaction. BM- = bone marrow–negative; BM+ = bone marrow–positive; R = randomization; S = standard of care.

significant interaction or observation of a treatment benefit in one subgroup but not in the other. Each subgroup should be randomized to ensure that potentially confounding variables are balanced across treatment groups, reducing the risk of bias in the estimation of treatment effects. We can wonder what the differences are between a study design with interaction and a study design with randomization stratified by the biomarker. The major difference between the 2 study designs is the primary criterion. The first investigates interaction effects between treatment and biomarker, whereas the second assesses the treatment effect on ensuring a balanced allocation across treatment groups within biomarker subgroups. Sample size should be calculated to ensure that the study is adequately powered to detect meaningful interaction effects.

Gregorc et al. (*28*) performed a biomarker-stratified randomized phase 3 trial in which the primary criterion was the existence of a significant interaction based on OS between the serum protein test classification and treatment. The interaction was significant ($P = 0.017$), with worse survival with erlotinib than with chemotherapy (standard treatment) in non–small cell lung cancer patients with a proteomic test classification of poor (HR, 1.72; 95% CI, 1.08–2.74; $P = 0.022$) and no significant difference in patients with a proteomic test classification of good (HR, 1.06; 95% CI, 0.77–1.46; $P = 0.71$). It was expected that patients treated by erlotinib would survive longer than chemotherapy patients in this last subgroup (classification of good), which was not the case. On the contrary, the results were paradoxic in showing a tendency for erlotinib to be inferior to chemotherapy on PFS (HR, 1.26; 95% CI, 0.94–1.69; $P = 0.129$). The trial did not show erlotinib to be superior in patients with a proteomic test classification of good, even if the interaction was significant. Therefore, it is not possible to conclude that this biomarker has predictive value. This example illustrates the fact that a significant interaction only is not sufficient to validate a predictive biomarker.

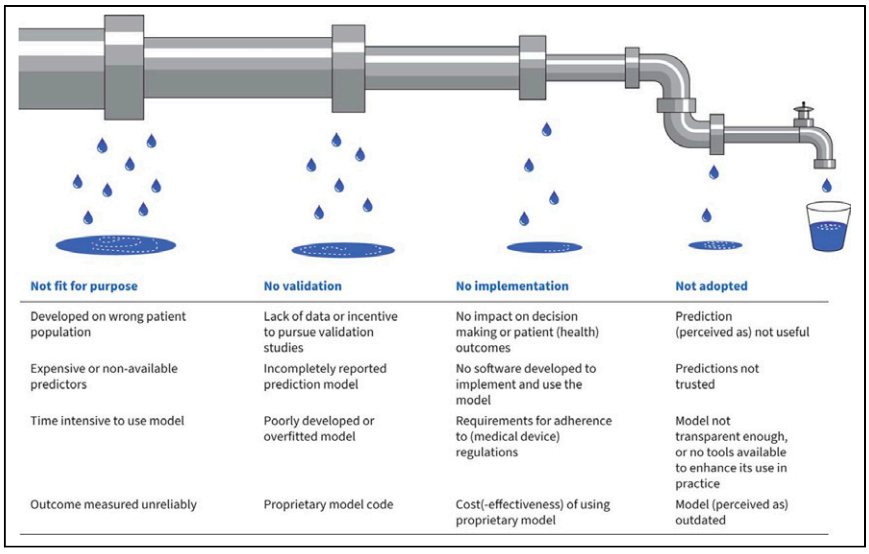

FIGURE 5. Prognostic model adoption pipeline. (Reprinted with permission of (*28*).)

Despite hundreds of publications on prognostic or predictive indices, relatively few of them find their way into routine clinical use. van Royen et al. (*29*) illustrates this loss of prognostic indices with a very interesting leaky pipeline for prognostic model adoption (Fig. 5).

CONCLUSION

This article provides an overview of the development and validation of a new prognostic index. It is crucial to understand that prognostic and predictive measures have distinct meanings and, therefore, different clinical uses. By systematically evaluating the discrimination, calibration, clinical utility, comparative performance, and validation of a prognostic index, we

will have relevant information about strengths, limitations, and potential applications in clinical practice. A prognostic index should not be used to decide on specific therapies for patients, in order to avoid incorrectly depriving them of a potentially useful treatment. A predictive biomarker can be used to decide on a treatment strategy, but a formal impact trial should be performed to obtain definite evidence of the usefulness of a prediction model for clinical decision-making. Future studies on predictive factors might give rise to novel individualized treatment strategies. It is necessary to develop a framework that allows the continuous updating of risk assessments as new data become available. Le Gouill et al. (*30*) mentioned that a PET-driven approach based on changes in SUV_{max} can provide early identification of patients with a high risk of relapse, for whom innovative therapeutic solutions are needed. Kurtz et al. (*31*) described a method to dynamically determine outcome probabilities for individual patients using risk predictors acquired over time. Continuous iteration and improvement of the dynamic risk profiling framework through incorporation of additional data may enhance accuracy and reliability. This process may involve recalibrating the prognostic index, retraining machine learning models, or adjusting statistical parameters to reflect the latest information.

REFERENCES

1. McShane LM, Altman DG, Sauerbrei W, et al. REporting recommendations for tumour MARKer prognostic studies (REMARK). *Br J Cancer.* 2005;93:387–391.
2. Ballman KV. Biomarker: predictive or prognostic? *J Clin Oncol.* 2015;33:3968–3971.
3. Vercellino L, Cottereau AS, Casasnovas O, et al. High total metabolic tumor volume at baseline predicts survival independent of response to therapy. *Blood.* 2020;135:1396–1405.
4. Mok TS, Wu YL, Thongprasert S, et al. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *N Engl J Med.* 2009;361:947–957.
5. Harrell FE, Califf RM, Pryor DB, et al. Evaluating the yield of medical tests. *JAMA.* 1982;247:2543–2546.
6. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med.* 1996;15:361–387.
7. Thieblemont C, Chartier L, Dührsen U, et al. A tumor volume and performance status model to predict outcome before treatment in diffuse large B-cell lymphoma. *Blood Adv.* 2022;6:5995–6004.
8. Jelicic J, Juul-Jensen K, Bukumiric Z, et al. Prognostic indices in diffuse large B-cell lymphoma: a population-based comparison and validation study of multiple models. *Blood Cancer J.* 2023;13:157.
9. Graf E, Schmoor C, Sauerbrei W, et al. Assessment and comparison of prognostic classification schemes for survival data. *Stat Med.* 1999;18:2529–2545.
10. International Non-Hodgkin's Lymphoma Prognostic Factors Project. A predictive model for aggressive non-Hodgkin's lymphoma. *N Engl J Med.* 1993;329:987–994.
11. Hoster E, Dreyling M, Klapper W, et al. A new prognostic index (MIPI) for patients with advanced-stage mantle cell lymphoma. *Blood.* 2008;111:558–565.
12. Mikhaeel NG, Heymans MW, Eertink JJ, et al. Proposed new dynamic prognostic index for diffuse large B-cell lymphoma: International Metabolic Prognostic Index. *J Clin Oncol.* 2022;40:2352–2360.
13. Cook NR. Quantifying the added value of new biomarkers: how and how not. *Diagn Progn Res.* 2018;2:14.
14. Bachy E, Maurer MJ, Habermann TM, et al. A simplified scoring system in de novo follicular lymphoma treated initially with immunochemotherapy. *Blood.* 2018;132:49–58.
15. Pencina MJ, D'Agostino RB, Steyerberg EW, et al. Extension of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med.* 2011;30:11–21.
16. Leening MJG, Vedder MM, Witteman JCM, et al. Net reclassification improvement: computation, interpretation, and controversies—a literature review and clinician's guide. *Ann Intern Med.* 2014;160:122–131.
17. Kerr KF, Wang Z, Janes H, et al. Net reclassification indices for evaluating risk prediction instruments: a critical review. *Epidemiology.* 2014;25:114–121.
18. Pepe MS, Fan J, Feng Z, et al. The net reclassification index (NRI): a misleading measure of prediction improvement even with independent test data sets. *Stat Biosci.* 2015;7:282–295.
19. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making.* 2006;26:565–574.
20. Vickers AJ, Calster BV, Steyerberg E. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ.* 2016;352:i6.
21. van Geloven N, Giardiello D, Bonneville EF, et al. Validation of prediction models in the presence of competing risks: a guide through modern methods. *BMJ.* 2022;377:e069249.
22. Ramspek CL, Jager KJ, Dekker FW, et al. External validation of prognostic models: what, why, how, when and where? *Clin Kidney J.* 2020;14:49–58.
23. Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med.* 2000;19:453–473.
24. Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med.* 2015;162:W1-73.
25. Maemondo M, Inoue A, Kobayashi K, et al. Gefitinib or chemotherapy for non-small-cell lung cancer with mutated EGFR. *N Engl J Med.* 2010;362:2380–2388.
26. Mitsudomi T, Morita S, Yatabe Y, et al. Gefitinib versus cisplatin plus docetaxel in patients with non-small-cell lung cancer harbouring mutations of the epidermal growth factor receptor (WJTO3405): an open label, randomised phase 3 trial. *Lancet Oncol.* 2010;11:121–128.
27. Cappuzzo F, Ciuleanu T, Stelmakh L, et al. Erlotinib as maintenance treatment in advanced non-small-cell lung cancer: a multicentre, randomised, placebo-controlled phase 3 study. *Lancet Oncol.* 2010;11:521–529.
28. Gregorc V, Novello S, Lazzari C, et al. Predictive value of a proteomic signature in patients with non-small-cell lung cancer treated with second-line erlotinib or chemotherapy (PROSE): a biomarker-stratified, randomised phase 3 trial. *Lancet Oncol.* 2014;15:713–721.
29. van Royen FS, Moons KGM, Geersing GJ, et al. Developing, validating, updating and judging the impact of prognostic models for respiratory diseases. *Eur Respir J.* 2022;60:2200250.
30. Le Gouill S, Ghesquières H, Oberic L, et al. Obinutuzumab vs rituximab for advanced DLBCL: a PET-guided and randomized phase 3 study by LYSA. *Blood.* 2021;137:2307–2320.
31. Kurtz DM, Esfahani MS, Scherer F, et al. Dynamic risk profiling using serial tumor biomarkers for personalized outcome prediction. *Cell.* 2019;178:699–713.e19.