APPENDIX

Data Preparation

The NLST data set provided by the National Cancer Institute was in .csv format. Two separate files were used: one containing patient characteristics (prsn.csv) and the other scan abnormalities (sctabn.csv). The former was organized with each row representing a unique participant ID (PID); in the latter, each row represented a unique nodule or other scan feature from one scan. This meant that many participants were represented by multiple rows, whereas some PIDs were not present in the data set because no abnormalities were found in these participants throughout the three screening rounds. In order to merge the two data sets in R, the scan abnormalities data set was modified.

A new .csv file was made (sctabn_modified.csv) containing the scan outcome data where each row represented a unique PID: each variable was transformed into three new variables, one per screening round; cells were left blank in the screening rounds where no abnormalities were found. Note that PIDs in which no abnormalities were found in any of the scans were not included in the data set; when merging (explained below), these variables had a default null value.

As mentioned above, many participants had multiple abnormalities in one screening round; to be able to implement this into the new data set, only the length of the longest nodule diameter (SCT_LONG_DIA) and perpendicular diameter (SCT_PERP_DIA) of the nodule with the greatest SCT_LONG_DIA value per PID per screening round was copied to the new data set. In the case of a tie, the longest SCT_PERP_DIA was taken. Some new variables were created based on variables from sctabn.csv, as shown in Appendix Table 1. Specifically, presence of emphysema (sct_emphy), a nonsolid nodule (sct_anynonsolid), a partsolid nodule (sct_anypartsolid), a spiculated nodule (sct_anyspic), or a nodule located in the upper lobe (sct_anyupperlobe) were given as binary variables. Additionally, the total number of nodules minus one was given as a continuous integer variable (sct_extranodules). The NLST data set dictionaries can be downloaded from https://biometry.nci.nih.gov/cdas/datasets/nlst/. Note that these new variables

represented the presence of features in any of the multiple nodules detected (if applicable) in the entire CT scan. For example, if one scan contained a spiculated 6x6mm nodule, a partsolid 8x5mm nodule, and a third 4x3mm nodule located in the upper lobe, all three variables would be considered present for that scan, and the longest diameter and perpendicular diameter used for our models would have been 8mm and 5mm, respectively. Similarly, two new binary variables were created using the provided patient characteristics variables to notate whether lung cancer was diagnosed in a first-degree family members (fam) and whether the participant had been previously diagnosed with any form of cancer (cancbin).

Statistical Analysis

Statistical analysis was performed using the R statistical analysis package version 3.3.2. The latest version available by 31 May 2017 of each package mentioned were used. Multiple logistic regression was performed under guidance from Mangiafico in the "Multiple Logistic Regression" chapter.[1]

The prsn.csv file was imported into the R program first, where a new data frame was created containing only the participants who fit the inclusion criteria ("ELIG" = 2 AND "rndgroup" = 1 AND "scr_res0" < 10 AND "cancyr" ≠ 0 AND "scr_res1" < 11). Some data cleaning was necessary to correctly and consistently label missing and null values as such, as well as defining variable classes (e.g. factor, numeric, integer, string). Next, sctabn_modified.csv was imported and merged with the participant characteristics data frame. This data set was also cleaned.

Out of all variables used in this study, only the variable "cancbin" had 57 (0.2%) missing data points; these were replaced using multiple imputations using the "mice()" function from the "mice" package.

The NLST data set contained over 300 variables, but it would be methodologically incorrect and ultimately impractical to attempt creating models containing all variables. A rule of thumb allows the pre-selection of, at most, one variable per 10 outcome cases (in this case, lung cancer); therefore, only 17 variables could be selected. Six models are reported. For the base parsimonious model (subsequently referred to as the "base model"), a bidirectional stepwise multiple logistic regression analysis was

performed using the “step” function on a set of 17 variables deemed important; variables which had a Pearson's Chi-squared p-value less than 0.20 were included in the base model. The number of variables The beta coefficients were determined using model-fitting function “brglm()” from the “brglm” package. In the case that one or more variables had a p-value of 0.20 or greater after being fitted to a model, the variable with the highest p-value was removed and the model-fitting function was run again; this process was repeated as many times as necessary. For each variable, odds ratios with 95% two-tailed confidence intervals and p-value were calculated using the code “exp(cbind(coef(x), confint(x)))” partially from the “MASS” package, where “x” represents the model. For more information on odds ratios, please refer to McHugh for the basic interpretation and Sperandei for its interpretation in the context of logistic regression analysis.[1,2]

For the parsimonious polynomial model (“polynomial model”), the continuous variables were transformed into polynomials to the second power (“poly” function); subsequently, the same statistical analysis as used for the base model was repeated to create the polynomial model. The parsimonious patient characteristics model (“patient characteristics model”) analysis was performed on 11 patient characteristics variables. The diameter and Patz models were created using only one variable each: “SCT_LONG_DIA” (longest nodule diameter) and a binary variant of “scr_res0” (T0 screening outcome), respectively. Finally, the beta coefficients as described by McWilliams et al. were used to calculate and apply the PanCan model (full model with spiculation) risk scores.[3] Although the PanCan model was designed for predicting lung cancer risk in individual nodules, the same scan variables were used as in our models (Appendix Table 1). Originally, a PanCan score was nodule-dependent and therefore ignored the characteristics of other nodules in the same scan. Consequently, each of our PanCan version’s scores are equal to or greater than what the original PanCan scores would be, though this did not significantly affect the results.

Bootstrapping was performed following the guide by Bodarenko to validate the models and test for overfitting.[4,5] Each model's bootstrap area under the receiver operating characteristic (ROC) curve (AUC) was calculated by fitting each model to a bootstrap sample 1000 times and applying it, separately, to the original sample and the bootstrap sample itself; this required the “predict(),” “prediction(),” and “performance()” functions from the “ROCR” package. The same functions were used to create the continuous table of risk scores including the number of false positives and false negatives. The ROC curves (Figure 2 and Appendix Figure 1) were created using “ggplot()” from the “ggplot2” package; the graph using ROC components (Figure 3) was utilized “plot_roc_components” from the “rmda” package. The 95% confidence intervals of predictive values (sensitivity, specificity, positive predictive value, and negative predictive value) were calculated using the Wilson score method corrected for continuity.[6] The Hosmer-Lemeshow test was used to test model calibration by determining whether the differences between the observed and expected outcomes are statistically significant (a p-value <0.05 would indicate poor calibration).[7] For this, the “plotCalibration()” function from the “PredictABEL” package was used. Decision curve analysis was performed as a means to evaluate the models by incorporating the consequences of false positives and negatives; the methods are described by Vickers and Elkin (2006).[8] In R, this was carried out using the “decision_curve” function from the “rmda” package; the decision curve analysis using the polynomial model is shown in Appendix Figure 2.

REFERENCES

1. McHugh ML. The odds ratio: calculation, usage, and interpretation. Biochem Med. 2009;19(2):120–6.
2. Sperandei S. Understanding logistic regression analysis. Biochem Med. 2014 Feb;24(1):12–8.
3. Mangiafico SS. An R companion for the handbook of biological statistics, version 1.3.0 [Internet]. New Brunswick: Rutgers Cooperative Extension; 2015 [date accessed 2017 Apr 28]. 242–55 p. Available from: rcompanion.org/documents/RCompanionBioStatistics.pdf
4. Bodarenko V. (2015, July 15). The bootstrap approach to managing model uncertainty [Internet]. FI Consulting; 2015 Jul 15 [date accessed 2017 Apr 28]. Available from: https://rstudio-pubs-static.s3.amazonaws.com/90467_c70206f3dc864d53bf36072207ee011d.html#area-under-the-roc-curve
5. Efron B, Tibshirani R. An Introduction to the Bootstrap. United States of America: Chapman and Hall. 1993.
6. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. Stat Med. 1998;17:857–72.
7. Hosmer DW, Lemeshow S. A goodness-of-fit test for the multiple logistic regression model. Commun Statist – Theor Method. 1980 Jan;9:1043–69.
8. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6);565-74.

Appendix Table 1: New scan abnormalities variables created for our models

New variable name	Description	Logic criteria
Scan abnormalities		
sct_emphy	Emphysema on scan	"sct_ab_desc" = 59 (Emphysema)
sct_anyspic	Presence of a spiculated nodule	"sct_margins" = 1 (Spiculated (Stellate))
sct_extranodules	Number of additional nodules	Count of "sct_ab_desc" = 51 (Non-calcified nodule or mass (opacity >= 4 mm diameter)) minus 1
sct_anynonsolid	Presence of a nonsolid nodule	"sct_pre_att" = 2 (Ground glass)
sct_anypartsolid	Presence of a partsolid nodule	"sct_pre_att" = 3 (Mixed)
sct_anyupperlobe	Presence of a nodule located in the upper lobe	"sct_epi_loc" = 1 (Right Upper Lobe) OR 2 (Right Middle Lobe) OR 4 (Left upper Lobe)
Participant characteristics		
fam	Lung cancer diagnosed in first-degree family members	"fambrother" = 1 OR "famchild" = 1 OR "famfather" = 1 OR "fammother" = 1 OR "famsister" = 1
cancbin	Prior diagnosis of cancer	"cancblad"= 1 OR "cancbrea" = 1 OR "canccerv" = 1 OR "canccolo" = 1 OR "cancesop" = 1 OR "canckidn" = 1 OR "canclary" = 1 OR "canclung" = 1 OR "cancnasa = 1 OR "cancoral" = "1" OR "cancpanc" = 1 OR "cancphar" = 1 OR "cancstom" = 1 OR "cancthyr" = 1 OR "canctran" = 1

Appendix Table 2: Base model versus polynomial model variables

Predictor variables	Base model			Polynomial model		
	Odds ratio (95% CI)	P-value	Beta coefficient	Odds ratio (95% CI)	P-value	Beta coefficient
Model constant	N/A	N/A	-10.29	N/A	N/A	-28.15
Age, per yr	1.05 (1.01–1.10)	.014	0.05292	1.79 (0.88–3.93)	.110	0.5845
Age^2, per yr^2	N/A	N/A	N/A	1.00 (0.99–1.00)	.158	-0.004026
Gender, female	N/A	N/A	N/A	N/A	N/A	N/A
Lung cancer in family	-	-	-	-	-	-
Prior diagnosis of cancer	1.76 (0.88–3.04)	.059	0.5623	1.74 (0.88–3.02)	.062	0.5555
Smoking status, active	-	-	-	1.66 (1.22–2.28)	.001	0.5046
Pack years, per pack yr	1.01 (1.00–1.01)	.002	0.008134	1.04 (1.02–1.07)	<.001	0.03922
Pack years^2, per pack yr^2	N/A	N/A	N/A	1.00 (1.00–1.00)	.013	-1.632×10^{-4}
Smoking duration, per yr	1.03 (0.99–1.07)	.160	0.02570	-	-	-
Quit duration, per yr	0.96 (0.92–1.01)	.105	-0.03929	-	-	-
Prior diagnosis of COPD	1.50 (0.84–2.44)	.115	0.4067	1.51 (0.85–2.46)	.110	0.4144
Prior diagnosis of pneumonia	N/A	N/A	N/A	N/A	N/A	N/A
Occupational exposure to asbestos	1.57 (0.81–2.67)	.119	0.4496	-	-	-
Emphysema on scan	1.27 (0.92–1.73)	.135	0.2366	-	-	-
Longest nodule diameter, per mm	0.95 (0.90–0.99)	.028	-0.05149	-	-	-
Longest perpendicular diameter, per mm	1.13 (1.06–1.23)	.001	0.1262	1.11 (1.06–1.15)	<.001	0.09962
Longest perpendicular diameter^2, per mm^2	N/A	N/A	N/A	1.00 (1.00–1.00)	.015	-6.524×10^{-4}
Presence of nonsolid nodule	1.70 (1.03–2.67)	.023	0.5314	1.52 (0.92–2.42)	.077	0.4217
Presence of partsolid nodule	3.05 (1.69–5.07)	<.001	1.116	2.49 (1.36–4.21)	.001	0.9108
Presence of nodule in upper lobe	1.79 (1.19–2.65)	.004	0.5803	1.60 (1.05–2.39)	.021	0.4685
Presence of spiculated nodule	2.27 (1.37–3.61)	.001	0.8195	2.12 (1.29–3.37)	.002	0.7512
Nodule count per scan, per additional nodule	-	-	-	1.67 (0.99–2.98)	.049	0.5128
Nodule count per scan^2, per additional nodule^2	N/A	N/A	N/A	0.82 (0.66–0.97)	.019	-0.1947

The dash signifies that the variable was included in the initial regression analysis but not in the final model and thus does not have a value in the equation.

Abbreviations: CI = confidence interval; COPD = chronic obstructive pulmonary disease; N/A = not applicable (was not included in the initial regression analysis)

Appendix Table 3: Results of the base and polynomial models applied at several number of delayed diagnoses thresholds points

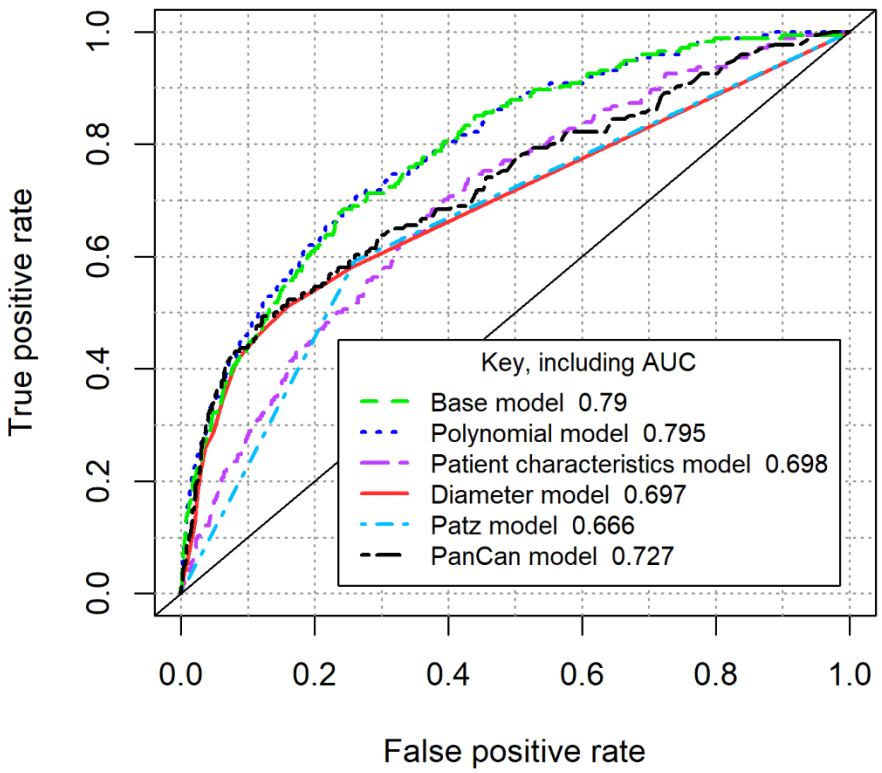
Model	Delayed diagnoses (%) (CI)	Scans avoided (%, CI)	Sensitivity (%) (CI)	Positive predictive value (%) (CI)	Negative predictive value (%) (CI)	Number needed to diagnose (CI)	TNM I	TNM II	TNM III	TNM IV	Missing TNM
No T1 scan	174 (100.0)	0	0.0	100.0	0.0	-	100	22	21	22	9
Base model	0 (0.0, 0.0-2.7)	248 (1.0, 0.9-1.1)	100.0 (97.3-100.0)	0.7 (0.6-0.8)	100.0 (98.1-100.0)	140 (139-144)	0	0	0	0	0
	9 (5.2, 2.5-9.9)	7816 (31.8, 31.3-32.4)	94.8 (90.1-97.5)	1.0 (0.8-1.1)	99.9 (99.8-99.9)	101 (98-108)	5	0	3	1	0
	17 (9.8, 6.0-15.4)	10459 (42.6, 42.0-43.2)	90.2 (84.6-94.0)	1.1 (0.9-1.3)	99.8 (99.7-99.9)	90 (85-97)	10	2	3	2	0
	43 (24.7, 18.6-31.9)	16151 (65.8, 65.2-66.4)	75.3 (68.1-81.4)	1.5 (1.3-1.8)	99.7 (99.6-99.8)	64 (58-72)	23	4	6	7	3
	71 (40.8, 33.5-48.5)	19877 (81.0, 80.5-81.5)	59.2 (51.5-66.5)	2.2 (1.8-2.6)	99.6 (99.5-99.7)	45 (39-53)	32	12	10	13	4
Polynomial model	0 (0.0, 0.0-2.7)	2558 (10.4, 10.0-10.8)	100.0 (97.3-100.0)	0.8 (0.7-0.9)	100.0 (99.8-100.0)	126 (126-130)	0	0	0	0	0
	8 (4.6, 2.2-9.2)	7544 (30.7, 30.2-31.3)	95.4 (90.8-97.8)	1.0 (0.8-1.1)	99.9 (99.8-100.0)	102 (99-108)	5	1	2	0	0
	17 (9.8, 6.0-15.4)	10947 (44.6, 44.0-45.2)	90.2 (84.6-94.0)	1.1 (1.0-1.3)	99.8 (99.7-99.9)	87 (82-93)	9	3	2	1	2
	44 (25.3, 19.2-32.5)	16710 (68.1, 67.5-68.7)	74.7 (67.5-80.8)	1.6 (1.4-1.9)	99.7 (99.6-99.8)	60 (55-68)	21	5	5	9	4
	70 (40.2, 33.0-47.9)	20023 (81.6, 81.1-82.1)	59.8 (52.1-67.0)	2.2 (1.9-2.7)	99.7 (99.6-99.7)	43 (38-51)	35	9	9	13	4

Rows with the same color fill have a similar number of delayed diagnoses at this threshold.
Abbreviations: CI = 95% confidence interval; n = sample size; T1 = first annual follow-up screening round; TNM = the American Joint Committee on Cancer's 7th lung cancer TNM classification and staging system

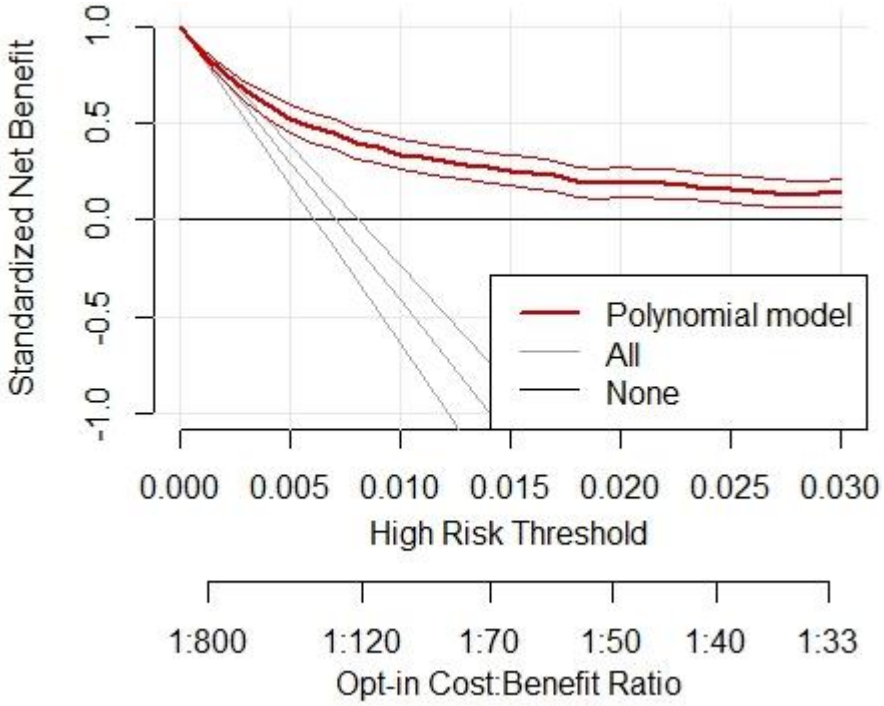
Appendix Table 4: ROC AUC of 1000 bootstrap sample fitted models applied to the original and the bootstrap samples, and the Hosmer-Lemeshow test p-values

Model	ROC AUC original sample (95% CI)	ROC AUC bootstrap sample (95% CI)	Hosmer-Lemeshow test p-value
Base model	0.783 (0.783–0.784)	0.795 (0.794–0.796)	0.359
Polynomial model	0.785 (0.784–0.786)	0.799 (0.797–0.800)	0.668
Patient characteristics model	0.693 (0.692–0.693)	0.703 (0.702–0.704)	0.588
Diameter model	0.697 (0.697–0.697)	0.697 (0.696–0.698)	<0.001
Patz model	0.666 (0.666–0.666)	0.667 (0.666–0.668)	1.000
PanCan model	0.727 (0.727–0.727)	0.727 (0.726–0.729)	<0.001

Abbreviations: CI = confidence interval; ROC AUC = receiver operating characteristic area under the curve



Appendix Figure 1: Logistic regression ROC curves of six risk prediction models
Abbreviations: AUC = area under the ROC curve; ROC = receiver operating characteristic



Appendix Figure 2: Decision curve analysis using the polynomial model
The two thinner lines encompassing each thicker line represent the 95% confidence intervals.