ORIGINAL ARTICLE

External validation and recalibration of the Brock model to predict probability of cancer in pulmonary nodules using NLST data

Audrey Winter, Denise R Aberle, William Hsu

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/thoraxjnl-2018-212413).

Department of Radiological Sciences, Medical Imaging Informatics, University of California, Los Angeles, California, USA

Correspondence to
Dr Audrey Winter, Department of Radiological Sciences, Medical Imaging Informatics, University of California, Los Angeles, CA 90095, USA;
audrey.winter89@gmail.com

Received 26 July 2018
Revised 29 January 2019
Accepted 4 February 2019
Published Online First
21 March 2019

ABSTRACT

Introduction We performed an external validation of the Brock model using the National Lung Screening Trial (NLST) data set, following strict guidelines set forth by the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis statement. We report how external validation results can be interpreted and highlight the role of recalibration and model updating.

Materials and methods We assessed model discrimination and calibration using the NLST data set. Adhering to the inclusion/exclusion criteria reported by McWilliams *et al*, we identified 7879 non-calcified nodules discovered at the baseline low-dose CT screen with 2 years of follow-up. We characterised differences between Pan-Canadian Early Detection of Lung Cancer Study and NLST cohorts. We calculated the slope on the prognostic index and the intercept coefficient by fitting the original Brock model to NLST. We also assessed the impact of model recalibration and the addition of new covariates such as body mass index, smoking status, pack-years and asbestos.

Results While the area under the curve (AUC) of the model was good, 0.905 (95% CI 0.882 to 0.928), a histogram plot showed that the model poorly differentiated between benign and malignant cases. The calibration plot showed that the model overestimated the probability of cancer. In recalibrating the model, the coefficients for emphysema, spiculation and nodule count were updated. The updated model had an improved calibration and achieved an optimism-corrected AUC of 0.912 (95% CI 0.891 to 0.932). Only pack-year history was found to be significant ($p<0.01$) among the new covariates evaluated.

Conclusion While the Brock model achieved a high AUC when validated on the NLST data set, the model benefited from updating and recalibration. Nevertheless, covariates used in the model appear to be insufficient to adequately discriminate malignant cases.

Key messages

What is the key question?

► How does following the guidelines set forth in the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement help appraise the suitability of using the Brock model by assessing both its discrimination and calibration on a target screening population?

What is the bottom line?

► While the Brock model achieves high area under the curve (AUC) on an external data set (National Lung Screening Trial), further inspection of the model's discrimination and calibration reveals opportunities to improve the transportability of the model through recalibration, revision and extension.

Why read on?

► We perform an external validation of the Brock model, adhering to the guidelines set forth by the TRIPOD statement. While metrics such as the AUC are consistent between the derivation and validation data sets, a lack of calibration was noted. Moreover, the positive predictive value, influenced by the disease prevalence, was low. We demonstrate the impact of model recalibration, revision and/or extension when seeking to adopt the Brock model on a target population that is different from the derivation population.

INTRODUCTION

Lung cancer screening plays a critical role along with smoking cessation in reducing the mortality associated with lung cancer. Based on the results of the National Lung Screening Trial (NLST),[1] in which participants underwent up to three low-dose CT (LDCT) screening exams, a 20% reduction mortality was observed in comparison with individuals who underwent comparable screening with chest X-ray. However, using a nodule size threshold of a 4 mm diameter to define positive screens, the trial also reported 96.4% of positive screenings were actually false positive results in the LDCT arm. It is well recognised that false positive screens are major contributors to the potential harms of screening due to unnecessary downstream invasive procedures that may be associated with complications. A reliable diagnostic prediction model that estimates the probability of lung cancer in the setting of LDCT-detected indeterminate nodules would better define the level of risk of lung cancer and inform clinical decision-making. To aid radiologists with assessing the clinical significance of nodules identified on an LDCT or chest X-ray exam, several diagnostic prediction models have been proposed to estimate the probability of lung cancer for a patient given the presence of one or

To cite: Winter A, Aberle DR, Hsu W. *Thorax* 2019;**74**:551–563.

Table 1 Articles about application of the full Brock model[2] on external data sets

Articles	Nodules, n	Period	Events, n	Follow-up	Discrimination	Calibration
Zhao *et al*[15]	52	December 2007 onwards	5	2–6 years	AUC	Hosmer-Lemeshow test
Al-Ameri *et al*[16]	244 patients*	2008–2013	99	2 years	AUC	–
Nair *et al*†‡[17]	6568 patients	2002–2007	220	1 year	AUC	Calibration plot
She *et al*[18]	1798 patients*	2014–2015	1198	–	AUC	–
Schreuder *et al*†[19]	24 542	2002–2007	174	1–2 years	AUC	Hosmer-Lemeshow test
Winkler Wille *et al*[20]	1152	2004–2006	233	–	AUC	–
Talwar *et al*[21]	702*	2012–2013	323	2 years	AUC	Calibration plot
White *et al*†[22]	4431	2002–2007	116	Up to 7 years	AUC	–
Chung *et al*[23]	872 for centre A and 974 for centre B*	2004–2012	177 for centre A and 264 for centre B	Up to 10 years	AUC	–
Yang *et al*[24]	242*	2015	187	–	AUC	–

*Non-screening population.
†Articles using the National Lung Screening Trial (NLST) data set.
AUC, area under the curve.

more nodules.[2–10] However, individual radiologists and institutions must decide if the model is appropriate for their patient population. Determining whether or not a prediction model is capable of being generalised to a new population is called external validation. External validation is strongly recommended for all prediction models, as stated in the 'Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis' (TRIPOD) statement.[11] TRIPOD provides guidelines for authors on key information needed to clearly describe a diagnostic or prognostic prediction model and for readers on how to interpret the validity and generalisability of these models. When evaluating the performance of a model, two fundamental aspects should be considered: (1) discrimination, which is the ability to discriminate among different outcomes (eg, patients predicted to be at higher risk should exhibit higher event rates than those considered at lower risk); and (2) calibration, which is the agreement between predicted and observed probabilities (eg, a well-calibrated risk score or prediction rule assigns the correct event probability at all levels of predicted risk).[12–14] In practice, however, many of these models are not externally validated and even when external validation is purportedly performed, the process and metrics used to perform external validation may not be fully accurate.

While a number of studies have attempted to externally validate prediction models for pulmonary nodules,[9 15–27] none of them completely adhere to the methods recommended by TRIPOD.[11] Among these models, the Brock model[2] has been recommended by organisations such as the British Thoracic Society[28] and has the most external validation studies reported.[15–24] Table 1 provides a summary of what has been done in those studies.

Only four of those studies[15 17 19 21] present both discrimination and calibration metrics when applying the model to a new population. In addition, only five of these studies reported sufficient sample size,[17 18 20 21 23] which recommends ideally 200 examples each of positive and negative events.[29] Moreover, five of these studies[16 18 21 23 24] applied the Brock model on non-screening patients, which is not consistent with the exclusion/inclusion criteria of the original model.

In this study, we evaluate the external validity of the Brock model[2] using the NLST data set. Compared with prior studies that have conducted a similar analysis,[15–24] we assess approaches for measuring the discrimination and calibration of the Brock model[2] and explore how issues identified by these metrics can be addressed through model recalibration, revision and extension,[14 30] following the TRIPOD statement[11] (checklist in online supplementary file 1).

MATERIALS AND METHODS

Study population

The NLST was a randomised controlled trial in which participants underwent three screenings with either LDCT or chest X-ray. Eligible participants were between 55 and 74 years of age at the time of randomisation; had a minimum of 30 pack-years (total years smoked × cigarettes per day/20) smoking; and if a former smoker had quit within the previous 15 years. All screening exams were completed from August 2002 through September 2007. Participants who had previously received a diagnosis of lung cancer, had undergone chest LDCT within 18 months before enrolment or had experienced haemoptysis or unexplained weight loss of more than 6.8 kg (15 lb) in the preceding year were excluded. A total of 53 452 participants were enrolled; 26 722 were randomly assigned to screening with LDCT and 26 730 to screening with chest radiography.[1] Participants were followed for lung cancer diagnoses that occurred through 31 December 2009.[1]

From the collected NLST data set, we identified the relevant data elements related to participant demographics, smoking history and LDCT screening results that were used as covariates in the Brock model (details of how data elements were identified are provided in the online supplementary table S1). As a guiding principle, the model should be applied in exactly the same way in the validation data set (NLST) as in the derivation data set (Pan-Canadian Early Detection of Lung Cancer Study [PanCan]),[12] following the same inclusion/exclusion criteria from McWilliams *et al*[2] and using the same covariates and duration of follow-up. Nodules with baseline LDCT screens that had missing values for covariates were excluded from our analysis. Notably, the NLST data set had sufficient screen-detected lung cancers to be appropriate for validation, ideally 200 (or more) events.[29]

One challenge was that data from the NLST were captured at the participant level whereas the PanCan study captured data at the nodule level. Thus, the abnormality that was directly related

Table 2 Recalibrating and model revision methods considered for logistic regression models for a prognostic index (PI) with n predictors (ie, methods 1–5) and extension methods considering k new predictors; not included in the original model (ie, methods 6–8)[14 30]

	No	Updating method	Description	Interpretation
	1	No adjustment	Fit the original model in exactly the same way it was constructed in the derivation data set. No parameters are estimated.	This method makes no revisions to the model but evaluates the performance of the original scores when applied to an independent cohort.
Recalibration	2	α	Correct the 'calibration in the large' by updating the intercept α.	If a lack of calibration is noted during the external validation process (ie, $\hat{\alpha} \neq 0$, $\hat{\beta}_{overall} \neq 1$), an initial 're-calibration step' is performed, in which α more or less $\beta_{overall}$ can be updated without changing the PI itself. Updating only the intercept (method 2) addresses problems arising from overly high or low predicted probabilities in the validation data set through only a change in the intercept ($\hat{\alpha} < 0$ or $\hat{\alpha} > 0$). Beyond updating the intercept, in method 3, the calibration slope can also be updated, in which all original regression coefficients can be multiplied by a constant. This addresses the problem of coefficients in the original model being too large: $\beta_{overall} < 1$ (or too small: $\beta_{overall} > 1$).
	3	$\alpha + \beta_{overall}$	Re-estimate both calibration slope $\beta_{overall}$ and intercept α. This step is called 'logistic calibration'.	
Model revision	4	$\alpha + \beta_{overall} + \gamma_{1\ldots n \mid AIC}$	Building on method 3, each covariate from the original model is included in a forward stepwise procedure using Akaike information criterion (AIC), including only those that lower (improve) the AIC. Their coefficients γ_i (for i in 1 to n) are then estimated. Of note, their updated regression coefficients β_i are then: $\beta_i = \gamma_i + \beta_{overall}\,\beta_i\ _{(original\ model)}$.	In the 're-calibration step' above, the PI is revised globally and equally for all regression coefficients estimated in the derivation data set. In method 4, covariates are included one by one if they improve the relative goodness of the fit of the model (ie, lower the AIC), which identifies covariates having a significantly different effect in the validation versus derivation data sets and then to recalibrate the PI by increasing ($\gamma_i > 0$) or decreasing ($\gamma_i < 0$) their weights in the model. Notably, in clinical practice, this procedure can account for big changes in covariates over time by recalibrating the model, for example. In method 5, all coefficients are re-estimated in the validation data set. This step also highlights covariates that have a different influence in derivation and validation data sets by comparing α and $\beta_{1\ldots n}$ with the original ones. With this method, the original PI is not preserved.
	5	$\alpha + \beta_{1\ldots n}$	Re-estimate coefficients of all the covariates: $\beta_{1\ldots n}$ included in the original model and the intercept α in the validation data set without any selection.	
Model extension	6	$\alpha + \beta_{overall} + \gamma_{1\ldots n \mid AIC} + \beta_{n+1\ldots n+k \mid AIC}$	Re-estimate the intercept α, the slope on the PI: $\beta_{overall}$ and the coefficients of the covariates included in the model: $\gamma_{1\ldots n \mid AIC}$, as well as coefficients of new covariates: $\beta_{n+1\ldots n+k \mid AIC}$ in a stepwise manner (as in method 4).	In the 'model revision' step, only the original covariates were used. In the 'model extension' step, new covariates not originally included in the model are added. Method 6 is similar to method 4 but adds new covariates only if they increase the relative goodness of the fit of the model (ie, lower AIC). For example, in clinical practice, this process enables the incorporation of new covariates into the model should they become available. Method 7 approximates method 5 but, as in method 6, adds new covariates only if they decrease the AIC of the model. Finally, in method 8, we create a completely new model without considering the original one by estimating all the regression coefficients (for covariates originally included as well as new covariates).
	7	$\alpha + \beta_{1\ldots n} + \beta_{n+1\ldots n+k \mid AIC}$	Fit the exact same model as in the original one; re-estimating all coefficients of the original covariates: $\alpha + \beta_{1\ldots n}$, without selection process, and adding new covariates $\beta_{n+1\ldots n+k \mid AIC}$ only if they improve model performance in the validation data set.	
	8	$\alpha + \beta_{1\ldots n+k}$	Fit a model which allows the estimation of all the n+k coefficients: $\beta_{1\ldots n+k}$ and the intercept α without selection process.	

to the lung cancer diagnosis could not be matched with certainty. To determine which nodules were malignant or not within an interval of 2 years from the baseline screening, we created an 'event' covariate that denoted whether the abnormality was lung cancer or not based on the reported anatomical location for the abnormality and the diagnosed cancer. Online supplementary table S1 shows the covariates used to create the event covariate. Follow-up times began at the time of the baseline screen and ended with whichever of the following events came first: diagnosis or 2-year follow-up.

Given the variety of data elements collected on NLST participants, we considered four additional covariates beyond what was used in the Brock model as part of model extension: body mass index, the smoking status at randomisation, number of pack-years and occupational exposure to asbestos for 1 or more years. We specifically targeted these covariates based on their association with lung cancer risk in previous prediction models.[2–4] Participants with missing values (n=20 patients) for these additional covariates were excluded from analysis; thus, the final data set is shown in online supplementary table S2.

Prognostic index

The main output of a logistic regression model, the type of model underlying the Brock model, is a prognostic index (PI), which is a weighted sum of the covariates X_i in the model, where the weights β_i are the regression coefficients and α is the intercept:

$$PI = \alpha + \beta_1 X_1 + \ldots + \beta_n X_n.$$

The validity of a logistic regression model can be evaluated using both qualitative (visual) and quantitative assessments.

Visual assessment of discrimination and calibration

To assess visually the discrimination and calibration of the Brock model when applied to NLST, we used three different visualisations[13]:

- Histograms of the PI per outcome (Y=0 or Y=1), where perfect discrimination would appear as non-overlapping histograms.
- A receiver operating characteristic (ROC) curve, which plots sensitivity (SE) against one minus specificity (SPC). The discrimination performance of an ROC curve is quantified

by the area under the curve (AUC) and CI. AUC values range from 0 to 1; the closer the AUC is to 1, the better the discrimination performance.

- A calibration plot, which reports graphically predicted outcome probabilities (on the x-axis) against observed outcome (on the y-axis).[11] A well-calibrated prediction implies that the curve lies on the diagonal, which means perfect agreement between the predicted probabilities and the observed outcomes.

Comparison of the derivation and validation data sets

As a quantitative assessment, we compared the distributions of covariates between PanCan and NLST data sets used in the Brock model using the Student's t-test, Wilcoxon-Mann-Whitney test, or χ^2 test, when appropriate. For each test, an effect size (ES) was also reported. For χ^2 tests, Cramér's V (ϕ_C) was calculated (ie, magnitude of the ES; small: $0.1 \leq ES < 0.3$, medium: $0.3 \leq ES < 0.5$ and large: $ES > 0.5$) and for Student's and Wilcoxon-Mann-Whitney tests r^2 and Cohen's r^2 (ie, magnitude of the ES; small: $0.01 \leq ES < 0.09$, medium: $0.09 \leq ES < 0.25$ and large: $ES > 0.25$), respectively.[31–33] The larger the ES is, the more important the impact of the findings are with all other factors being equal.

Assessment of the model calibration

To perform external validation of the Brock model, we estimated the regression coefficient on the PI and the intercept coefficient by fitting a logistic regression in the validation data set[14 30] as follows:

$$Y = \hat{\alpha} + \hat{\beta}_{overall}\, PI.$$

If perfectly calibrated, $\hat{\alpha} = 0$ and $\hat{\beta}_{overall} = 1$. If $\hat{\alpha}$ and/or $\hat{\beta}_{overall}$ are significantly different from 0 and 1 based on a likelihood ratio test, recalibration is needed.[14 30]

Of note, $\hat{\alpha} < 0$ (or $\hat{\alpha} > 0$) means that the predicted probabilities in the validation data set were too high (or too low).[14 30] With respect to $\hat{\beta}_{overall}$, if the calibration slope is smaller than 1, then

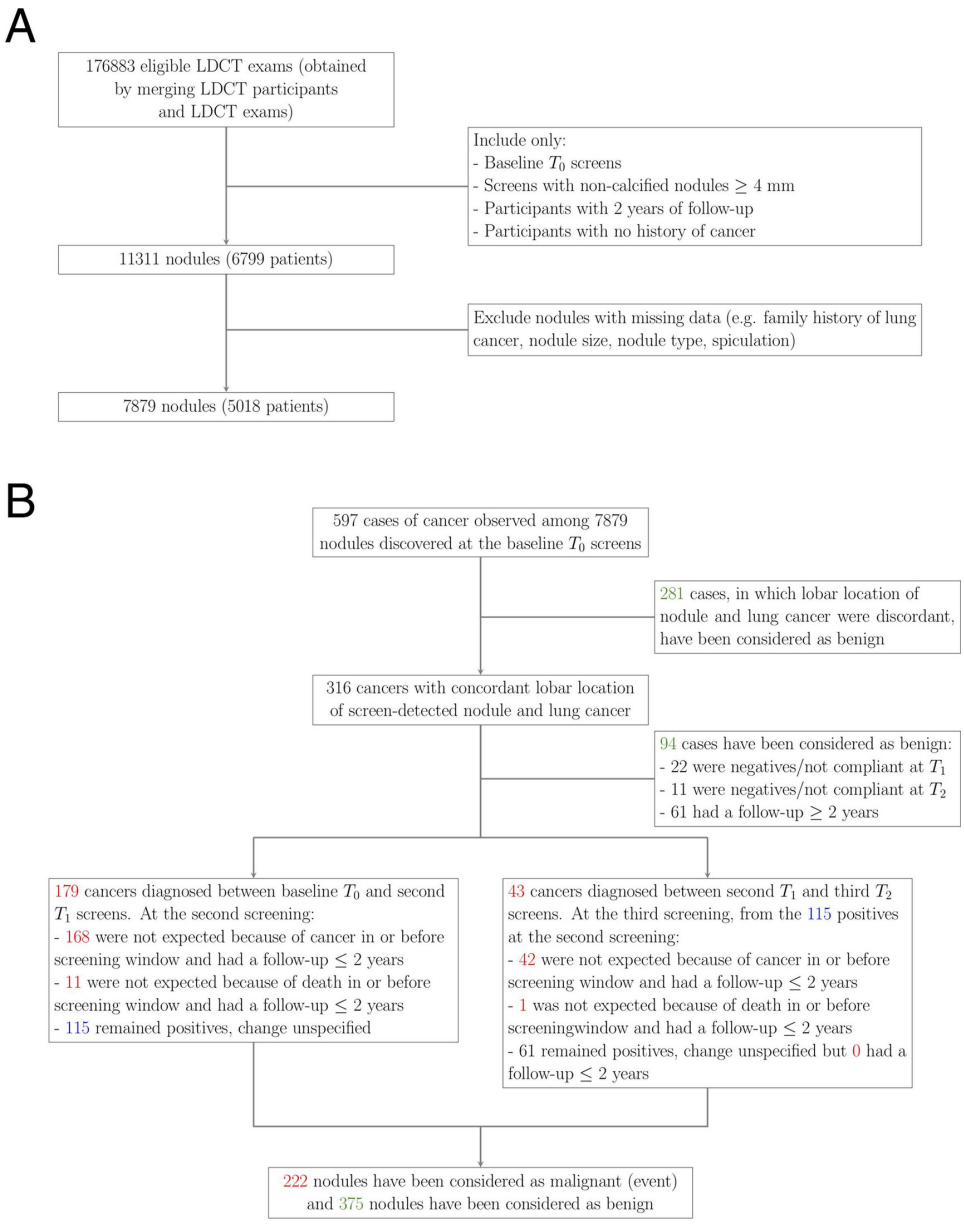

Figure 1 (A) Flow diagram and (B) event covariate construction (details of how data elements were identified are provided in the online supplementary table S1, and in NLST article).[1] LDCT, low-dose CT.

Table 3 Distribution of nodule and participant covariates according to cancer status and comparison in the NLST data set.[1] Mean and SD are reported for quantitative covariates; number and percentage are reported for qualitative covariates. Student's t-test or Wilcoxon-Mann-Whitney test or χ^2 test used when appropriate

Nodule covariates	Malignant nodules (n=222)	Benign nodules (n=7657)	Total (n=7879)	P value
Nodule size (mm)	18.45 (13.89)	6.56 (4.47)	6.89 (5.35)	<0.01
Nodule type				<0.01
Solid	195 (87.84%)	7021 (91.69%)	7216 (91.59%)	
Non-solid	10 (4.5%)	475 (6.2%)	485 (6.16%)	
Part-solid	17 (7.66%)	161 (2.1%)	178 (2.26%)	
Nodule location				<0.01
Right upper lobe	90 (40.54%)	1612 (21.05%)	1702 (21.6%)	
Right middle lobe	11 (4.95%)	1179 (15.4%)	1190 (15.10%)	
Right lower lobe	39 (17.57%)	1925 (25.14%)	1964 (24.93%)	
Left upper lobe	56 (25.23%)	952 (12.43%)	1008 (12.79%)	
Lingula	0 (0%)	339 (4.43%)	339 (4.30%)	
Left lower lobe	26 (11.71%)	1604 (20.95%)	1630 (20.69%)	
Other	0 (0%)	46 (0.6%)	46 (0.58%)	
Nodule upper location	146 (65.77%)	2564 (33.49%)	2710 (34.40%)	<0.01
Nodule count at baseline	1.87 (1.1)	2.45 (1.64)	2.44 (1.63)	<0.01
Spiculation	159 (71.62%)	873 (11.4%)	1032 (13.1%)	<0.01
Brock score	0.32 (0.25)	0.03 (0.08)	0.04 (0.10)	<0.01
Participants covariates	**Cancer (n=194)**	**Benign (n=4824)**	**Total (n=5018)**	**P value**
Age (years)	63.74 (5.35)	61.94 (5.10)	62.01 (5.12)	<0.01
Gender				0.99
Female	75 (38.66%)	1872 (38.81%)	1947 (38.80%)	
Male	119 (61.34%)	2952 (61.19%)	3071 (61.20%)	
Family history of lung cancer	58 (29.90%)	1096 (22.72%)	1154 (23.0%)	0.02
Emphysema	83 (42.78%)	1578 (32.71%)	1661 (33.10%)	<0.01

NLST, National Lung Screening Trial.

the coefficients derived from the original model were too large, which results in overestimation of the probability of lung cancer for participants at high risk or underestimation of the probability for participants at low risk of lung cancer. Conversely, if the calibration slope is bigger than 1, then the original regression coefficients were too low, and the calculated probabilities of lung cancer will be falsely low.

Model recalibration, revision and extension

When a lack of calibration was observed, we updated the model using several methods described in table 2 based on recommendations outlined.[14 30]

These recommendations are categorised into four main methods: no model adjustment (method 1); recalibration methods (methods 2 and 3); revision methods (methods 4 and 5); and extension methods, considering new predictors not included in the original model (methods 6–8).

After updating the model using methods 2–8, we calculated three measures of performance: (1) the Brier score as a measure of overall performance; (2) the concordance statistic (c-statistic) as a measure of discrimination; and (3) the Akaike information criterion (AIC) as a measure of relative goodness of fit. The Brier score measures the accuracy (mean squared difference), between predicted probabilities and actual outcomes, where the lower the Brier score, the better the fit; a perfect score is 0. The Brier score provides a measure of the absolute performance of the model. The c-statistic is equivalent to the area under the ROC curve, which varies between 0 and 1. The more closely the c-statistic approximates 1, the better the model distinguishes between those with and without lung cancer. The AIC estimates the relative goodness of fit, where the preferred model has the lowest AIC value; as such it cannot provide information on the absolute goodness of the fit of a given model.

Of note, regression coefficients' variances estimated through logistic regression models (external validation and recalibration) were adjusted with the use of the Huber-White[34] robust variance estimator in order to follow what was done by McWilliams *et al*.[2]

Model performance for follow-up intervals of 3 and 4 years

As the follow-up in the derivation cohort used by McWilliams *et al* to train the Brock model ranged from 2.1 to 4.3 years,[2] we performed additional analysis examining discrimination and calibration when predicting malignancy at 3 and 4 years in the NLST population. We created two new subsets of the data, one for each follow-up duration: only patients who were followed for this period were included. Lung cancer diagnoses that occurred outside of this period were not considered.

All analyses were performed using R software, V.3.3.0 (R Development Core Team, A Language and Environment for Statistical Computing, Vienna, Austria, 2016. URL: https://www.R-project.org/).

RESULTS

Data

A flow diagram of the participants selected for the study is presented in figure 1. A detailed flow diagram of how the events were defined is provided in figure 1. The mapping between covariates and the NLST data elements is shown in online supplementary table S1.

Patient characteristics

Characteristics of participants and nodules, with a comparison by group of diagnosis (ie, cancer or no cancer) used in this study, are described in table 3 (and online supplementary table S2, for new covariates).

PI for Brock model

We initially calculated the Brock model[2] (online supplementary figure S1) as follows:

$$\frac{e^{Brock-PI}}{1+e^{Brock-PI}},$$

With: Brock-PI = −6.7892+0.0287*(Age −62)
+0.6011*(1 if female, 0 otherwise)
+0.2961*(1 if family history of lung cancer, 0 otherwise)
+0.2953*(1 if emphysema, 0 otherwise)
+0.377*(1 if nodule type is part-solid, 0 otherwise)
−0.1276*(1 if nodule type is non-solid, 0 otherwise)
+0.6581*(1 if nodule is in upper lung, 0 otherwise)
−0.0824*(Nodule count −4)
+0.7729*(1 if spiculated, 0 otherwise)

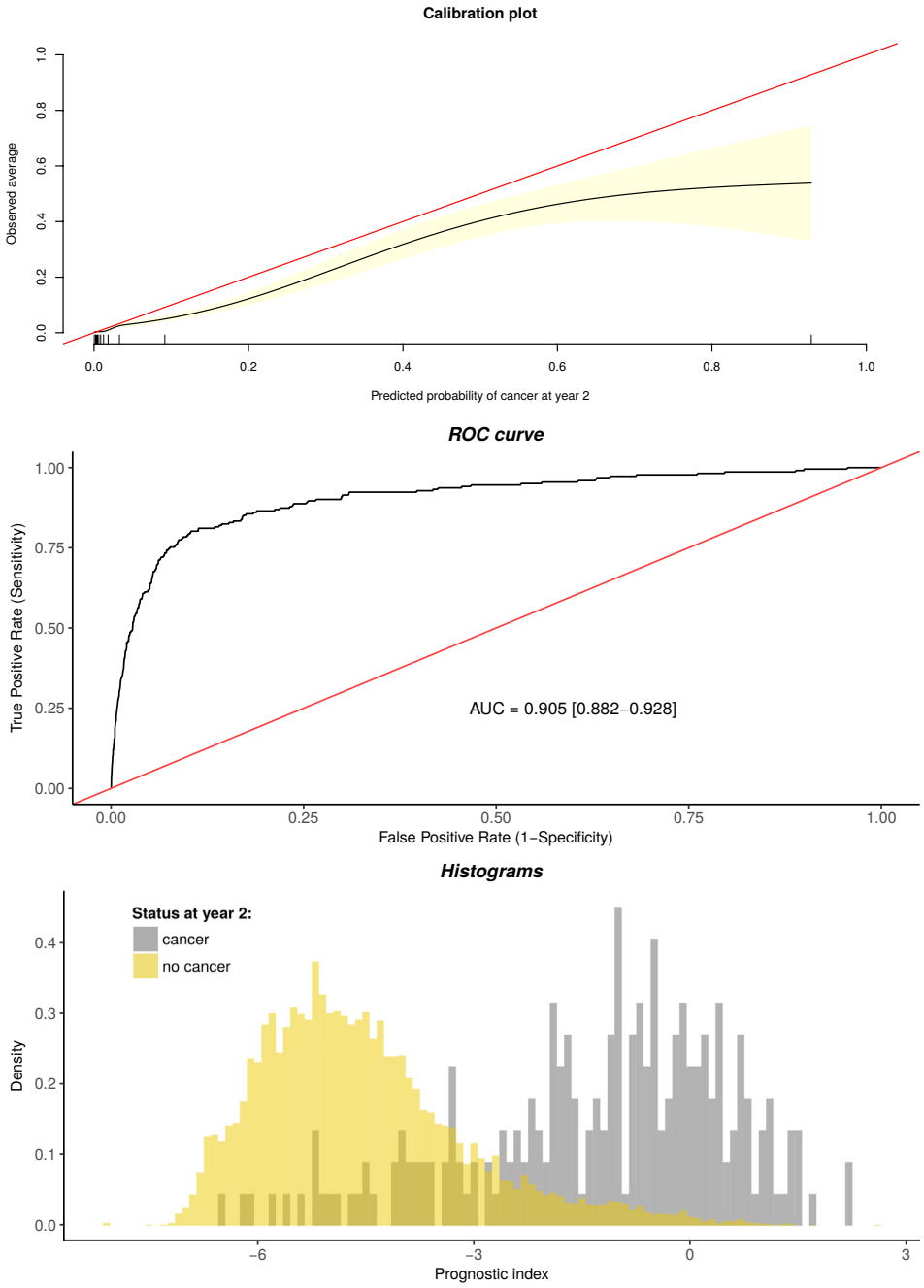

Figure 2 Visual assessment of calibration and discrimination of the Brock model in the National Lung Screening Trial (NLST) data set. AUC, area under the curve; ROC, receiver operating characteristic.

$$-5.3854 * \left(\left\{ \frac{nodule\ size}{10} \right\}^{-0.5} + 1.58113883 \right).$$

Brock model achieves good AUC but has low positive predictive value

The ROC curve showed a good discrimination with an AUC=0.905 (95% CI 0.882 to 0.928) (figure 2). The optimal combination of model SE and SPC (online supplementary figure S2) was at a cut-off probability of 6.5%.[35] However, histograms in figure 2 show that the distribution of the PI, although shifted slightly to the right in the lung cancer group, is insufficient to provide strong discrimination between cases with and without lung cancer. In our study, only 2.8% of the nodules were malignant, which is not accounted for in the AUC. As such, we also examined metrics such as positive predictive value (PPV) and negative predictive value (NPV), which are influenced by disease prevalence. The original model in the NLST data set with an optimal cut-off of 6.5% achieved an SE of 80.18% (74.94–85.42) and an SPC of 89.63% (88.95–90.31) with an excellent NPV of 99.36% (99.2–99.53) but low PPV of 18.31% (16.92–19.7). Finally, the calibration plot showed significant overestimation of lung cancer with the greatest overestimation occurring at higher levels of risk (figure 2).

Covariates were different between NLST and PanCan

When we compared participants' and nodules' characteristics between PanCan and NLST data sets, we found that

the distributions of all covariates were significantly different (table 4), reflecting basic differences between cohorts that would compromise the fit of the model to the NLST data set. The biggest differences were observed for emphysema and spiculation (p<0.001) with medium and small ES, respectively. A large ES was noted for nodule count at the baseline covariate.

Brock model overestimates the probability of cancer in NLST data set

After fitting the original logistic regression model, we found that a likelihood ratio test showed that a significant difference between the model intercept and slope and $\hat{\alpha} = 0$ and $\hat{\beta}_{overall} = 1$ (p<0.01; $\hat{\alpha} = -0.59\ [-0.81, -0.36]$ and $\hat{\beta}_{overall} = 0.97\ [0.88, 1.06]$). The slope was very close to 1, but the intercept was too far from 0 (lower than 0, which is consistent with the overestimation observed in the calibration plot), and as a result, model recalibration is needed.

Table 4 Differences in nodule and participant characteristics between the PanCan data set[2] and the NLST data set[1] (p values, Student's t-test or Wilcoxon-Mann-Whitney test or χ^2 test used when appropriate). For each test, an effect size (ES) was also reported. For χ^2 tests, Cramér's V or ϕ_C was calculated (ie, magnitude of the ES; small: $0.1 \leq ES < 0.3$, medium: $0.3 \leq ES < 0.5$ and large: $ES > 0.5$) and for Student's t-test and Wilcoxon-Mann-Whitney tests, r^2 and Cohen's r^2 (ie, magnitude of the ES; small: $0.01 \leq ES < 0.09$, medium: $0.09 \leq ES < 0.25$ and large: $ES > 0.25$), respectively

Nodule covariates	PanCan (n=7008)	NLST (n=7879)	P value	Effect size
Nodule size (mm)	4.1 (3.1)	6.89 (5.35)	<0.001	0.003
Nodule type*			<0.001 †	0.122
Solid	5511 (78.9%)	7216 (91.59%)		
Non-solid	1105 (15.8%)	485 (6.16%)		
Part-solid	303 (4.3%)	178 (2.26%)		
Perifissural	70 (1.0%)	–		
Nodule location (upper lobe)‡	3879 (55.70%)	2710 (34.40%)	<0.001	0.21
Nodule count at baseline	6.2 (4)	2.44 (1.63)	<0.001	0.733
Spiculation	197 (2.8%)	1032 (13.1%)	<0.001	0.186
Participant covariates	**PanCan (n=1871)**	**NLST (n=5018)**	**P value**	**Effect size**
Age (years)	62.4 (5.8)	62.01 (5.12)	<0.001	0.027
Gender			<0.001	0.077
Female	886 (47.4%)	1947 (38.80%)		
Male	985 (52.6%)	3071 (61.20%)		
Family history of lung cancer	597 (32.4%)	1154 (23.0%)	<0.001	0.091
Emphysema	1110 (60.4%)	1661 (33.10%)	<0.001	0.238

*Only for 6989 nodules in PanCan data set.
†Test without perifissural category.
‡Only for 6964 nodules in PanCan data set.
NLST, National Lung Screening Trial; PanCan, Pan-Canadian Early Detection of Lung Cancer Study.

Recalibration and extension of Brock model improves overall performance

Table 5 summarises the results of the various recalibration and extension methods. The c-statistic, Brier score and AIC were very similar across the seven methods.

Based on method 5, we observed that some of the regression coefficients from the NLST database were similar to those of the PanCan database, such as age, family history and nodule location, while others were very different, most notably emphysema, gender and spiculation. Based on the results of method 4, only the coefficients of gender, emphysema and spiculation require revision. This could be explained by differences in participant/nodule characteristics highlighted in table 4. Indeed, significant differences, with small ES, were observed for emphysema (60.4% vs 33.10%, p<0.01, ES=0.238) and spiculation (2.8% vs 13.1%, p<0.01, ES=0.186). Based on methods 6–8, new covariates were added to the model following the methodology described in McWilliams *et al*.[2] Namely, if quantitative covariates had a non-linear relationship with lung cancer, they were transformed using fractional polynomials.[36] However, only pack-years contributed to model performance. The results of method 6 achieved the highest c-statistic, 0.914 (95% CI 0.892 to 0.936), and the lowest AIC (1288) and Brier score (0.021). Emphysema, spiculation and nodule count regression coefficients were updated, and pack-year was added to the model. According to this recalibrated model, the probability of having lung cancer 2 years after the LDCT baseline is $\frac{e^{PI}}{1+e^{PI}}$, with:

PI=−1.88+0.83*(Brock-PI)

- −0.35*(1 if emphysema, 0 otherwise)
- −0.10*(Nodule count −4)
- +0.85*(1 if spiculation, 0 otherwise)
- $+0.38^{*}\left(\frac{pack\ year\ -57.32}{25.01}+1.1\right)$.

A likelihood ratio test that was performed between the recalibrated and original models (online supplementary table S2; p<0.01) suggests a significant improvement in model performance after recalibration. Notably, the AUC reported for the Brock model[2] was 0.942 [95% CI 0.909 to 0.967]; the overlap of CIs between the original and our revised models suggests that they are very close in performance. When we visually compared the calibration and discrimination of the original model with our recalibrated model (figure 3), calibration clearly improved, although discrimination did not change (0.914 [95% CI 0.892 to 0.936] vs 0.905 [95% CI 0.882 to 0.928], p=0.59 based on Delong test[37]). With an optimal cut-off of 3.4%, the recalibrated model achieved an SE of 82.43% [77.43–87.44] and an SPC of 88.49% [87.78–89.21]. These results are associated with an NPV and a PPV of 99.43% [99.26–99.59] and 17.25% [16.01–18.49], respectively.

Moreover, as internal validation, we performed 500 bootstrap replications for the model generated following method 6. The optimism-corrected performance was 0.912 (95% CI 0.891 to 0.932).[11]

Discrimination and calibration performance declines when predicting cases with longer follow-up

For the 3-year follow-up model, we retained 7769 nodules where 255 were lung cancer cases. For the 4-year follow-up, there were 7660 nodules where 260 were lung cancer cases. Figure 4 shows the visual performance assessment. Calibration seemed to improve with the increase of the follow-up (and hence the increase of lung cancer events). However, the greater the years of follow-up, the greater the drop-in discrimination. We also assessed external validation on the data sets

Table 5 Results of the updating methods[14 30] for the Brock model.[2] Variances of logistic regression models' coefficients were adjusted with the use of the Huber-White robust variance estimator.[34] The reference groups are the same as in McWilliams *et al*.[2] For each model, the c-statistic, the Brier score and the Akaike information criterion (AIC) were calculated

Updating methods		Discrimination (c-statistic)	Overall performance (Brier score)	Relative goodness of fit (AIC)
Method 1: No adjustment		0.905 (0.882, 0.928)	0.023	1385
Method 2: Recalibration		0.905 (0.882, 0.928)	0.022	1333
α	−0.54 (−0.71, 0.36)			
Method 3: Recalibration		0.905 (0.882, 0.928)	0.022	1335
α	−0.59 (−0.81, −0.36)			
$\beta_{overall}$	0.97 (0.88, 1.06)			
Method 4: Revision		0.907 (0.885, 0.930)	0.021	1317
α	−1.0 (−1.48, −0.52)			
$\beta_{overall}$	0.85 (0.75, 0.96)			
γ_{age}	–			
γ_{gender}	−0.34 (−0.69, 0.01)			
$\gamma_{family\ history}$	–			
$\gamma_{emphysema}$	−0.28 (−0.65, 0.09)			
$\gamma_{nodule\ size}$	–			
$\gamma_{nodule\ type}$	–			
$\gamma_{nodule\ upper}$	–			
$\gamma_{baseline\ nodule\ number}$	–			
$\gamma_{spiculation}$	0.81 (0.40, 1.21)			
Method 5: Revision		0.901 (0.879–0.924)	0.022	1324
α	−6.92 (−7.5, −6.35)			
β_{age}	0.04 (0, 0.07)			
β_{gender}	0.19 (−0.17, 0.55)			
$\beta_{family\ history}$	0.16 (−0.24, 0.55)			
$\beta_{emphysema}$	−0.06 (−0.42, 0.31)			
$\beta_{nodule\ size}$	−4.50 (−5.12, −3.87)			
$\beta_{nodule\ type:\ non-solid}$	−0.19 (−0.93, 0.54)			
$\beta_{nodule\ type:\ part-solid}$	0.01 (−0.61, 0.64)			
$\beta_{nodule\ upper}$	0.70 (0.33, 1.07)			
$\beta_{baseline\ nodule\ number}$	−0.15 (−0.28, −0.02)			
$\beta_{spiculation}$	1.46 (1.08, 1.83)			
Method 6: Extension		0.914 (0.892, 0.936)	0.021	1288
α	−1.88 (−2.48, −1.29)			
$\beta_{overall}$	0.83 (0.72, 0.94)			
γ_{age}	–			
γ_{gender}	–			
$\gamma_{family\ history}$	–			
$\gamma_{emphysema}$	−0.35 (−0.72, 0.03)			
$\gamma_{nodule\ size}$	–			
$\gamma_{nodule\ type}$	–			
$\gamma_{nodule\ upper}$	–			
$\gamma_{baseline\ nodule\ number}$	−0.10 (−0.23, 0.04)			
$\gamma_{spiculation}$	0.85 (0.45, 1.25)			
β_{BMI}	–			
$\beta_{pack-year}$	0.38 (0.22, 0.54)			

Continued

Table 5 Continued

Updating methods		Discrimination (c-statistic)	Overall performance (Brier score)	Relative goodness of fit (AIC)
$\beta_{\text{asbestos work}}$	–			
$\beta_{\text{smoking status}}$	–			
Method 7: Extension		0.910 (0.888, 0.931)	0.022	1383
α	−7.98 (−10.27, −5.68)			
β_{age}	0.03 (0, 0.07)			
β_{gender}	0.34 (−0.04, 0.71)			
$\beta_{\text{family history}}$	0.15 (−0.25, 0.54)			
$\beta_{\text{emphysema}}$	−0.10 (−0.46, 0.27)			
$\beta_{\text{nodule size}}$	0.10 (0.06, 0.13)			
$\beta_{\text{nodule type: non-solid}}$	−0.14 (−0.85, 0.58)			
$\beta_{\text{nodule type: part-solid}}$	0.20 (−0.47, 0.87)			
$\beta_{\text{nodule upper}}$	0.75 (0.37, 1.12)			
$\beta_{\text{baseline nodule number}}$	−0.18 (−0.31, −0.04)			
$\beta_{\text{spiculation}}$	2.07 (1.65, 2.48)			
β_{BMI}	–			
$\beta_{\text{pack-year}}$	0.35 (0.20, 0.50)			
$\beta_{\text{asbestos work}}$	–			
$\beta_{\text{smoking status}}$	–			
Method 8: Extension		0.909 (0.887, 0.931)	0.022	1388
α	−7.99 (−10.41, −5.56)			
β_{age}	0.03 (0, 0.07)			
β_{gender}	0.32 (−0.06, 0.69)			
$\beta_{\text{family history}}$	0.14 (−0.25, 0.54)			
$\beta_{\text{emphysema}}$	−0.10 (−0.48, 0.28)			
$\beta_{\text{nodule size}}$	0.10 (0.06, 0.13)			
$\beta_{\text{nodule type: non-solid}}$	−0.13 (−0.85, 0.59)			
$\beta_{\text{nodule type: part-solid}}$	0.23 (−0.43, 0.89)			
$\beta_{\text{nodule upper}}$	0.75 (0.38, 1.12)			
$\beta_{\text{baseline nodule number}}$	−0.18 (−0.31, −0.04)			
$\beta_{\text{spiculation}}$	2.07 (1.65, 2.49)			
β_{BMI}	0.001 (−0.21, 0.21)			
$\beta_{\text{pack-year}}$	0.36 (0.20, 0.51)			
$\beta_{\text{asbestos work}}$	0.03 (−0.33, 0.40)			
$\beta_{\text{smoking status}}$	−0.33 (−1.32, 0.66)			

In methods 6–8, we used the data set presented in online supplementary table S2. The references for those covariates are always yes versus no. As in McWilliams *et al*,[2] age was centred on 62 years, nodule size was centred on 4 mm and nodule count was calculated as $(\{\text{nodule size}/10\}^{-0.5} - 1.58113883)$.
Pack-year and body mass index (BMI) covariates were normalised. They had a non-linear relationship with lung cancer and were transformed using a fractional polynomial.[36] After transformation, BMI was ((BMI−27.64)/4.8)+3 and pack-year was ((pack-year−57.32)/25.01)+1.1.

of extended follow-ups of 3 and 4 years. Likelihood tests were significant at 5% for both follow-up periods ($p<0.01$; $\hat{\alpha} = -0.51\ [-0.73, -0.29]$ and $\hat{\beta}_{overall} = 0.90\ [0.82, 0.98]$, $\hat{\alpha} = -0.46\ [-0.69, -0.24]$ and $\hat{\beta}_{overall} = 0.90\ [0.82, 0.98]$ for 3 and 4 years, respectively).

DISCUSSION

The aim of this work was to assess the external validity of a diagnostic prediction model issued from McWilliams *et al*[2] in the NLST database and to assess how various recalibration methods influenced model performance. We assessed both model discrimination and calibration. Visually, despite a good AUC (0.91, 95% CI 0.89 to 0.93), a lack of calibration and discrimination has been noted. This lack was confirmed by the estimation of the intercept and the slope on the PI in the original data set, which highlighted a need of recalibration due to high predicted probabilities in the original model. Emphysema, nodule count at baseline and nodule spiculation had a different effect in the

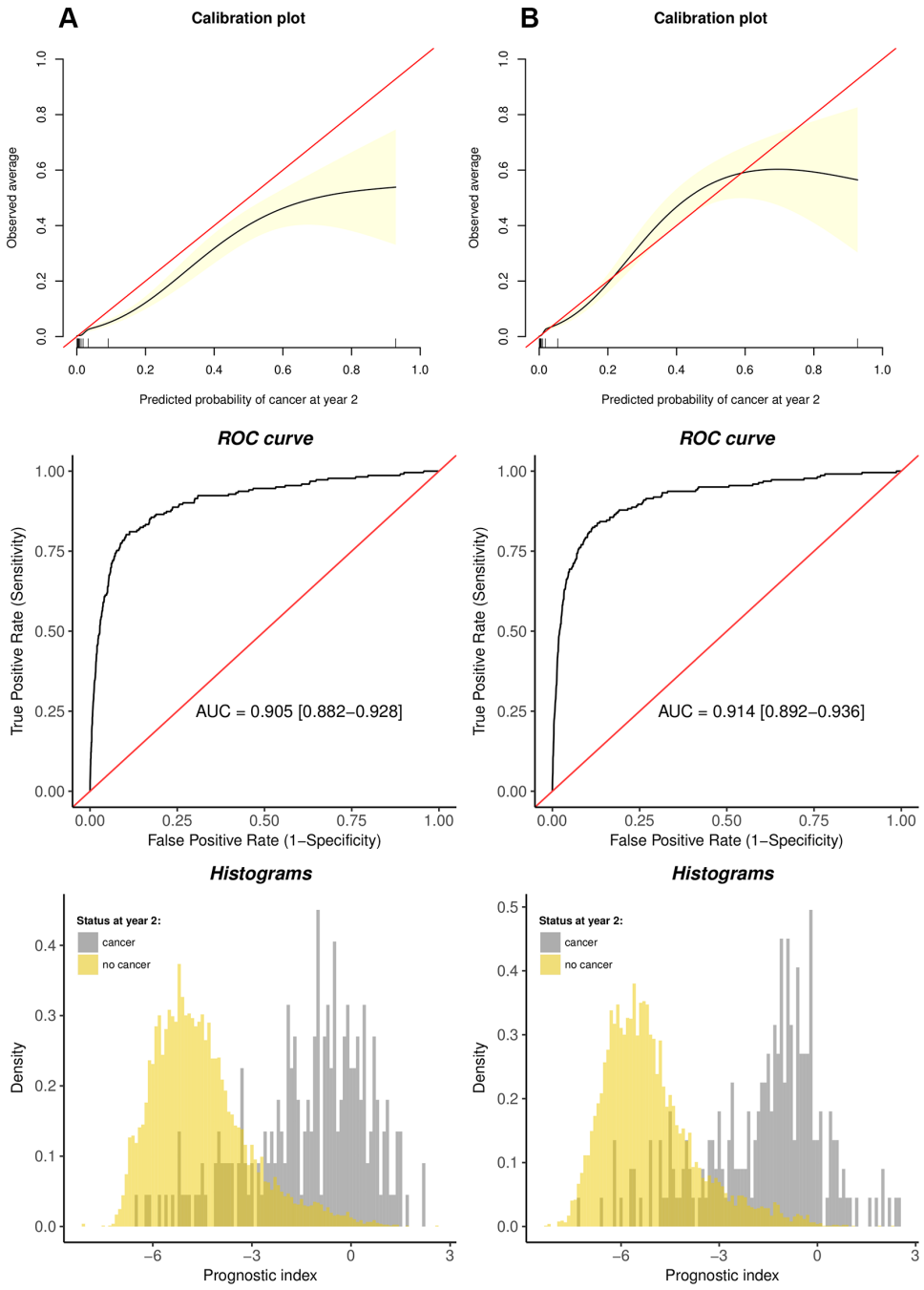

Figure 3 Visual discrimination and calibration of the original model (A) compared with the model recalibrated through method 6 (B). AUC, area under the curve; ROC, receiver operating characteristic.

NLST cohort compared with the PanCan cohort. Among the new covariates, only the pack-year history was added. Finally, by using recalibration methods, we proposed a new prediction score based on recalibration of the Brock model.

Three groups applied the McWilliams model to the NLST data set; while some applied it on external data sets.[17 19 22] External validation of a prediction model requires the exploration of both discrimination and calibration. As noted previously, AUC is insufficient to fully assess the discrimination of a prediction model. This measure can only provide a general overview of the apparent discrimination but does not constitute an external validation by itself. Moreover, the assessment of calibration is equally important to model performance when applying an existing prediction model to an external data set. Of note, the Hosmer-Lemeshow test is not recommended to assess calibration because of its limited power and interpretability.[38–40] To make an external validation properly, the recommended method is to test the fit of the original model on an independent external data set by estimating the intercept and the slope on the PI.[11 14 30] For this purpose, the external validation data set has to be 'comparable' to the original data set in terms of exclusion and inclusion criteria and follow-up duration. This last point is challenging to follow because, as noted, the NLST database was patient based, while the PanCan was nodule based. In their article, Nair and colleagues[17] considered only the largest nodule observed on baseline screening LDCT to be malignant and excluded all patients with lung cancer diagnoses made after the T1 screen (first incidence screen). Schreuder *et al*[19] created their own

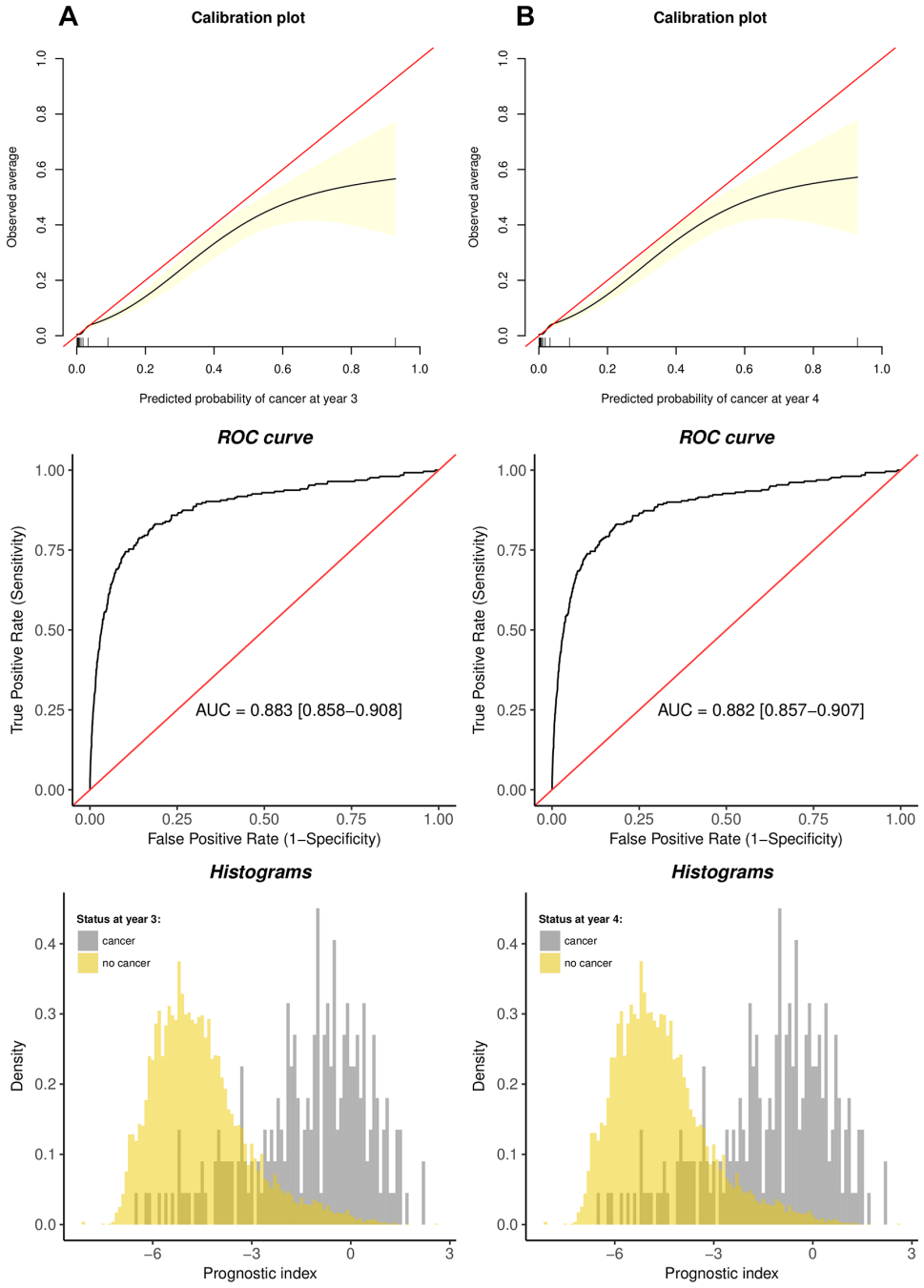

Figure 4 Visual assessment of the Brock model for 3 years (A) and 4 years (B) of follow-up. AUC, area under the curve; ROC, receiver operating characteristic.

prediction model using the NLST data by aggregating all nodule characteristics at the participant level. Their outcome was lung cancer diagnosis during the year following the T1 screen. White *et al*[22] excluded from their analysis all participants with multiple nodules only in the cancer group and used all the follow-up data available in the NLST data set. From our perspective, the exclusion of all patients with cancer with multiple observed nodules or all participants diagnosed after the T1 screen, or inclusion of only one occurrence per participant, led to substantial selection bias. Moreover, it is not clear that they considered censorship given the use of a logistic regression model. Furthermore, exclusion and inclusion criteria were not explicitly described, which means that the model may not have been applied exactly in the same way as in the derivation data set. Finally, recalibration was not considered in any of the prior validation studies.

Some limitations of our study should be noted. To apply the Brock score, we assumed that the diagnostic score was derived from patients followed for 2 years (and only 2 years) after their baseline screen. Logistic regression models require that all participants be followed for the entire time period,[11] which has to be the same for all participants and thus do not account for censorship. Indeed, survival models are dependent on time (ie, a Cox model, eg, $\lambda\left(t, X_1 \ldots X_n\right) = \lambda_0\left(t\right) \exp\left(\sum_{i=1}^{n} \beta_i X_i\right)$) and analyse time to event outcomes; in contrast, logistic regression models address

only binary responses. Moreover, survival models explain more variability in the data than logistic regression models,[41] which is why logistic regression only focuses on short follow-ups while survival models are preferred for longer follow-ups.[11] The task of predicting probabilities of cancer at 2, 3 or 4 years is not the same statistically. As such, care should be given when attempting to estimate the probability of cancer beyond a 2-year follow-up period. Moreover, while the probability of malignancy may increase over a longer follow-up, the type of cancer identified over longer periods may be biologically distinct (more indolent).

Moreover, as NLST is a patient-based data set, we were not able to qualify the status of each nodule. We made some assumptions based on nodule locations and ultimate lung cancer locations to construct our 'event covariate'. Based on those assumptions, some nodules were classified as not malignant, but we cannot access the veracity of those assumptions, and this may have contributed to selection bias. Similarly, 3432 (30%) of LDCT screenings were excluded due to missing data, which may also have contributed to selection bias. In PanCan data, LDCT screenings were classified as positive for nodules ≥1 mm which was not the case in NLST (≥4 mm). This could explain the differences noticed between the derivation and the validation data sets. Differences between NLST and PanCan data sets have been highlighted (table 4); nevertheless, we temper this result, which is partly due to ES. Finally, given that this was an exercise in external validation, we applied the prediction model in the exact same way as in the NLST data set. As such, we did not verify original assumptions made on the covariates included in the original logistic regression model, such as the linearity of a covariate with the outcome, which may have impaired the interpretation of the model.

At the patient level in the target population, the original score avoids 165 unnecessary screenings for one missing cancer diagnosis (eg, true negative/false negative rate). This number reaches a value of 176 in the recalibrated model. In both models, the PPV is low and could lead, in clinical practice, to a lot of false positive cases. Nevertheless, the NPV is very good and limits the number of false negative cases, which could help prevent unnecessary screening. Of note, thresholds are useful to clinical decision-making but should be used with precaution and interpreted in their context according to their own population. Indeed, they are not universal. For example, using the Brock model, van Riel *et al*,[42] in a Danish population, used two thresholds to define three risk groups of increasing risk (<6%: low risk, 6%–30%: intermediate risk and ≥30%: high risk), whereas the British Thoracic Society[28] recommended a positron emission tomography-CT scan for patients with an estimated probability greater than 10%. Thresholds given in this article are examples but can be increased or decreased, either to limit the number of false negatives or the number of false positives and thus improve PPV or NPV, respectively, according to the needs and context of each. Examples of thresholds and their impact on SE, SPC, NPV and PPV are given in online supplementary table S3.

As mentioned in ref 43, there are two types of models, validated statistically and clinically. Toll *et al*[44] distinguished three phases in multivariable prediction research: (1) development of a prediction score; (2) external validation of this score; and (3) the study of the clinical impact. In this article, we outlined the importance of the statistical external validation of prediction models. This procedure assesses the validity of an existing prediction score in an independent external data set. Our study focused on a geographical external validation.[45] As such, the Brock model was applied in the exact same way in the validation data set as in the derivation data set. More than a statistical exercise, we seek to provide insights on how to interpret the results of external validation, providing readers with practical guidelines on how to evaluate and adopt published prediction models in clinical decision-making. As the number of prediction models increases, the importance of following reporting guidelines such as TRIPOD cannot be understated as the model design and reported results become increasingly more complex and difficult to reproduce and interpret. Of note, our intent was not to promote a new prediction model over the current Brock model, but rather, the effect of following recalibration and updating techniques on model performances in the NLST data set. As with any revisions to a prediction model, further validation is required before use. In conclusion, while the Brock model achieved a high AUC when validated on NLST, overall model performance benefited from both updating and recalibration without difference in the model discrimination, which cannot be altered during the recalibration process.[12] Recalibrating an existing model with data from the target population can improve its transportability to other individuals.[45] We found that the process of external validation, hence the generalisability of a prediction model, requires thoughtful accounting of both discrimination and calibration and the consideration of recalibration methods.

Correction notice This article has been corrected since it was published. The third sentence in the Introduction section was re-worded.

Acknowledgements The authors thank the National Cancer Institute (NCI) for access to the NCI's data collected by the National Lung Screening Trial as part of an existing data transfer agreement (NLST-93). The authors also thank Panayiotis Petousis and Lew Andrada for their comments.

Contributors Substantial contributions to the conception or design of the work, or the acquisition, analysis or interpretation of data: AW, DRA, WH. Drafting the work or revising it critically for important intellectual content: AW, DRA, WH. Final approval of the version published: AW, DRA, WH.

Funding Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under award number R01CA210360 and by the Department of Radiological Sciences under the Integrated Diagnostics Program.

Disclaimer The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement These data are restricted and cannot be publicly available, but permission access can be requested through this website (https://biometry.nci.nih.gov/cdas/).

REFERENCES

1 Aberle DR, Adams AM, Berg CD, *et al*. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med* 2011;365:395–409.
2 McWilliams A, Tammemagi MC, Mayo JR, *et al*. Probability of cancer in pulmonary nodules detected on first screening CT. *N Engl J Med* 2013;369:910–9.
3 Swensen SJ, Silverstein MD, Ilstrup DM, *et al*. The probability of malignancy in solitary pulmonary nodules. Application to small radiologically indeterminate nodules. *Arch Intern Med* 1997;157:849–55.
4 Gould MK, Ananth L, Barnett PG, *et al*. A clinical model to estimate the pretest probability of lung cancer in patients with solitary pulmonary nodules. *Chest* 2007;131:383–8.
5 Yonemori K, Tateishi U, Uno H, *et al*. Development and validation of diagnostic prediction model for solitary pulmonary nodules. *Respirology* 2007;12:856–62.
6 Gurney JW. Determining the likelihood of malignancy in solitary pulmonary nodules with Bayesian analysis. Part I. Theory. *Radiology* 1993;186:405–13.
7 Cummings SR, Lillington GA, Richard RJ. Estimating the probability of malignancy in solitary pulmonary nodules. A Bayesian approach. *Am Rev Respir Dis* 1986;134:449–52.
8 Deppen SA, Blume JD, Aldrich MC, *et al*. Predicting lung cancer prior to surgical resection in patients with lung nodules. *J Thorac Oncol* 2014;9:1477–84.

9 Soardi GA, Perandini S, Larici AR, *et al*. Multicentre external validation of the BimC model for solid solitary pulmonary nodule malignancy prediction. *Eur Radiol* 2017;27:1929–33.
10 Li Y, Wang J. A mathematical model for predicting malignancy of solitary pulmonary nodules. *World J Surg* 2012;36:830–5.
11 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–W73.
12 Royston P, Altman DG. External validation of a COX prognostic model: principles and methods. *BMC Med Res Methodol* 2013;13.
13 Royston P, Altman DG. Visualizing and assessing discrimination in the logistic regression model. *Stat Med* 2010;29:2508–20.
14 Steyerberg EW, Borsboom GJJM, van Houwelingen HC, *et al*. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med* 2004;23:2567–86.
15 Zhao H, Marshall HM, Yang IA, *et al*. Screen-detected subsolid pulmonary nodules: long-term follow-up and application of the PanCan lung cancer risk prediction model. *Br J Radiol* 2016;89.
16 Al-Ameri A, Malhotra P, Thygesen H, *et al*. Risk of malignancy in pulmonary nodules: a validation study of four prediction models. *Lung Cancer* 2015;89:27–30.
17 Nair VS, Sundaram V, Desai M, *et al*. Accuracy of models to identify lung nodule cancer risk in the National Lung Screening trial. *Am J Respir Crit Care Med* 2018;197:1220–3.
18 She Y, Zhao L, Dai C, *et al*. Development and validation of a nomogram to estimate the pretest probability of cancer in Chinese patients with solid solitary pulmonary nodules: a multi-institutional study. *J Surg Oncol* 2017;116:756–62.
19 Schreuder A, Schaefer-Prokop CM, Scholten ET, *et al*. Lung cancer risk to personalise annual and biennial follow-up computed tomography screening. *Thorax* 2018;73:626–33.
20 Winkler Wille MM, van Riel SJ, Saghir Z, *et al*. Predictive accuracy of the PanCan lung cancer risk prediction model -external validation based on CT from the Danish lung cancer screening trial. *Eur Radiol* 2015;25:3093–9.
21 Talwar A, Rahman NM, Kadir T, *et al*. A retrospective validation study of three models to estimate the probability of malignancy in patients with small pulmonary nodules from a tertiary oncology follow-up centre. *Clin Radiol* 2017;72:177.e1–177.e8.
22 White CS, Dharaiya E, Campbell E, *et al*. The Vancouver lung cancer risk prediction model: assessment by using a subset of the National Lung Screening trial cohort. *Radiology* 2017;283:264–72.
23 Chung K, Mets OM, Gerke PK, *et al*. Brock malignancy risk calculator for pulmonary nodules: validation outside a lung cancer screening population. *Thorax* 2018;73:857–63.
24 Yang B, Jhun BW, Shin SH, *et al*. Comparison of four models predicting the malignancy of pulmonary nodules: a single-center study of Korean adults. *PLoS One* 2018;13:e0201242.
25 Perandini S, Soardi GA, Larici AR, *et al*. Multicenter external validation of two malignancy risk prediction models in patients undergoing 18F-FDG-PET for solitary pulmonary nodule evaluation. *Eur Radiol* 2017;27:2042–6.
26 Schultz EM, Sanders GD, Trotter PR, *et al*. Validation of two models to estimate the probability of malignancy in patients with solitary pulmonary nodules. *Thorax* 2008;63:335–41.
27 Herder GJ, van Tinteren H, Golding RP, *et al*. Clinical prediction model to characterize pulmonary nodules. *Chest* 2005;128:2490–6.
28 Callister MEJ, Baldwin DR, Akram AR, *et al*. British Thoracic Society guidelines for the investigation and management of pulmonary nodules. *Thorax* 2015;70(Suppl 2):ii1–54.
29 Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study: sample size considerations for validating a prognostic model. *Statistics in Medicine* 2016;35:214–26.
30 Janssen KJM, Moons KGM, Kalkman CJ, *et al*. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol* 2008;61:76–86.
31 Fritz CO, Morris PE, Richler JJ. Effect size estimates: current use, calculations, and interpretation. *J Exp Psychol Gen* 2012;141:2–18.
32 Tomczak M, Tomczak E. The need to report effect size estimates revisited. *An overview of some recommended measures of effect size* 2014;1:19–25.
33 Cohen J. In: Hillsdale NJ, ErlbaumAssociates L, eds. *Statistical power analysis for the behavioral sciences*. 2nd, 1988.
34 White H. A Heteroskedasticity-Consistent covariance matrix estimator and a direct test for Heteroskedasticity. *Econometrica* 1980;48:817–38.
35 Youden WJ. Index for rating diagnostic tests. *Cancer* 1950;3:32–5.
36 Sauerbrei W, Meier-Hirmer C, Benner A, *et al*. Multivariable regression model building by using fractional polynomials: description of SAS, STATA and R programs. *Computational Statistics Data Analysis* 2006;50:3464–85.
37 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44:837–45.
38 Harrell FE. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer Series in Statistics, 2013.
39 Peek N, Arts DGT, Bosman RJ, *et al*. External validation of prognostic models for critically ill patients required substantial sample sizes. *J Clin Epidemiol* 2007;60:491.e1–491.e13.
40 Steyerberg EW, Vickers AJ, Cook NR, *et al*. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010;21:128–38.
41 Green MS, Symons MJ. A comparison of the logistic risk function and the proportional hazards model in prospective epidemiologic studies. *J Chronic Dis* 1983;36:715–23.
42 van Riel SJ, Ciompi F, Winkler Wille MM, *et al*. Malignancy risk estimation of pulmonary nodules in screening CTS: comparison between a computer model and Human observers. *PLoS One* 2017;12:e0185032.
43 Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000;19:453–73.
44 Toll DB, Janssen KJM, Vergouwe Y, *et al*. Validation, updating and impact of clinical prediction rules: a review. *J Clin Epidemiol* 2008;61:1085–94.
45 Moons KGM, Kengne AP, Grobbee DE, *et al*. Risk prediction models: II. external validation, model updating, and impact assessment. *Heart* 2012;98:691–8.