Assessing the Accuracy of a Deep Learning Method to Risk Stratify Indeterminate Pulmonary Nodules

Pierre P. Massion, Sanja Antic, Sarim Ather, Carlos Arteta, Jan Brabec, Heidi Chen, Jerome Declerck, David Dufek, William Hickes, Timor Kadir, Jonas Kunst, Bennett A. Landman, Reginald F. Munden, Petr Novotny, Heiko Peschl, Lyndsey C. Pickup, Catarina Santos, Gary T. Smith, Ambika Talwar, and Fergus Gleeson

ONLINE DATA SUPPLEMENT

Online Supplementary Methodological Materials:

Model implementation details: The LCP-CNN follows the design of the DenseNet architechture with 5 dense blocks, each containing 4 composite functions BN-ReLU-Conv[1×1]-BN-ReLU-Conv[3×3]. The input to the LCP-CNN is resampled at a resolution of 0.25mm x 0.25mm x 1mm, which are augmented at training time using random cropping, flipping and rotations. A 2.5D model is used as it was found in early development that the 3D model did not provide sufficient gains to warrant its use. This is most likely due to the heterogeneity of our dataset in terms of imaging protocols and other factors. The LCP-CNN system was pre-trained using >130,000 images selected and curated to optimally prime the network for subsequent training. Class balancing was used in the CNN training to account for the lower proportion of malignant nodules in the training dataset; without this the resulting CNN would be tuned to benign nodules. During training, the input patches are sampled from the training data such that approximately an equal number of samples from each class (benign and malignant) is used. The network training procedure attempts to optimize a cross-entropy loss function, and the network parameters are updated according to the ADAM optimizer until convergence. The LCP-CNN has been developed using the PyTorch framework for machine learning (1).

An eight-fold cross-validation strategy was used for training and validation on the NLST data where the datasets were split into eight approximately equal subsets, as in Supplementary Figure 1. In each fold, six subsets were used for training, one as an auxiliary split (to check for over-fitting and set internal parameters), and one for (internal) validation; this was repeated eight times selecting a different subset for validation. This approach ensured that in each of the eight folds, each patient appeared only in the training, auxiliary or test set. While cross-validation was used for the internal validation on the NLST data, for the external validation experiments, a single model was created.

1. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Kopf A, Yang A, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang F, Bai J, Chintala S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *NIPS* 2019: 8024-8035.

Supplementary Tables:

Supplementary Table 1. Patient and nodule characteristics for the internal NLST dataset and two external validation datasets, Vanderbilt University Medical Center (Vanderbilt) and Oxford University Hospitals (Oxford).

Supplementary Table 2. Prediction comparison for the LCP-CNN against the two indicated clinical risk models (Brock for the screening dataset, Mayo for the two incidental validation sets). This is a numerical annotation of Figure 3 presented as confusion matrices.

Supplementary Table 3. Net reclassification index numbers with confidence intervals and *P*-values for various thresholds. The p-values represent the significance of the NRI, but the sign of the NRI is important for determining which classifier performs better. This is an extended version of Table 1 built for users of other thresholds.

Supplementary Table 4. Scanner manufacturer and dose information for scans in the two validation sets. Note that due to deidentification procedures and variations in what each scanner model/manufacturer records in the DICOM initially, not all fields are available for all scans, which is why the upper table contains a count of the number of records over which each median and range is calculated. For values which vary over the scan volume, the median value was first recorded for each scan.

Supplementary Figures:

Supplementary Figure 1. (A) Overview of data selection, curation process and composition of the eight-way cross-validation approach for the NLST dataset. Starting with the complete dataset, first all patients randomized to the CT screening arm were selected. Those who did not develop lung cancer within the 7-year trial window were assigned to the benign group, and the others to the cancer group. From the benign group, all studies with reported nodules were examined, and all screen-reported nodules that could be identified on these CTs were marked up. All cancer group nodules that could be unambiguously identified as the diagnosed malignancies reported in NLST metadata were marked up using an extensive multi-pass review process, whether these corresponded to nodules reported at screening time. For illustration, only one-fold of the eight cross-validation rounds is shown. In each fold, the Training set is used to train the LCP-CNN model, the Auxiliary set used to test for over-fitting and to set internal parameters, and the Internal Validation set is used to evaluate the performance of the model. This is then repeated another seven times selecting a different subset for Internal Validation and the remainder for Training and Auxiliary. Summary of (B) Vanderbilt and (C) Oxford datasets.

Supplementary Figure 2. Typical results from the LCP-CNN on incidental nodule findings in the external validation cohorts. Note the different reconstruction kernels: (d) is a very soft scan, while (e) is very hard. Note also the presence of image noise, as in (p), which does not hamper the LCP-CNN.

Supplementary Figure 3. Nodules that are still classified as intermediate by the LCP-CNN. Top: Oxford benign IPNs whose appearance may be quite concerning. Second row: Oxford cancers scoring in the intermediate range. Third row: Vanderbilt benign nodules that caused concern. Nodule (h) is a granuloma related to histoplasmosis. Bottom row: Vanderbilt cancers that scored in the intermediate zone. Case (j) was diagnosed as cancer only 573 days after this CT.

Supplementary Figure 4. Challenging cases for the LCP-CNN. Left half of the figure: benign nodules scoring over 80. Case (e) resolved on subsequent imaging, and (f) was removed with a wedge resection and found to be an abscess. Right half of the figure: cancer nodules with low scores. Note that (h) is still just over 15%, so is still far from the rule-out threshold. Case (g) is a cancer that was not diagnosed for 606 days; the patient was undergoing surveillance for an entirely different non-cancerous lesion. Case (h) took 1872 days to diagnose. The exact diagnosis and time-to-diagnosis information in not available for the Oxford dataset.

Supplementary Figure 5. The three lowest-scoring cancer cases, along with the images of the same cancers closest to the point at which they were diagnosed. (a) Cancer diagnosed 606 days after initial scan; follow-up scores cannot be computed due to missing CT slice data. (b) Cancer diagnosed after 1872 days; LCP-CNN score just before diagnosis date was 93.3. (c) Cancer diagnosed 573 days after initial scan; LCP-CNN score two months before diagnosis was 88.7.

Supplementary Figure 6. Collection of plot and tables for Mayo model results on the internal validation NLST data, including reclassification with respect to Mayo.

Supplementary Table 1.

	NLST		Vanderbilt		Oxford	
	Malignant	Benign	Malignant	Benign	Malignant	Benign
#patients	575	5972	64	52	62	365
#CT volumes	892	10928	64	52	62	367
#nodules	932	14761	64	52	63	400
Sex: Male (%)	336 (58.4)	3683 (61.7)	36 (56.3)	34 (65.4)	32 (51.6)	187 (51.2)
Sex: Female (%)	239 (41.6)	2289 (38.3)	28 (43.8)	18 (34.6)	30 (48.4)	178 (48.8)
Age in Years: Mean (SD)	63.7 (5.3)	62.1 (5.1)	67.7 (8.6)	63.6 (8.1)	69.0 (9.7)	67.7 (12.1)
Nodule size in mm						
Median	15	7.8	19	10	11.7	6.9
IQR	10.0-23.0	6.8-9.9	13.0-21.0	6.0-16.0	9.1-14.3	6.0-8.0
Margin: Smooth (%)	247 (26.5)	(70.6)	1 (1.6)	20 (38.5)	30 (47.6)	341 (85.3)
Margin: Spiculated (%)	462 (49.6)	1749 (11.8)	24 (37.5)	11 (21.2)	20 (31.7)	27 (6.8)
Margin: Other/unreported	223 (23.9)	2592 (17.6)	39 (60.9)	21 (40.4)	13 (20.6)	32 (8.0)
Location: RUL	343 (36.8)	3157 (21.4)	23 (35.9)	16 (30.8)	20 (31.7)	78 (19.5)
Location: RML	52 (5.6)	2212 (15.0)	5 (7.8)	11 (21.2)	6 (9.5)	68 (17.0)
Location: RLL	156 (16.7)	3720 (25.2)	13 (20.3)	14 (26.9)	15 (23.8)	114 (28.5)
Location: LUL	215 (23.1)	1769 (12.0)	17 (26.6)	6 (11.5)	14 (22.2)	32 (8.0)
Location: LLL	141 (15.1)	3150 (21.3)	6 (9.4)	5 (9.6)	7 (11.1)	95 (23.8)
Location: Other/Unreported	25 (2.7)	753 (5.1)	0 (0.0)	0 (0.0)	1 (1.6)	13 (3.3)
Hist: Adenocarcinoma	326 (56.7)		39 (60.9)		39 (62.9)	
Hist: Squamous Cell	133 (23.1)		9 (14.1)		9 (14.5)	
Hist: NSCLC	72 (12.5)		1 (1.6)		4 (6.5)	
Hist: Small-cell Lung Cancer	30 (5.2)		5 (7.8)		1 (1.6)	
Hist: Other	11 (1.9)		10 (15.6)		8 (12.9)	
Hist: Unreported	3 (0.5)		0 (0.0)		1 (1.6)	

Supplementary Table 2: Confusion matrix-style results

		Brock			
	NLST Cases	**Low≤5%**	**Intermediate**	**High≥65%**	Total
LCP-CNN	**Low≤5%**	26	15	0	41
	Intermediate	80	218	17	315
	High≥65%	20	485	71	576
		126	718	88	932

		Brock			
	NLST Benign	**Low≤5%**	**Intermediate**	**High≥65%**	Total
LCP-CNN	**Low≤5%**	7507	1771	12	9290
	Intermediate	2206	2472	49	4727
	High≥65%	101	609	34	744
		9814	4852	95	14761

		Mayo			
	VUMC Cases	**Low≤5%**	**Intermediate**	**High≥65%**	Total
LCP-CNN	**Low≤5%**	0	1	0	1
	Intermediate	1	16	1	18
	High≥65%	0	30	15	45
		1	47	16	64

		Mayo			
	VUMC Benign	**Low≤5%**	**Intermediate**	**High≥65%**	Total
LCP-CNN	**Low≤5%**	6	17	0	23
	Intermediate	0	15	3	18
	High≥65%	0	9	2	11
		6	41	5	52

		Mayo			
	Oxford Cases	**Low≤5%**	**Intermediate**	**High≥65%**	Total
LCP-CNN	**Low≤5%**	0	2	0	2
	Intermediate	0	38	0	38
	High≥65%	0	20	3	23
		0	60	3	63

		Mayo			
	Oxford Benign	**Low≤5%**	**Intermediate**	**High≥65%**	Total
LCP-CNN	**Low≤5%**	9	248	0	257
	Intermediate	3	130	0	133
	High≥65%	0	9	1	10
		12	387	1	400

Supplementary Table 3.

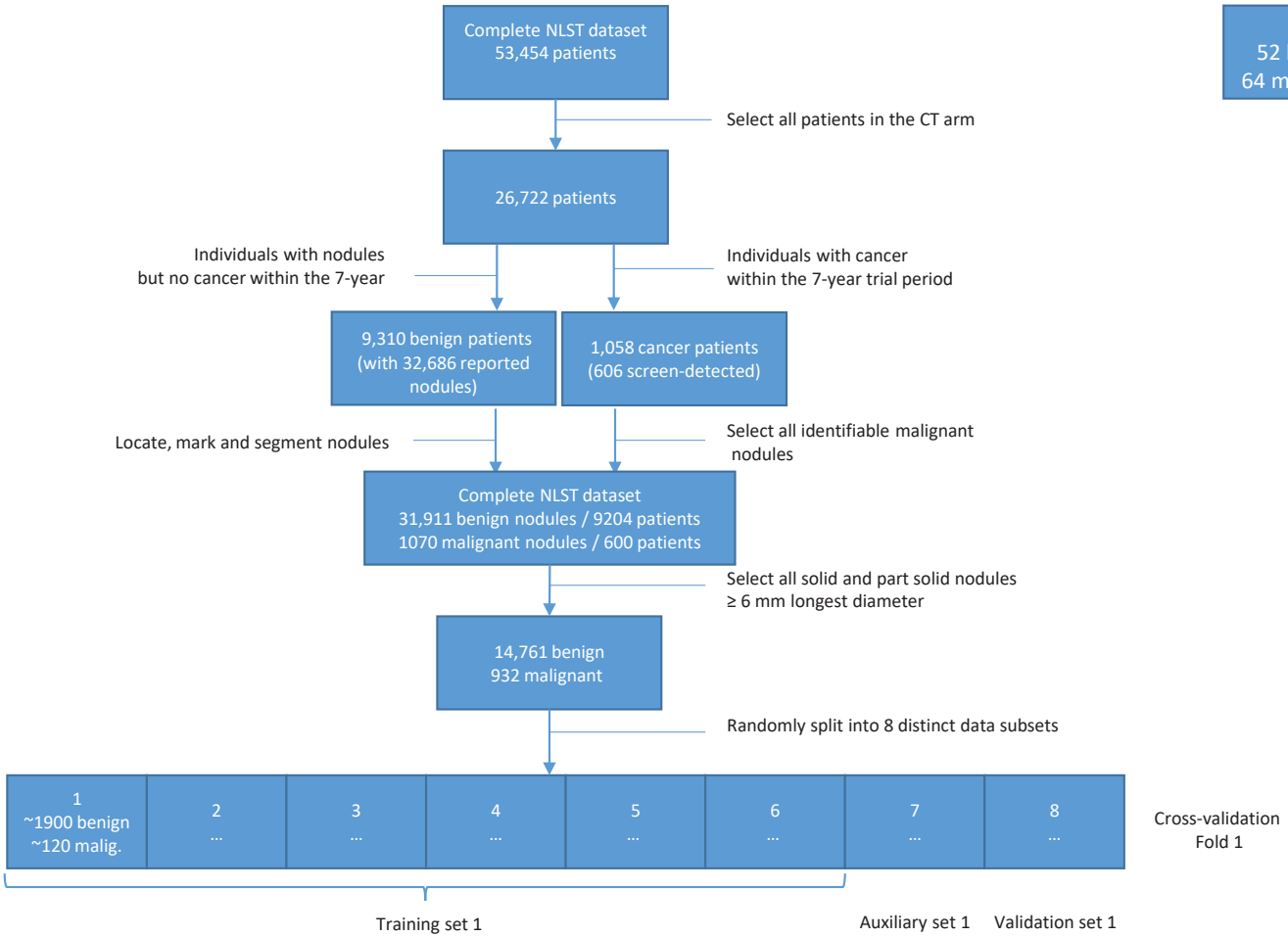
Vanderbilt Reclassification: Comparing to Mayo (Incidental population)										
Thr.	Cancer up	Cancer down	Net Cancer	P-val Cancer	Benign up	Benign down	Net Benign	P-val Ben.	Overall	P-val Overall
5	0.02 (0.00-0.05)	0.02 (0.00-0.05)	0.00 (-0.05-0.05)	0.3423	0.00 (0.00-0.00)	0.33 (0.19-0.46)	0.33 (0.19-0.46)	<.0001	0.33 (0.20-0.47)	<.0001
10	0.06 (0.02-0.13)	0.02 (0.00-0.05)	0.05 (-0.02-0.13)	0.046	0.04 (0.00-0.10)	0.15 (0.06-0.25)	0.12 (0.00-0.23)	0.0155	0.16 (0.03-0.30)	0.0074
15	0.08 (0.02-0.16)	0.02 (0.00-0.05)	0.06 (0.00-0.14)	0.024	0.10 (0.02-0.17)	0.08 (0.02-0.15)	-0.02 (-0.13-0.10)	0.4324	0.04 (-0.09-0.18)	0.26
65	0.47 (0.34-0.59)	0.02 (0.00-0.05)	0.45 (0.33-0.58)	<.0001	0.17 (0.08-0.29)	0.06 (0.00-0.13)	-0.12 (-0.25-0.00)	0.0439	0.34 (0.15-0.52)	0.0004
70	0.50 (0.38-0.63)	0.00 (0.00-0.00)	0.50 (0.38-0.63)	<.0001	0.13 (0.06-0.23)	0.06 (0.00-0.13)	-0.08 (-0.19-0.04)	0.123	0.42 (0.25-0.59)	<.0001
80	0.50 (0.38-0.63)	0.00 (0.00-0.00)	0.50 (0.38-0.63)	<.0001	0.13 (0.06-0.23)	0.04 (0.00-0.10)	-0.10 (-0.21-0.02)	0.0588	0.40 (0.24-0.57)	<.0001
Oxford Reclassification: Comparing to Mayo (Incidental population)										
Thr.	Cancer up	Cancer down	Net Cancer	P-val Cancer	Benign up	Benign down	Net Benign	P-val Ben.	Overall	P-val Overall
5	0.00 (0.00-0.00)	0.03 (0.00-0.08)	-0.03 (-0.08-0.00)	0.1364	0.01 (0.00-0.02)	0.62 (0.57-0.67)	0.61 (0.56-0.66)	<.0001	0.58 (0.51-0.64)	<.0001
10	0.02 (0.00-0.05)	0.06 (0.02-0.13)	-0.05 (-0.11-0.02)	0.1189	0.02 (0.01-0.04)	0.54 (0.49-0.59)	0.52 (0.47-0.57)	<.0001	0.47 (0.39-0.56)	<.0001
15	0.06 (0.02-0.13)	0.11 (0.05-0.19)	-0.05 (-0.16-0.05)	0.2177	0.04 (0.02-0.06)	0.42 (0.37-0.47)	0.38 (0.33-0.44)	<.0001	0.33 (0.22-0.45)	<.0001
65	0.32 (0.21-0.43)	0.00 (0.00-0.00)	0.32 (0.21-0.43)	<.0001	0.02 (0.01-0.04)	0.00 (0.00-0.00)	-0.02 (-0.04--0.01)	<.0001	0.29 (0.18-0.41)	<.0001
70	0.27 (0.16-0.38)	0.00 (0.00-0.00)	0.27 (0.16-0.38)	<.0001	0.02 (0.01-0.04)	0.00 (0.00-0.00)	-0.02 (-0.04--0.01)	0.0002	0.25 (0.14-0.36)	<.0001
80	0.24 (0.14-0.35)	0.00 (0.00-0.00)	0.24 (0.14-0.35)	<.0001	0.01 (0.00-0.02)	0.00 (0.00-0.00)	-0.01 (-0.02-0.00)	0.0504	0.23 (0.13-0.34)	<.0001
NLST Reclassification: Comparing to Brock (Screening population)										
Thr.	Cancer up	Cancer down	Net Cancer	P-val Cancer	Benign up	Benign down	Net Benign	P-val Ben.	Overall	P-val Overall
5	0.11 (0.09-0.13)	0.02 (0.01-0.02)	0.09 (0.07-0.11)	<.0001	0.16 (0.15-0.16)	0.12 (0.12-0.13)	-0.04 (-0.04--0.03)	<.0001	0.06 (0.03-0.08)	<.0001
10	0.18 (0.16-0.21)	0.02 (0.01-0.02)	0.17 (0.14-0.19)	<.0001	0.16 (0.15-0.17)	0.06 (0.06-0.07)	-0.10 (-0.11--0.09)	<.0001	0.07 (0.04-0.10)	<.0001
15	0.26 (0.23-0.28)	0.02 (0.01-0.02)	0.24 (0.21-0.27)	<.0001	0.15 (0.14-0.15)	0.04 (0.04-0.05)	-0.10 (-0.11--0.10)	<.0001	0.14 (0.11-0.17)	<.0001
65	0.54 (0.51-0.57)	0.02 (0.01-0.03)	0.52 (0.49-0.56)	<.0001	0.05 (0.04-0.05)	0.00 (0.00-0.01)	-0.04 (-0.05--0.04)	<.0001	0.48 (0.45-0.51)	<.0001
70	0.54 (0.51-0.57)	0.01 (0.01-0.02)	0.53 (0.49-0.56)	<.0001	0.04 (0.04-0.04)	0.00 (0.00-0.00)	-0.04 (-0.04--0.03)	<.0001	0.49 (0.45-0.52)	<.0001
80	0.45 (0.41-0.48)	0.00 (0.00-0.01)	0.44 (0.41-0.48)	<.0001	0.02 (0.02-0.03)	0.00 (0.00-0.00)	-0.02 (-0.02--0.02)	<.0001	0.42 (0.39-0.45)	<.0001

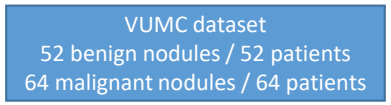
Supplementary Table 4: Scanner manufacturer and dose information for validation sites.

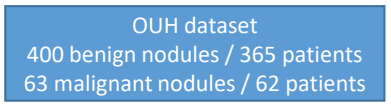
	Vanderbilt			Oxford		
Field	**#Records**	**Median**	**Range**	**#Records**	**Median**	**Range**
CTDIvol	91	7.2	1.2-27.0	52	14.6	2.2-32.8
Exposure	114	99	1.0-375.0	461	1	0.0-4805.0
KVP	114	120	100.0-120.0	461	120	100.0-140.0
XrayTubeCurrent	114	168.5	30.0-1228.0	461	110	20.0-700.0

Manufacturer	**#VUMC scans**	**Main Vanderbilt Models**	**#Oxford scans**	**Main Oxford Models**
GE	407	Discovery 690 (38), Discovery 710 (34), LightSpeed Pro 16 (37)	20	BrightSpeed (15), Discovery STE (5)
SIEMENS	2	SOMATOM Definition AS+ (1), Sensation 16 (1)	15	SOMATOM Definition AS (3), SOMATOM Force (11), Sensation 16 (1)
TOSHIBA	51	Aquilion (49), Aquilion ONE (2)	0	n/a
Philips	0	n/a	79	Brilliance 6 (1), Brilliance 64 (13), Ingenuity CT (8)

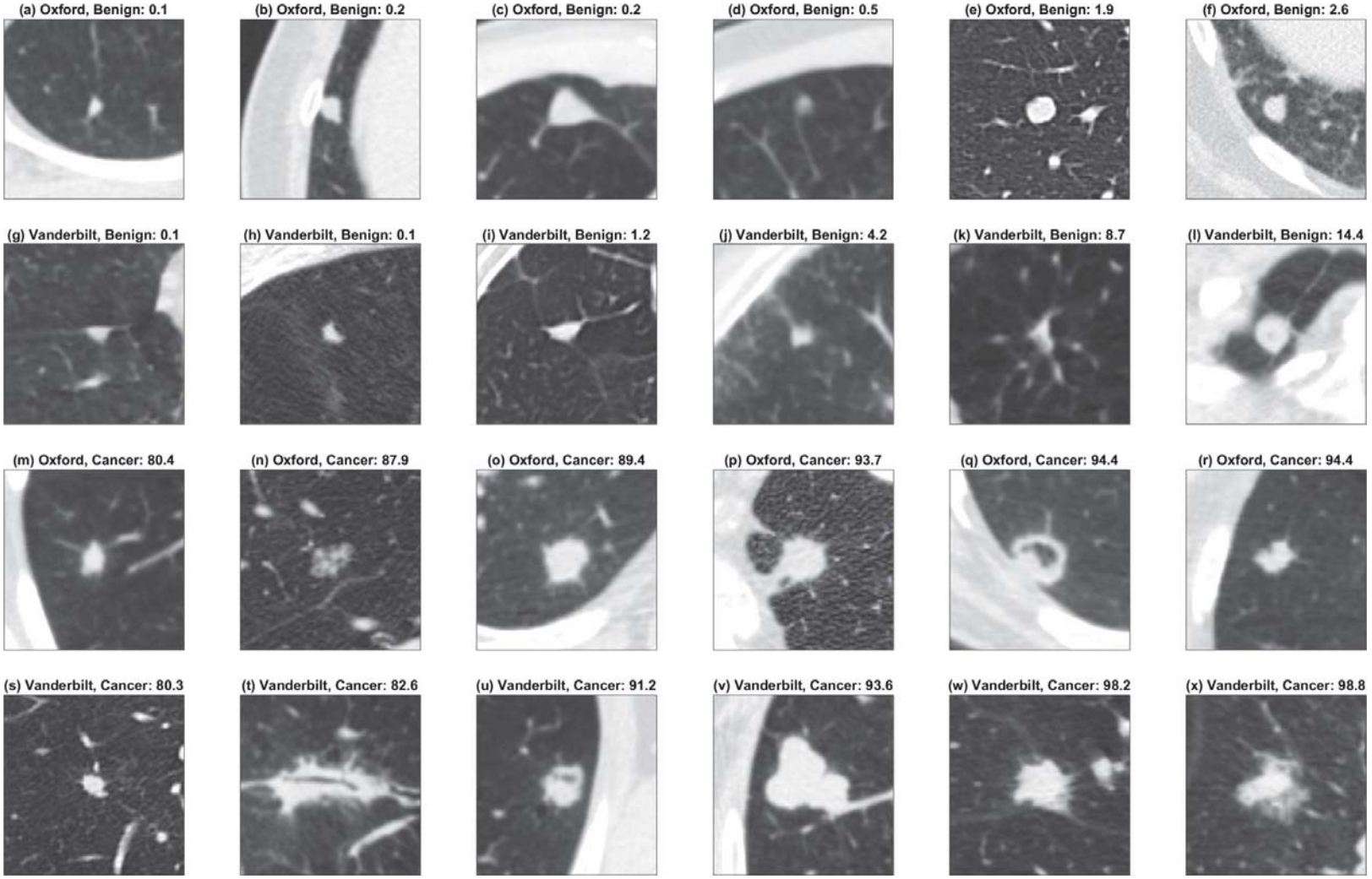
Supplementary Figure 1.



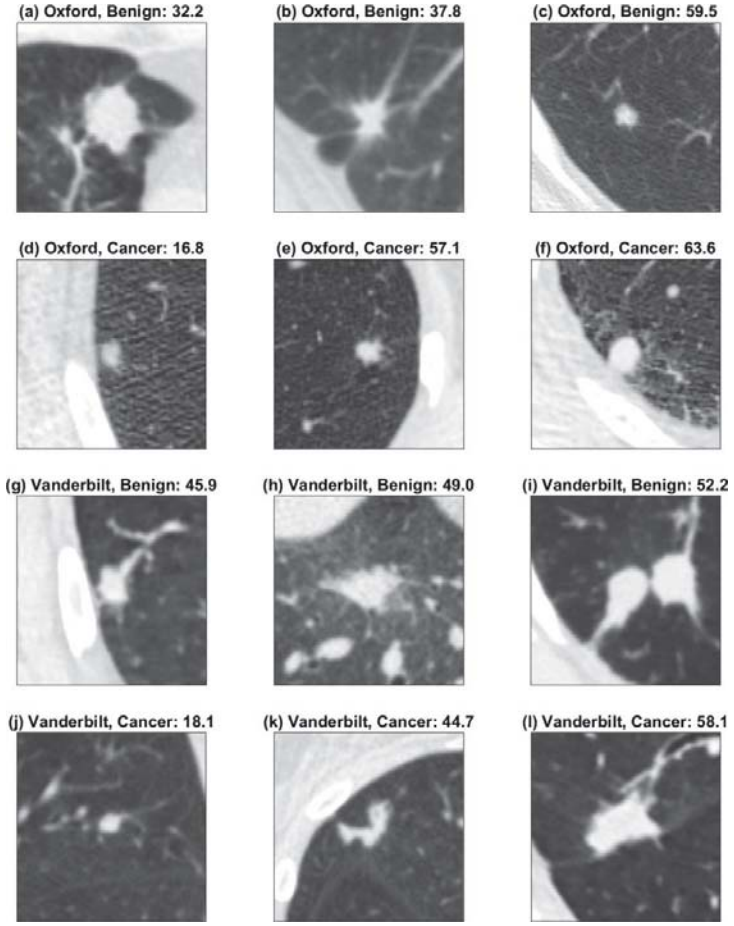




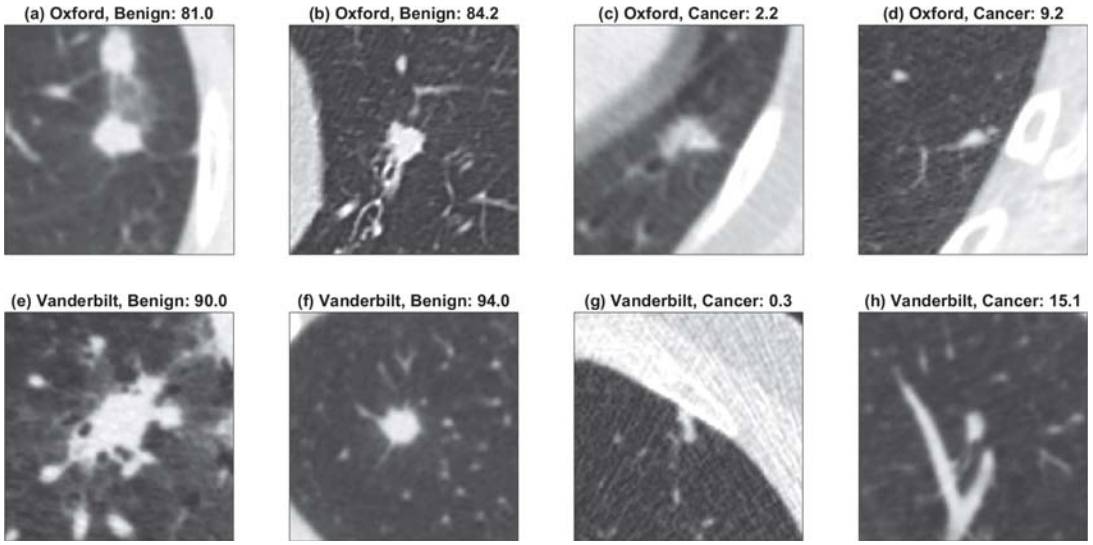
Supplementary Figure 2.



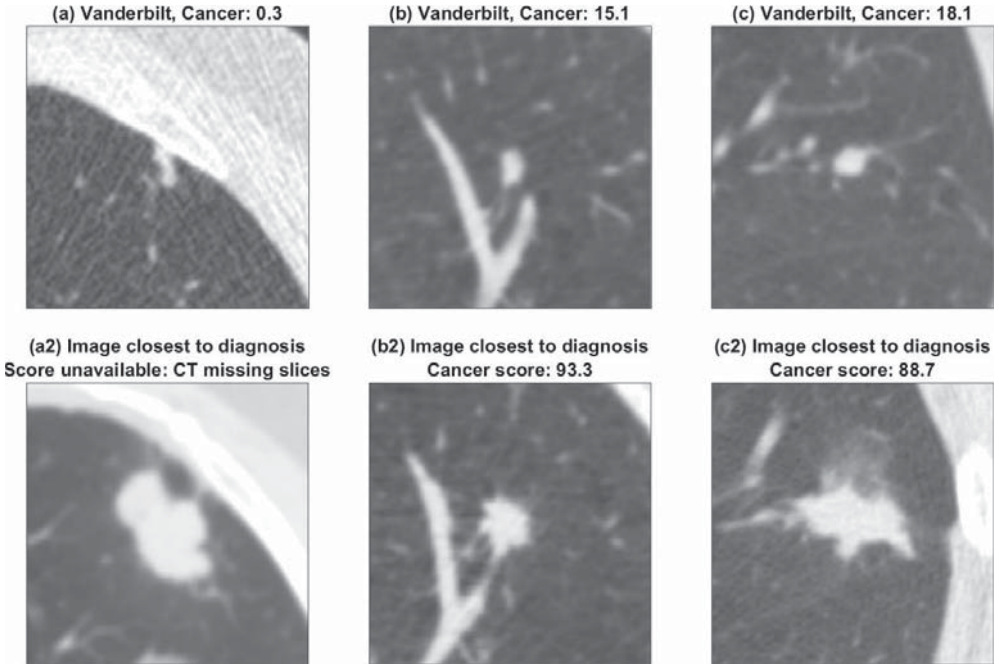
Supplementary Figure 3.



Supplementary Figure 4.



Supplementary Figure 5.



Supplementary Figure 6: NLST Mayo comparison result set

NLST Reclassification: Comparing to Mayo (for comparison to other datasets)

Target	Cancer up (95%CI)	Cancer down (95%CI)	Net Cancer (95%CI)	Net Cancer *P*-val	Benign up (95%CI)	Benign down (95%CI)	Net Benign (95%CI)	Net Benign *P*-val	Overall (95%CI)	Overall *P*-val
5%	0.00 (0.00-0.00)	0.04 (0.03-0.06)	-0.04 (-0.06--0.03)	<.0001	0.01 (0.00-0.01)	0.59 (0.59-0.60)	0.59 (0.58-0.60)	<.0001	0.54 (0.53-0.56)	<.0001
65%	0.42 (0.38-0.45)	0.05 (0.04-0.06)	0.37 (0.33-0.40)	<.0001	0.04 (0.04-0.04)	0.02 (0.02-0.02)	-0.02 (-0.02--0.02)	<.0001	0.35 (0.31-0.38)	<.0001

	Mayo		
Threshold	**Sensitivity**	**Specificity**	**DLR-**
5%	100.0 (100.0-100.0)	4.1 (3.8-4.5)	0.00 (0.00-0.00)
Threshold	**Sens**	**Specificity**	**DLR+**
65%	25.1 (22.3-27.9)	96.9 (96.6-97.1)	7.99 (6.90-9.23)