SUPPLEMENTARY INFORMATION
https://doi.org/10.1038/s41591-019-0447-x

nature medicine

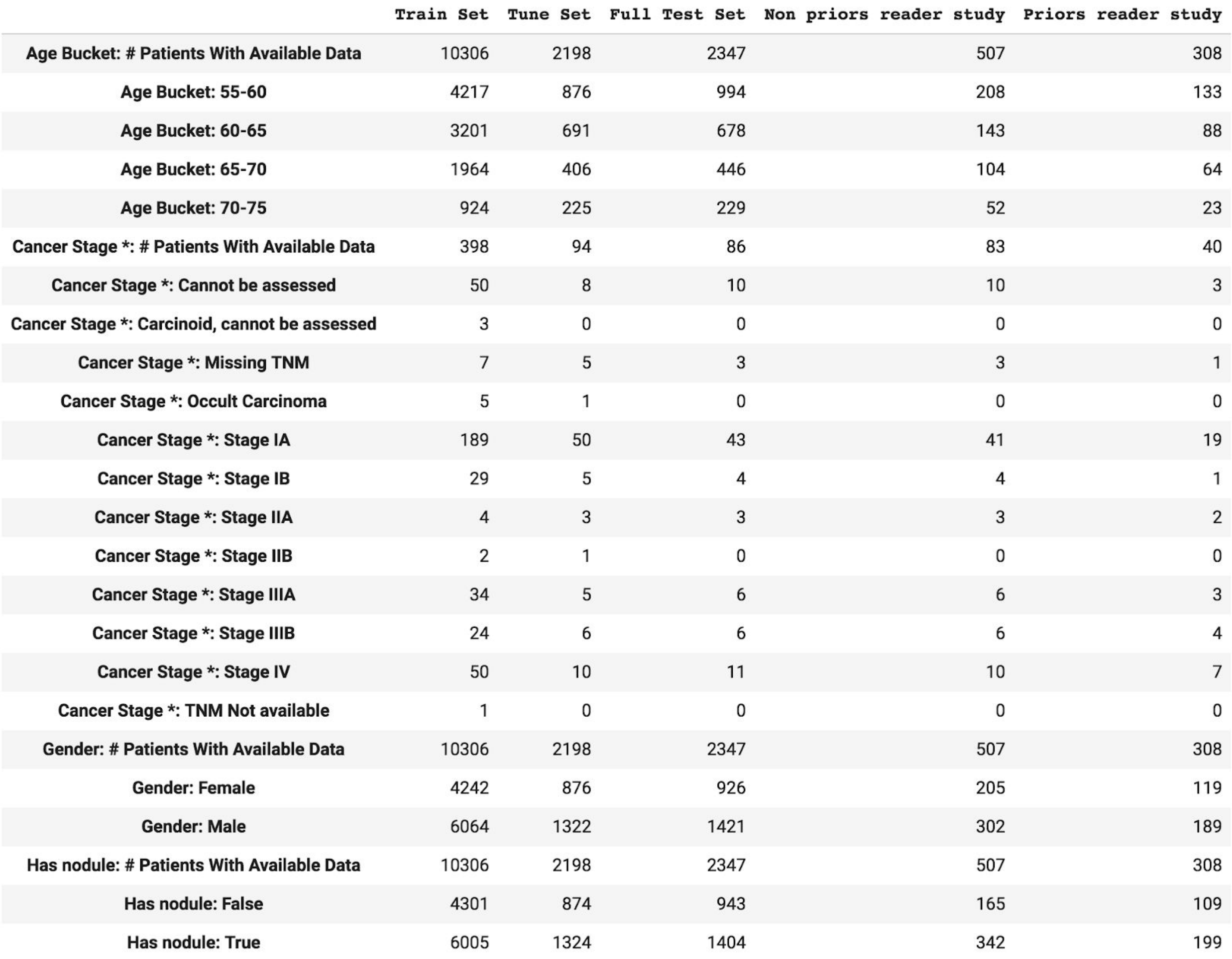
In the format provided by the authors and unedited.

End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography

Diego Ardila[1,5], Atilla P. Kiraly[1,5], Sujeeth Bharadwaj[1,5], Bokyung Choi[1,5], Joshua J. Reicher[2], Lily Peng[1], Daniel Tse[1*], Mozziyar Etemadi[3], Wenxing Ye[1], Greg Corrado[1], David P. Naidich[4] and Shravya Shetty[1]

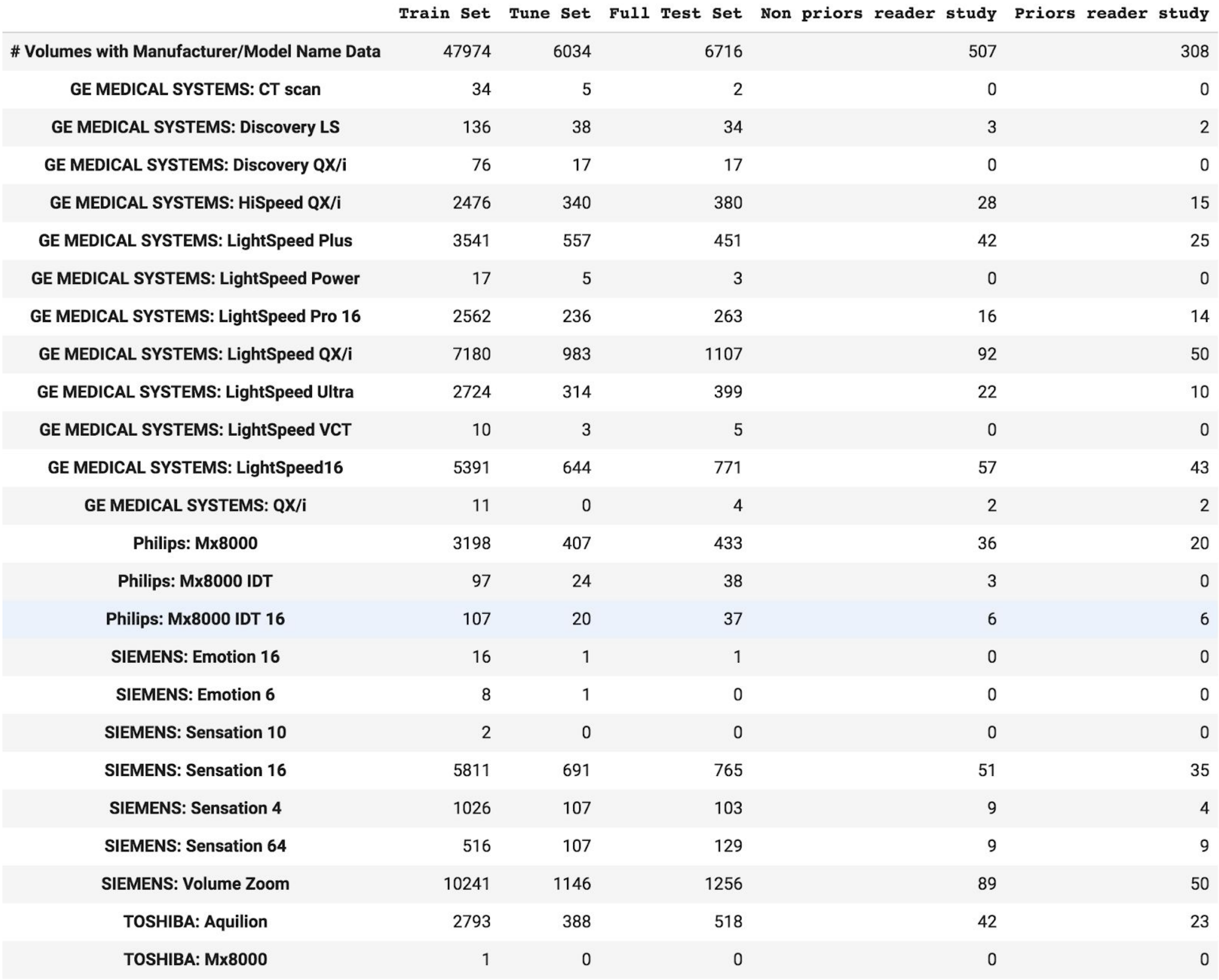
[1]Google AI, Mountain View, CA, USA. [2]Stanford Health Care and Palo Alto Veterans Affairs, Palo Alto, CA, USA. [3]Northwestern Medicine, Chicago, IL, USA. [4]New York University-Langone Medical Center, Center for Biological Imaging, New York City, NY, USA. [5]These authors contributed equally: Diego Ardila, Atilla P. Kiraly, Sujeeth Bharadwaj, Bokyung Choi. *e-mail: tsed@google.com

SUPPLEMENTARY INFORMATION



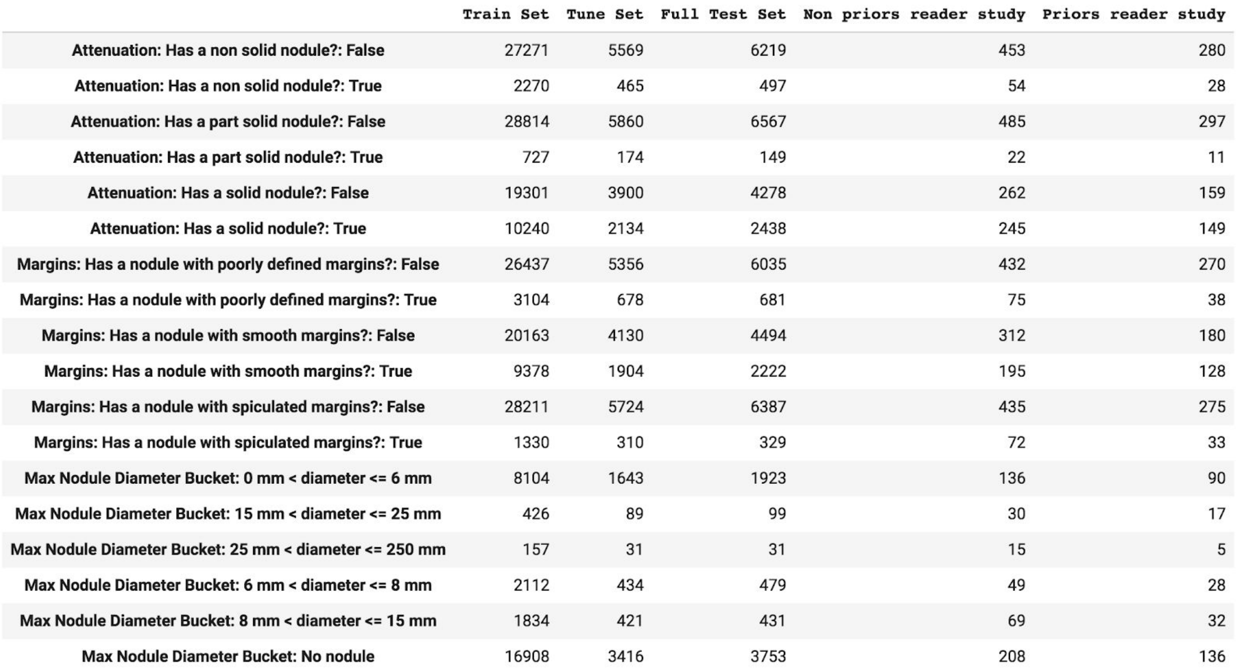
	Train Set	Tune Set	Full Test Set	Non priors reader study	Priors reader study
Age Bucket: # Patients With Available Data	10306	2198	2347	507	308
Age Bucket: 55-60	4217	876	994	208	133
Age Bucket: 60-65	3201	691	678	143	88
Age Bucket: 65-70	1964	406	446	104	64
Age Bucket: 70-75	924	225	229	52	23
Cancer Stage *: # Patients With Available Data	398	94	86	83	40
Cancer Stage *: Cannot be assessed	50	8	10	10	3
Cancer Stage *: Carcinoid, cannot be assessed	3	0	0	0	0
Cancer Stage *: Missing TNM	7	5	3	3	1
Cancer Stage *: Occult Carcinoma	5	1	0	0	0
Cancer Stage *: Stage IA	189	50	43	41	19
Cancer Stage *: Stage IB	29	5	4	4	1
Cancer Stage *: Stage IIA	4	3	3	3	2
Cancer Stage *: Stage IIB	2	1	0	0	0
Cancer Stage *: Stage IIIA	34	5	6	6	3
Cancer Stage *: Stage IIIB	24	6	6	6	4
Cancer Stage *: Stage IV	50	10	11	10	7
Cancer Stage *: TNM Not available	1	0	0	0	0
Gender: # Patients With Available Data	10306	2198	2347	507	308
Gender: Female	4242	876	926	205	119
Gender: Male	6064	1322	1421	302	189
Has nodule: # Patients With Available Data	10306	2198	2347	507	308
Has nodule: False	4301	874	943	165	109
Has nodule: True	6005	1324	1404	342	199

Supplementary Table [1]: Demographics and cancer staging breakdown of patients in NLST subsets used. The total number of patients with cancer staging information (affected rows indicated by an asterisk *) is greater than the number of cancer patients reported in this table because some patients received a cancer diagnosis after the initial 3 years of screening, and were therefore considered cancer negative patients in our main analysis. In this table we only show patients that were considered cancer positive in our main analysis.



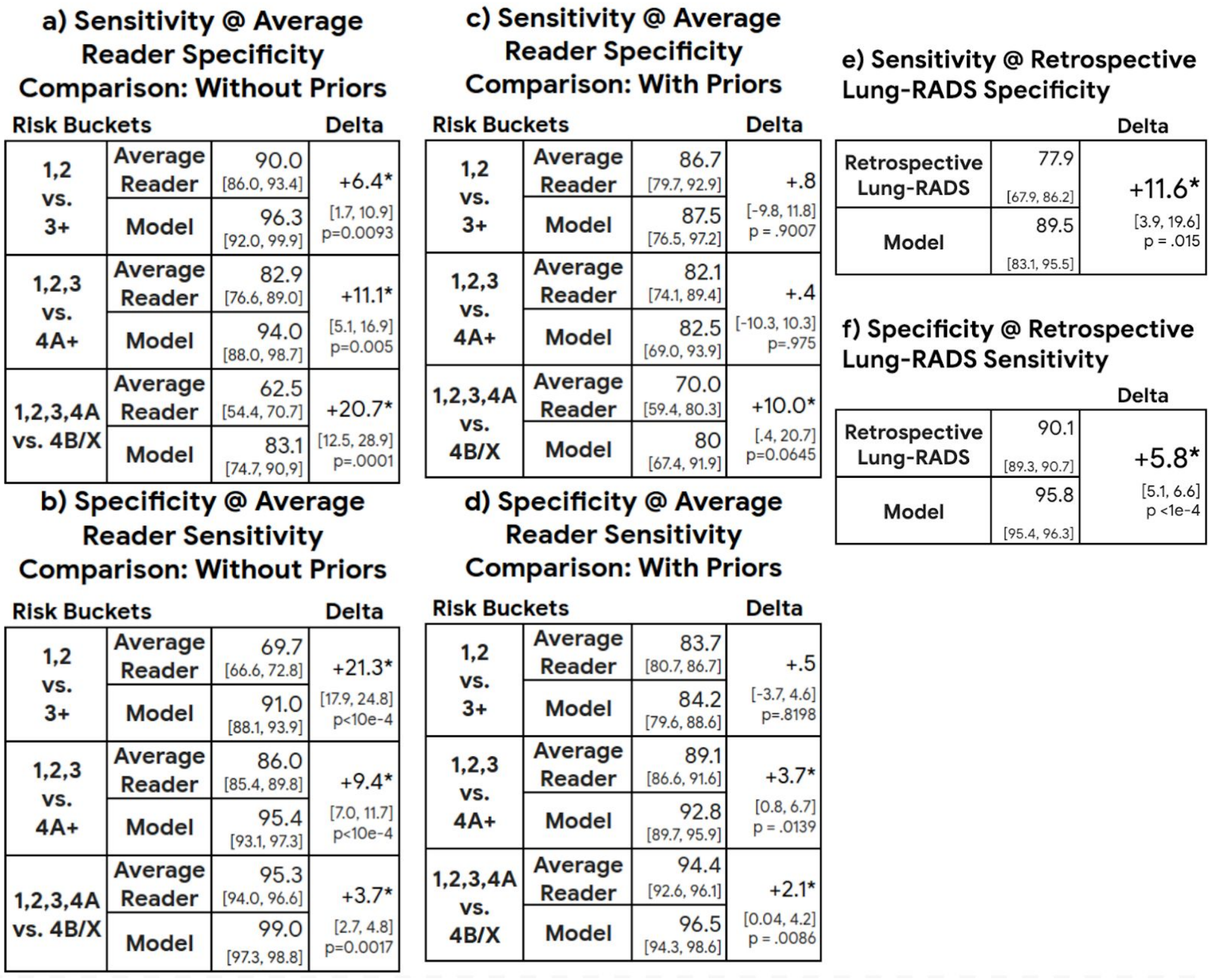
	Train Set	Tune Set	Full Test Set	Non priors reader study	Priors reader study
# Volumes with Manufacturer/Model Name Data	47974	6034	6716	507	308
GE MEDICAL SYSTEMS: CT scan	34	5	2	0	0
GE MEDICAL SYSTEMS: Discovery LS	136	38	34	3	2
GE MEDICAL SYSTEMS: Discovery QX/i	76	17	17	0	0
GE MEDICAL SYSTEMS: HiSpeed QX/i	2476	340	380	28	15
GE MEDICAL SYSTEMS: LightSpeed Plus	3541	557	451	42	25
GE MEDICAL SYSTEMS: LightSpeed Power	17	5	3	0	0
GE MEDICAL SYSTEMS: LightSpeed Pro 16	2562	236	263	16	14
GE MEDICAL SYSTEMS: LightSpeed QX/i	7180	983	1107	92	50
GE MEDICAL SYSTEMS: LightSpeed Ultra	2724	314	399	22	10
GE MEDICAL SYSTEMS: LightSpeed VCT	10	3	5	0	0
GE MEDICAL SYSTEMS: LightSpeed16	5391	644	771	57	43
GE MEDICAL SYSTEMS: QX/i	11	0	4	2	2
Philips: Mx8000	3198	407	433	36	20
Philips: Mx8000 IDT	97	24	38	3	0
Philips: Mx8000 IDT 16	107	20	37	6	6
SIEMENS: Emotion 16	16	1	1	0	0
SIEMENS: Emotion 6	8	1	0	0	0
SIEMENS: Sensation 10	2	0	0	0	0
SIEMENS: Sensation 16	5811	691	765	51	35
SIEMENS: Sensation 4	1026	107	103	9	4
SIEMENS: Sensation 64	516	107	129	9	9
SIEMENS: Volume Zoom	10241	1146	1256	89	50
TOSHIBA: Aquilion	2793	388	518	42	23
TOSHIBA: Mx8000	1	0	0	0	0

Supplementary Table [2]: Manufacturer and model distributions for cases in the NLST subsets used. The number of volumes per manufacturer CT scanner model is shown for each column.



	Train Set	Tune Set	Full Test Set	Non priors reader study	Priors reader study
Attenuation: Has a non solid nodule?: False	27271	5569	6219	453	280
Attenuation: Has a non solid nodule?: True	2270	465	497	54	28
Attenuation: Has a part solid nodule?: False	28814	5860	6567	485	297
Attenuation: Has a part solid nodule?: True	727	174	149	22	11
Attenuation: Has a solid nodule?: False	19301	3900	4278	262	159
Attenuation: Has a solid nodule?: True	10240	2134	2438	245	149
Margins: Has a nodule with poorly defined margins?: False	26437	5356	6035	432	270
Margins: Has a nodule with poorly defined margins?: True	3104	678	681	75	38
Margins: Has a nodule with smooth margins?: False	20163	4130	4494	312	180
Margins: Has a nodule with smooth margins?: True	9378	1904	2222	195	128
Margins: Has a nodule with spiculated margins?: False	28211	5724	6387	435	275
Margins: Has a nodule with spiculated margins?: True	1330	310	329	72	33
Max Nodule Diameter Bucket: 0 mm < diameter <= 6 mm	8104	1643	1923	136	90
Max Nodule Diameter Bucket: 15 mm < diameter <= 25 mm	426	89	99	30	17
Max Nodule Diameter Bucket: 25 mm < diameter <= 250 mm	157	31	31	15	5
Max Nodule Diameter Bucket: 6 mm < diameter <= 8 mm	2112	434	479	49	28
Max Nodule Diameter Bucket: 8 mm < diameter <= 15 mm	1834	421	431	69	32
Max Nodule Diameter Bucket: No nodule	16908	3416	3753	208	136

Supplementary Table [3]: Attenuation, margin, and diameter volume counts of relevant subsets of data. These were generated using the nodule annotations in NLST, which happen once per patient year. The NLST data did not provide a reliable way of linking these back to a specific volume, but for all sets besides the training sets we used heuristics to select a single stack per screening year which was most likely to have been the stack that was used to generate the nodule annotations. However, this means that the volume counts are likely to be less accurate for the training set, where we did not apply heuristics to select a single stack per screening year.



a) Sensitivity @ Average Reader Specificity Comparison: Without Priors

Risk Buckets			Delta
1,2 vs. 3+	Average Reader	90.0 [86.0, 93.4]	+6.4* [1.7, 10.9] p=0.0093
	Model	96.3 [92.0, 99.9]	
1,2,3 vs. 4A+	Average Reader	82.9 [76.6, 89.0]	+11.1* [5.1, 16.9] p=0.005
	Model	94.0 [88.0, 98.7]	
1,2,3,4A vs. 4B/X	Average Reader	62.5 [54.4, 70.7]	+20.7* [12.5, 28.9] p=.0001
	Model	83.1 [74.7, 90,9]	

b) Specificity @ Average Reader Sensitivity Comparison: Without Priors

Risk Buckets			Delta
1,2 vs. 3+	Average Reader	69.7 [66.6, 72.8]	+21.3* [17.9, 24.8] p<10e-4
	Model	91.0 [88.1, 93.9]	
1,2,3 vs. 4A+	Average Reader	86.0 [85.4, 89.8]	+9.4* [7.0, 11.7] p<10e-4
	Model	95.4 [93.1, 97.3]	
1,2,3,4A vs. 4B/X	Average Reader	95.3 [94.0, 96.6]	+3.7* [2.7, 4.8] p=0.0017
	Model	99.0 [97.3, 98.8]	

c) Sensitivity @ Average Reader Specificity Comparison: With Priors

Risk Buckets			Delta
1,2 vs. 3+	Average Reader	86.7 [79.7, 92.9]	+.8 [-9.8, 11.8] p = .9007
	Model	87.5 [76.5, 97.2]	
1,2,3 vs. 4A+	Average Reader	82.1 [74.1, 89.4]	+.4 [-10.3, 10.3] p=.975
	Model	82.5 [69.0, 93.9]	
1,2,3,4A vs. 4B/X	Average Reader	70.0 [59.4, 80.3]	+10.0* [.4, 20.7] p=0.0645
	Model	80 [67.4, 91.9]	

d) Specificity @ Average Reader Sensitivity Comparison: With Priors

Risk Buckets			Delta
1,2 vs. 3+	Average Reader	83.7 [80.7, 86.7]	+.5 [-3.7, 4.6] p=.8198
	Model	84.2 [79.6, 88.6]	
1,2,3 vs. 4A+	Average Reader	89.1 [86.6, 91.6]	+3.7* [0.8, 6.7] p = .0139
	Model	92.8 [89.7, 95.9]	
1,2,3,4A vs. 4B/X	Average Reader	94.4 [92.6, 96.1]	+2.1* [0.04, 4.2] p = .0086
	Model	96.5 [94.3, 98.6]	

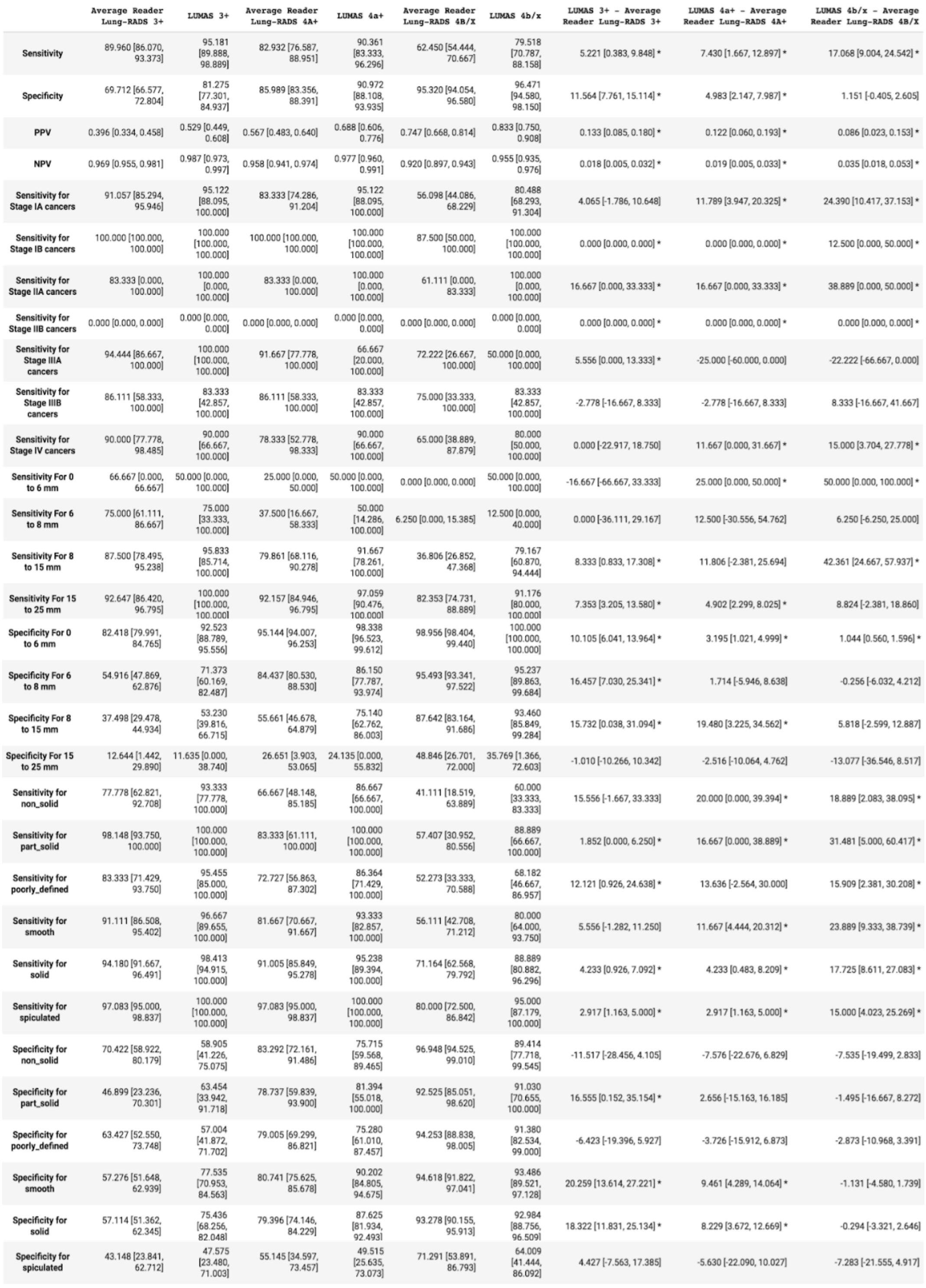
e) Sensitivity @ Retrospective Lung-RADS Specificity

		Delta
Retrospective Lung-RADS	77.9 [67.9, 86.2]	+11.6* [3.9, 19.6] p = .015
Model	89.5 [83.1, 95.5]	

f) Specificity @ Retrospective Lung-RADS Sensitivity

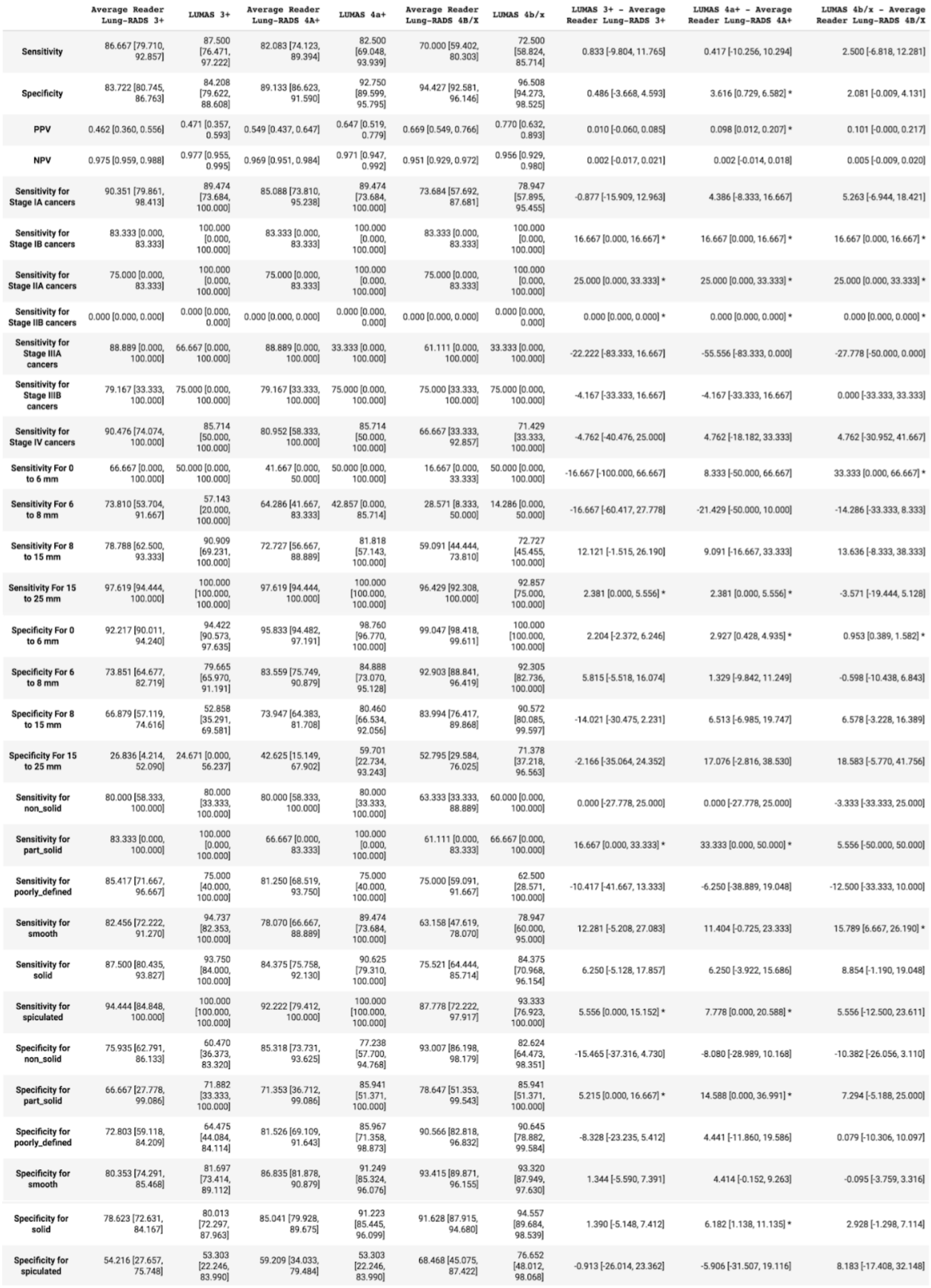
		Delta
Retrospective Lung-RADS	90.1 [89.3, 90.7]	+5.8* [5.1, 6.6] p <1e-4
Model	95.8 [95.4, 96.3]	

Supplementary Table [4]: Results with matched specificity and sensitivity. For the reader study without priors as an alternative to LUMAS, we (a) set the model's specificity to match the average reader specificity and then compared sensitivity and matched sensitivity and then (b) matched the average reader sensitivity and compared specificity. In both cases analysis was conducted with n=507 volumes from 507 patients. The same was done for the prior reader study in (c) and (d) with n=308 volumes from 308 patients. In the entire NLST test set, the matched specificity and sensitivity to NLST readers are shown for matched specificity (e) and matched sensitivity (f), comparing on n=6,716 cases. Note that these comparisons are more favorable to the model because they are based on operating points that maximize the delta. All comparisons in this table were made using a two-sided permutation test using 10,000 random resamplings of the data.



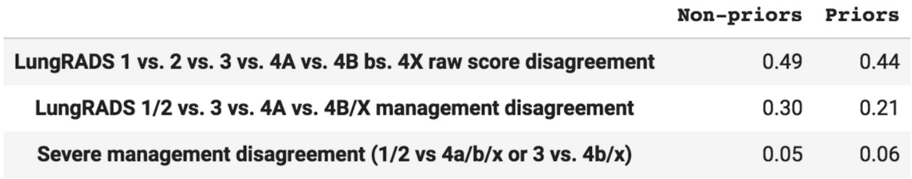
	Average Reader Lung-RADS 3+	LUMAS 3+	Average Reader Lung-RADS 4A+	LUMAS 4a+	Average Reader Lung-RADS 4B/X	LUMAS 4b/x	LUMAS 3+ - Average Reader Lung-RADS 3+	LUMAS 4a+ - Average Reader Lung-RADS 4A+	LUMAS 4b/x - Average Reader Lung-RADS 4B/X
Sensitivity	89.960 [86.070, 93.373]	95.181 [89.888, 98.889]	82.932 [76.587, 88.951]	90.361 [83.333, 96.296]	62.450 [54.444, 70.667]	79.518 [70.787, 88.158]	5.221 [0.383, 9.848] *	7.430 [1.667, 12.897] *	17.068 [9.004, 24.542] *
Specificity	69.712 [66.577, 72.804]	81.275 [77.301, 84.937]	85.989 [83.356, 88.391]	90.972 [88.108, 93.935]	95.320 [94.054, 96.580]	96.471 [94.580, 98.150]	11.564 [7.761, 15.114] *	4.983 [2.147, 7.987] *	1.151 [-0.405, 2.605]
PPV	0.396 [0.334, 0.458]	0.529 [0.449, 0.608]	0.567 [0.483, 0.640]	0.688 [0.606, 0.776]	0.747 [0.668, 0.814]	0.833 [0.750, 0.908]	0.133 [0.085, 0.180] *	0.122 [0.060, 0.193] *	0.086 [0.023, 0.153] *
NPV	0.969 [0.955, 0.981]	0.987 [0.973, 0.997]	0.958 [0.941, 0.974]	0.977 [0.960, 0.991]	0.920 [0.897, 0.943]	0.955 [0.935, 0.976]	0.018 [0.005, 0.032] *	0.019 [0.005, 0.033] *	0.035 [0.018, 0.053] *
Sensitivity for Stage IA cancers	91.057 [85.294, 95.946]	95.122 [88.095, 100.000]	83.333 [74.286, 91.204]	95.122 [88.095, 100.000]	56.098 [44.086, 68.229]	80.488 [68.293, 91.304]	4.065 [-1.786, 10.648]	11.789 [3.947, 20.325] *	24.390 [10.417, 37.153] *
Sensitivity for Stage IB cancers	100.000 [100.000, 100.000]	100.000 [100.000, 100.000]	100.000 [100.000, 100.000]	100.000 [100.000, 100.000]	87.500 [50.000, 100.000]	100.000 [100.000, 100.000]	0.000 [0.000, 0.000] *	0.000 [0.000, 0.000] *	12.500 [0.000, 50.000] *
Sensitivity for Stage IIA cancers	83.333 [0.000, 100.000]	100.000 [0.000, 100.000]	83.333 [0.000, 100.000]	100.000 [0.000, 100.000]	61.111 [0.000, 83.333]	100.000 [0.000, 100.000]	16.667 [0.000, 33.333] *	16.667 [0.000, 33.333] *	38.889 [0.000, 50.000] *
Sensitivity for Stage IIB cancers	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000] *	0.000 [0.000, 0.000] *	0.000 [0.000, 0.000] *
Sensitivity for Stage IIIA cancers	94.444 [86.667, 100.000]	100.000 [100.000, 100.000]	91.667 [77.778, 100.000]	66.667 [20.000, 100.000]	72.222 [26.667, 100.000]	50.000 [0.000, 100.000]	5.556 [0.000, 13.333] *	-25.000 [-60.000, 0.000]	-22.222 [-66.667, 0.000]
Sensitivity for Stage IIIB cancers	86.111 [58.333, 100.000]	83.333 [42.857, 100.000]	86.111 [58.333, 100.000]	83.333 [42.857, 100.000]	75.000 [33.333, 100.000]	83.333 [42.857, 100.000]	-2.778 [-16.667, 8.333]	-2.778 [-16.667, 8.333]	8.333 [-16.667, 41.667]
Sensitivity for Stage IV cancers	90.000 [77.778, 98.485]	90.000 [66.667, 100.000]	78.333 [52.778, 98.333]	90.000 [66.667, 100.000]	65.000 [38.889, 87.879]	80.000 [50.000, 100.000]	0.000 [-22.917, 18.750]	11.667 [0.000, 31.667] *	15.000 [3.704, 27.778] *
Sensitivity For 0 to 6 mm	66.667 [0.000, 66.667]	50.000 [0.000, 100.000]	25.000 [0.000, 50.000]	50.000 [0.000, 100.000]	0.000 [0.000, 0.000]	50.000 [0.000, 100.000]	-16.667 [-66.667, 33.333]	25.000 [0.000, 50.000] *	50.000 [0.000, 100.000] *
Sensitivity For 6 to 8 mm	75.000 [61.111, 86.667]	75.000 [33.333, 100.000]	37.500 [16.667, 58.333]	50.000 [14.286, 100.000]	6.250 [0.000, 15.385]	12.500 [0.000, 40.000]	0.000 [-36.111, 29.167]	12.500 [-30.556, 54.762]	6.250 [-6.250, 25.000]
Sensitivity For 8 to 15 mm	87.500 [78.495, 95.238]	95.833 [85.714, 100.000]	79.861 [68.116, 90.278]	91.667 [78.261, 100.000]	36.806 [26.852, 47.368]	79.167 [60.870, 94.444]	8.333 [0.833, 17.308] *	11.806 [-2.381, 25.694]	42.361 [24.667, 57.937] *
Sensitivity For 15 to 25 mm	92.647 [86.420, 96.795]	100.000 [100.000, 100.000]	92.157 [84.946, 96.795]	97.059 [90.476, 100.000]	82.353 [74.731, 88.889]	91.176 [80.000, 100.000]	7.353 [3.205, 13.580] *	4.902 [2.299, 8.025] *	8.824 [-2.381, 18.860]
Specificity For 0 to 6 mm	82.418 [79.991, 84.765]	92.523 [88.789, 95.556]	95.144 [94.007, 96.253]	98.338 [96.523, 99.612]	98.956 [98.404, 99.440]	100.000 [100.000, 100.000]	10.105 [6.041, 13.964] *	3.195 [1.021, 4.999] *	1.044 [0.560, 1.596] *
Specificity For 6 to 8 mm	54.916 [47.869, 62.876]	71.373 [60.169, 82.487]	84.437 [80.530, 88.530]	86.150 [77.787, 93.974]	95.493 [93.341, 97.522]	95.237 [89.863, 99.684]	16.457 [7.030, 25.341] *	1.714 [-5.946, 8.638]	-0.256 [-6.032, 4.212]
Specificity For 8 to 15 mm	37.498 [29.478, 44.934]	53.230 [39.816, 66.715]	55.661 [46.678, 64.879]	75.140 [62.762, 86.003]	87.642 [83.164, 91.686]	93.460 [85.849, 99.284]	15.732 [0.038, 31.094] *	19.480 [3.225, 34.562] *	5.818 [-2.599, 12.887]
Specificity For 15 to 25 mm	12.644 [1.442, 29.890]	11.635 [0.000, 38.740]	26.651 [3.903, 53.065]	24.135 [0.000, 55.832]	48.846 [26.701, 72.000]	35.769 [1.366, 72.603]	-1.010 [-10.266, 10.342]	-2.516 [-10.064, 4.762]	-13.077 [-36.546, 8.517]
Sensitivity for non_solid	77.778 [62.821, 92.708]	93.333 [77.778, 100.000]	66.667 [48.148, 85.185]	86.667 [66.667, 100.000]	41.111 [18.519, 63.889]	60.000 [33.333, 83.333]	15.556 [-1.667, 33.333]	20.000 [0.000, 39.394] *	18.889 [2.083, 38.095] *
Sensitivity for part_solid	98.148 [93.750, 100.000]	100.000 [100.000, 100.000]	83.333 [61.111, 100.000]	100.000 [100.000, 100.000]	57.407 [30.952, 80.556]	88.889 [66.667, 100.000]	1.852 [0.000, 6.250] *	16.667 [0.000, 38.889] *	31.481 [5.000, 60.417] *
Sensitivity for poorly_defined	83.333 [71.429, 93.750]	95.455 [85.000, 100.000]	72.727 [56.863, 87.302]	86.364 [71.429, 100.000]	52.273 [33.333, 70.588]	68.182 [46.667, 86.957]	12.121 [0.926, 24.638] *	13.636 [-2.564, 30.000]	15.909 [2.381, 30.208] *
Sensitivity for smooth	91.111 [86.508, 95.402]	96.667 [89.655, 100.000]	81.667 [70.667, 91.667]	93.333 [82.857, 100.000]	56.111 [42.708, 71.212]	80.000 [64.000, 93.750]	5.556 [-1.282, 11.250]	11.667 [4.444, 20.312] *	23.889 [9.333, 38.739] *
Sensitivity for solid	94.180 [91.667, 96.491]	98.413 [94.915, 100.000]	91.005 [85.849, 95.278]	95.238 [89.394, 100.000]	71.164 [62.568, 79.792]	88.889 [80.882, 96.296]	4.233 [0.926, 7.092] *	4.233 [0.483, 8.209] *	17.725 [8.611, 27.083] *
Sensitivity for spiculated	97.083 [95.000, 98.837]	100.000 [100.000, 100.000]	97.083 [95.000, 98.837]	100.000 [100.000, 100.000]	80.000 [72.500, 86.842]	95.000 [87.179, 100.000]	2.917 [1.163, 5.000] *	2.917 [1.163, 5.000] *	15.000 [4.023, 25.269] *
Specificity for non_solid	70.422 [58.922, 80.179]	58.905 [41.226, 75.075]	83.292 [72.161, 91.486]	75.715 [59.568, 89.465]	96.948 [94.525, 99.010]	89.414 [77.718, 99.545]	-11.517 [-28.456, 4.105]	-7.576 [-22.676, 6.829]	-7.535 [-19.499, 2.833]
Specificity for part_solid	46.899 [23.236, 70.301]	63.454 [33.942, 91.718]	78.737 [59.839, 93.900]	81.394 [55.018, 100.000]	92.525 [85.051, 98.620]	91.030 [70.655, 100.000]	16.555 [0.152, 35.154] *	2.656 [-15.163, 16.185]	-1.495 [-16.667, 8.272]
Specificity for poorly_defined	63.427 [52.550, 73.748]	57.004 [41.872, 71.702]	79.005 [69.299, 86.821]	75.280 [61.010, 87.457]	94.253 [88.838, 98.005]	91.380 [82.534, 99.000]	-6.423 [-19.396, 5.927]	-3.726 [-15.912, 6.873]	-2.873 [-10.968, 3.391]
Specificity for smooth	57.276 [51.648, 62.939]	77.535 [70.953, 84.563]	80.741 [75.625, 85.678]	90.202 [84.805, 94.675]	94.618 [91.822, 97.041]	93.486 [89.521, 97.128]	20.259 [13.614, 27.221] *	9.461 [4.289, 14.064] *	-1.131 [-4.580, 1.739]
Specificity for solid	57.114 [51.362, 62.345]	75.436 [68.256, 82.048]	79.396 [74.146, 84.229]	87.625 [81.934, 92.493]	93.278 [90.155, 95.913]	92.984 [88.756, 96.509]	18.322 [11.831, 25.134] *	8.229 [3.672, 12.669] *	-0.294 [-3.321, 2.646]
Specificity for spiculated	43.148 [23.841, 62.712]	47.575 [23.480, 71.003]	55.145 [34.597, 73.457]	49.515 [25.635, 73.073]	71.291 [53.891, 86.793]	64.009 [41.444, 86.092]	4.427 [-7.563, 17.385]	-5.630 [-22.090, 10.027]	-7.283 [-21.555, 4.917]

Supplementary Table [5]: Subset analysis on reader study on a single CT volume (without using priors).



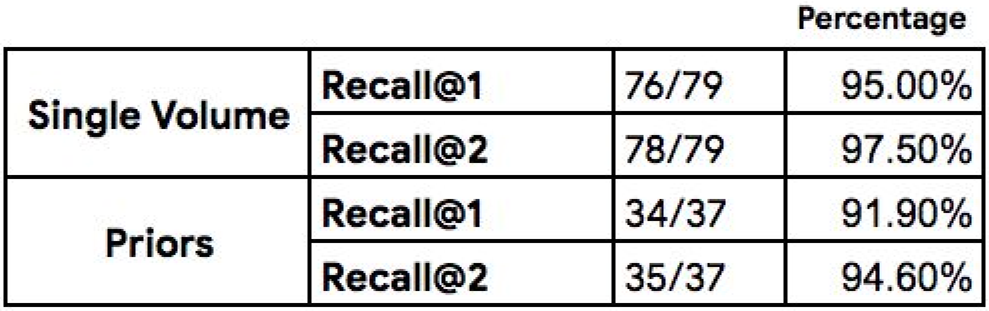
	Average Reader Lung-RADS 3+	LUMAS 3+	Average Reader Lung-RADS 4A+	LUMAS 4a+	Average Reader Lung-RADS 4B/X	LUMAS 4b/x	LUMAS 3+ - Average Reader Lung-RADS 3+	LUMAS 4a+ - Average Reader Lung-RADS 4A+	LUMAS 4b/x - Average Reader Lung-RADS 4B/X
Sensitivity	86.667 [79.710, 92.857]	87.500 [76.471, 97.222]	82.083 [74.123, 89.394]	82.500 [69.048, 93.939]	70.000 [59.402, 80.303]	72.500 [58.824, 85.714]	0.833 [-9.804, 11.765]	0.417 [-10.256, 10.294]	2.500 [-6.818, 12.281]
Specificity	83.722 [80.745, 86.763]	84.208 [79.622, 88.608]	89.133 [86.623, 91.590]	92.750 [89.599, 95.795]	94.427 [92.581, 96.146]	96.508 [94.273, 98.525]	0.486 [-3.668, 4.593]	3.616 [0.729, 6.582] *	2.081 [-0.009, 4.131]
PPV	0.462 [0.360, 0.556]	0.471 [0.357, 0.593]	0.549 [0.437, 0.647]	0.647 [0.519, 0.779]	0.669 [0.549, 0.766]	0.770 [0.632, 0.893]	0.010 [-0.060, 0.085]	0.098 [0.012, 0.207] *	0.101 [-0.000, 0.217]
NPV	0.975 [0.959, 0.988]	0.977 [0.955, 0.995]	0.969 [0.951, 0.984]	0.971 [0.947, 0.992]	0.951 [0.929, 0.972]	0.956 [0.929, 0.980]	0.002 [-0.017, 0.021]	0.002 [-0.014, 0.018]	0.005 [-0.009, 0.020]
Sensitivity for Stage IA cancers	90.351 [79.861, 98.413]	89.474 [73.684, 100.000]	85.088 [73.810, 95.238]	89.474 [73.684, 100.000]	73.684 [57.692, 87.681]	78.947 [57.895, 95.455]	-0.877 [-15.909, 12.963]	4.386 [-8.333, 16.667]	5.263 [-6.944, 18.421]
Sensitivity for Stage IB cancers	83.333 [0.000, 83.333]	100.000 [0.000, 100.000]	83.333 [0.000, 83.333]	100.000 [0.000, 100.000]	83.333 [0.000, 83.333]	100.000 [0.000, 100.000]	16.667 [0.000, 16.667] *	16.667 [0.000, 16.667] *	16.667 [0.000, 16.667] *
Sensitivity for Stage IIA cancers	75.000 [0.000, 83.333]	100.000 [0.000, 100.000]	75.000 [0.000, 83.333]	100.000 [0.000, 100.000]	75.000 [0.000, 83.333]	100.000 [0.000, 100.000]	25.000 [0.000, 33.333] *	25.000 [0.000, 33.333] *	25.000 [0.000, 33.333] *
Sensitivity for Stage IIB cancers	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000] *	0.000 [0.000, 0.000] *	0.000 [0.000, 0.000] *
Sensitivity for Stage IIIA cancers	88.889 [0.000, 100.000]	66.667 [0.000, 100.000]	88.889 [0.000, 100.000]	33.333 [0.000, 100.000]	61.111 [0.000, 100.000]	33.333 [0.000, 100.000]	-22.222 [-83.333, 16.667]	-55.556 [-83.333, 0.000]	-27.778 [-50.000, 0.000]
Sensitivity for Stage IIIB cancers	79.167 [33.333, 100.000]	75.000 [0.000, 100.000]	79.167 [33.333, 100.000]	75.000 [0.000, 100.000]	75.000 [33.333, 100.000]	75.000 [0.000, 100.000]	-4.167 [-33.333, 16.667]	-4.167 [-33.333, 16.667]	0.000 [-33.333, 33.333]
Sensitivity for Stage IV cancers	90.476 [74.074, 100.000]	85.714 [50.000, 100.000]	80.952 [58.333, 100.000]	85.714 [50.000, 100.000]	66.667 [33.333, 92.857]	71.429 [33.333, 100.000]	-4.762 [-40.476, 25.000]	4.762 [-18.182, 33.333]	4.762 [-30.952, 41.667]
Sensitivity For 0 to 6 mm	66.667 [0.000, 100.000]	50.000 [0.000, 100.000]	41.667 [0.000, 50.000]	50.000 [0.000, 100.000]	16.667 [0.000, 33.333]	50.000 [0.000, 100.000]	-16.667 [-100.000, 66.667]	8.333 [-50.000, 66.667]	33.333 [0.000, 66.667] *
Sensitivity For 6 to 8 mm	73.810 [53.704, 91.667]	57.143 [20.000, 100.000]	64.286 [41.667, 83.333]	42.857 [0.000, 85.714]	28.571 [8.333, 50.000]	14.286 [0.000, 50.000]	-16.667 [-60.417, 27.778]	-21.429 [-50.000, 10.000]	-14.286 [-33.333, 8.333]
Sensitivity For 8 to 15 mm	78.788 [62.500, 93.333]	90.909 [69.231, 100.000]	72.727 [56.667, 88.889]	81.818 [57.143, 100.000]	59.091 [44.444, 73.810]	72.727 [45.455, 100.000]	12.121 [-1.515, 26.190]	9.091 [-16.667, 33.333]	13.636 [-8.333, 38.333]
Sensitivity For 15 to 25 mm	97.619 [94.444, 100.000]	100.000 [100.000, 100.000]	97.619 [94.444, 100.000]	100.000 [100.000, 100.000]	96.429 [92.308, 100.000]	92.857 [75.000, 100.000]	2.381 [0.000, 5.556] *	2.381 [0.000, 5.556] *	-3.571 [-19.444, 5.128]
Specificity For 0 to 6 mm	92.217 [90.011, 94.240]	94.422 [90.573, 97.635]	95.833 [94.482, 97.191]	98.760 [96.770, 100.000]	99.047 [98.418, 99.611]	100.000 [100.000, 100.000]	2.204 [-2.372, 6.246]	2.927 [0.428, 4.935] *	0.953 [0.389, 1.582] *
Specificity For 6 to 8 mm	73.851 [64.677, 82.719]	79.665 [65.970, 91.191]	83.559 [75.749, 90.879]	84.888 [73.070, 95.128]	92.903 [88.841, 96.419]	92.305 [82.736, 100.000]	5.815 [-5.518, 16.074]	1.329 [-9.842, 11.249]	-0.598 [-10.438, 6.843]
Specificity For 8 to 15 mm	66.879 [57.119, 74.616]	52.858 [35.291, 69.581]	73.947 [64.383, 81.708]	80.460 [66.534, 92.056]	83.994 [76.417, 89.868]	90.572 [80.085, 99.597]	-14.021 [-30.475, 2.231]	6.513 [-6.985, 19.747]	6.578 [-3.228, 16.389]
Specificity For 15 to 25 mm	26.836 [4.214, 52.090]	24.671 [0.000, 56.237]	42.625 [15.149, 67.902]	59.701 [22.734, 93.243]	52.795 [29.584, 76.025]	71.378 [37.218, 96.563]	-2.166 [-35.064, 24.352]	17.076 [-2.816, 38.530]	18.583 [-5.770, 41.756]
Sensitivity for non_solid	80.000 [58.333, 100.000]	80.000 [33.333, 100.000]	80.000 [58.333, 100.000]	80.000 [33.333, 100.000]	63.333 [33.333, 88.889]	60.000 [0.000, 100.000]	0.000 [-27.778, 25.000]	0.000 [-27.778, 25.000]	-3.333 [-33.333, 25.000]
Sensitivity for part_solid	83.333 [0.000, 100.000]	100.000 [0.000, 100.000]	66.667 [0.000, 83.333]	100.000 [0.000, 100.000]	61.111 [0.000, 83.333]	66.667 [0.000, 100.000]	16.667 [0.000, 33.333] *	33.333 [0.000, 50.000] *	5.556 [-50.000, 50.000]
Sensitivity for poorly_defined	85.417 [71.667, 96.667]	75.000 [40.000, 100.000]	81.250 [68.519, 93.750]	75.000 [40.000, 100.000]	75.000 [59.091, 91.667]	62.500 [28.571, 100.000]	-10.417 [-41.667, 13.333]	-6.250 [-38.889, 19.048]	-12.500 [-33.333, 10.000]
Sensitivity for smooth	82.456 [72.222, 91.270]	94.737 [82.353, 100.000]	78.070 [66.667, 88.889]	89.474 [73.684, 100.000]	63.158 [47.619, 78.070]	78.947 [60.000, 95.000]	12.281 [-5.208, 27.083]	11.404 [-0.725, 23.333]	15.789 [6.667, 26.190] *
Sensitivity for solid	87.500 [80.435, 93.827]	93.750 [84.000, 100.000]	84.375 [75.758, 92.130]	90.625 [79.310, 100.000]	75.521 [64.444, 85.714]	84.375 [70.968, 96.154]	6.250 [-5.128, 17.857]	6.250 [-3.922, 15.686]	8.854 [-1.190, 19.048]
Sensitivity for spiculated	94.444 [84.848, 100.000]	100.000 [100.000, 100.000]	92.222 [79.412, 100.000]	100.000 [100.000, 100.000]	87.778 [72.222, 97.917]	93.333 [76.923, 100.000]	5.556 [0.000, 15.152] *	7.778 [0.000, 20.588] *	5.556 [-12.500, 23.611]
Specificity for non_solid	75.935 [62.791, 86.133]	60.470 [36.373, 83.320]	85.318 [73.731, 93.625]	77.238 [57.700, 94.768]	93.007 [86.198, 98.179]	82.624 [64.473, 98.351]	-15.465 [-37.316, 4.730]	-8.080 [-28.989, 10.168]	-10.382 [-26.056, 3.110]
Specificity for part_solid	66.667 [27.778, 99.086]	71.882 [33.333, 100.000]	71.353 [36.712, 99.086]	85.941 [51.371, 100.000]	78.647 [51.353, 99.543]	85.941 [51.371, 100.000]	5.215 [0.000, 16.667] *	14.588 [0.000, 36.991] *	7.294 [-5.188, 25.000]
Specificity for poorly_defined	72.803 [59.118, 84.209]	64.475 [44.084, 84.114]	81.526 [69.109, 91.643]	85.967 [71.358, 98.873]	90.566 [82.818, 96.832]	90.645 [78.882, 99.584]	-8.328 [-23.235, 5.412]	4.441 [-11.860, 19.586]	0.079 [-10.306, 10.097]
Specificity for smooth	80.353 [74.291, 85.468]	81.697 [73.414, 89.112]	86.835 [81.878, 90.879]	91.249 [85.324, 96.076]	93.415 [89.871, 96.155]	93.320 [87.949, 97.630]	1.344 [-5.590, 7.391]	4.414 [-0.152, 9.263]	-0.095 [-3.759, 3.316]
Specificity for solid	78.623 [72.631, 84.167]	80.013 [72.297, 87.963]	85.041 [79.928, 89.675]	91.223 [85.445, 96.099]	91.628 [87.915, 94.680]	94.557 [89.684, 98.539]	1.390 [-5.148, 7.412]	6.182 [1.138, 11.135] *	2.928 [-1.298, 7.114]
Specificity for spiculated	54.216 [27.657, 75.748]	53.303 [22.246, 83.990]	59.209 [34.033, 79.484]	53.303 [22.246, 83.990]	68.468 [45.075, 87.422]	76.652 [48.012, 98.068]	-0.913 [-26.014, 23.362]	-5.906 [-31.507, 19.116]	8.183 [-17.408, 32.148]

Supplementary Table [6]: Subset analysis on reader study using prior CT volumes.



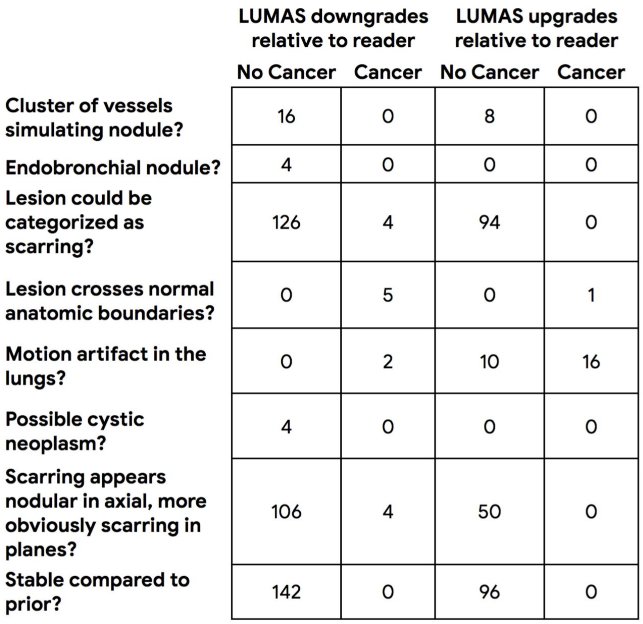
	Non-priors	Priors
LungRADS 1 vs. 2 vs. 3 vs. 4A vs. 4B bs. 4X raw score disagreement	0.49	0.44
LungRADS 1/2 vs. 3 vs. 4A vs. 4B/X management disagreement	0.30	0.21
Severe management disagreement (1/2 vs 4a/b/x or 3 vs. 4b/x)	0.05	0.06

Supplementary Table [7]: Reader disagreements. Numbers shown are fraction of cases with disagreements. We analyzed three different disagreement types: raw Lung-RADS score disagreement, management level disagreement (which groups Lung-RADS 1 and 2 as done for all other analyses presented) and large management disagreements where the disagreements were not in adjacent risk buckets (i.e. one reader reports Lung-RADS 2 and the other reports Lung-RADS 4A). This includes data from both reader studies (with priors and without priors).



			Percentage
Single Volume	Recall@1	76/79	95.00%
	Recall@2	78/79	97.50%
Priors	Recall@1	34/37	91.90%
	Recall@2	35/37	94.60%

Supplementary Table [8]: **Recall values on all cancer cases labeled with bounding boxes within the test dataset**. The numerator and denominator refer to the number of found vs total malignant nodules in both single volume cases and those with priors. The @1 and @2 suffixes refer to the top single detection and top two detections surfaced by the detection model, respectively. The corresponding HIT values shown in Figures 2 and 3 focus on the subset of correctly classified cancer cases. Since the HIT@2 metric achieved a 100% hit rate in both baseline and with priors cases, the missed detections may have impacted the classification for these two cancer cases.



	LUMAS downgrades relative to reader		LUMAS upgrades relative to reader	
	No Cancer	Cancer	No Cancer	Cancer
Cluster of vessels simulating nodule?	16	0	8	0
Endobronchial nodule?	4	0	0	0
Lesion could be categorized as scarring?	126	4	94	0
Lesion crosses normal anatomic boundaries?	0	5	0	1
Motion artifact in the lungs?	0	2	10	16
Possible cystic neoplasm?	4	0	0	0
Scarring appears nodular in axial, more obviously scarring in planes?	106	4	50	0
Stable compared to prior?	142	0	96	0

Supplementary Table [9]: Summary of the differences between the model and the consensus of the readers. Each disagreement with a reader in cases where the model disagreed with the consensus of the readers appears once for every time one of the questions in the leftmost column is true.

Kernel Selection

Each case often had multiple reconstruction kernels available. When running the model on a case we selected harder kernels commonly used in lung imaging (see list below). In the reader study, we used the same volumes that were chosen for the model. Within each case we chose the highest ranked kernel according to the following lists. There were no cases with more than one manufacturer.

- Siemens
  - 1. *B50f*, 2. *B45f*, 3. *B50s*, 4. *B40f*, 5. *B41s*, 6. *B60f*, 7. *B60s*, 8. *B70f*, 9. *B36f*, 10. *B35f*, 11. *B30f* 12. *B31s*
- GE
  - 1. *LUNG,* 2. *BONE,* 3. *BODY FILTER/BONE,* 4. *STANDARD,* 5. *BODY FILTER/STANDARD,* 6. *SOFT,* 7. *EXPERIMENTAL7,* 8. *BODY FILTER/EXPERIMENTAL7*

- Philips
    - 1. *D*, 2. *C*, 3. *B*, 4. *A*
- Toshiba
    - 1. *FC51*, 2. *FC50*, 3. *FC52*, 4. *FC53*, 5. *FC30*, 6. *FC11*, 7. *FC10*, 8. *FC82*, 9. *FL04*, 10. *FC02*, 11. *FC01*, 12. *FL01*

Additional Modeling Details

At a high level, our model begins with lung segmentation, followed by detection, and ending with classification, an approach that has been described in past research. However, for each of these components, we upgraded the specific techniques to state of art approaches (at the time of publication) for the general computer vision tasks: MaskRCNN[44] for instance segmentation, Retinanet[47] for object detection, and I3D[47,49] for action recognition from video (also a volume classification task).

Different lung segmentation approaches vary in terms of quality and computational cost. In our case, the approach used was solely to determine a center point of the lung segmentation bounding box and therefore precise lung boundaries were not a critical factor in the final results. One advantage of the MaskRCNN approach is that the segmentation is performed on two-dimensional (2D) slices and is independent of slice spacing.

The cancer ROI detection model was trained on LIDC first, and then we collected additional labels on NLST to fine tune the model to only detect malignant nodules instead of all nodules.

For classification, we found on our tune set that I3D alone performed well when predicting cancer directly. We then sought to combine this full volume approach with our two-stage approach (see Methods - *Model Development and Training*). We used I3D as the base feature extractor for classification tasks after determining it outperformed several other feature extractors on our tune set. We used the spatial resolutions shown in Extended Data Figure 10, which were the highest resolutions allowed by commercial hardware. For the cancer ROI detection and cancer risk prediction model, we were able to train on subvolumes smaller than the whole volume, which allowed us to use 1.4 mm x 0.7 mm^2 resolution images. While this may introduce additional “partial voluming” effects seen in clinical radiology, what the algorithm “sees” is generally quite different compared to human perception, and the training and evaluation was performed solely on these resampled volumes. For the full volume model we used a resampled 1.5 mm^3 resolution. Detecting cancer alone for the full volume model is difficult due to the wide range of appearance and locations of nodules; therefore, the model was also trained to detect the presence of nodules. In order to assess the contribution of the full volume model, we retrospectively computed an AUC of 89.0% on the test set. Additionally, the subjective analysis showed evidence that the model focused on nodules (see Supplementary - “*Subjective Analysis* and *Review of Results”*). These results demonstrate that the full volume model effectively collected features relevant for cancer detection.

Subjective Analysis

We analyzed subsets in which LUMAS differed from the majority vote of our six readers. We first examined disagreements between our model and Retrospective-Lung-RADS on the tune set with three radiologists to generate hypotheses to pursue on the test cases. They generated nine hypotheses framed as questions with categorical answers. We then labeled all cases in the without prior reader study where the LUMAS bucket disagreed with the consensus reader bucket. Upon labeling disagreements, the most commonly present hypotheses were “Lesion could be categorized as scarring?,” “Stable compared to prior?” (only for cases with priors), “Scarring appears nodular in axial, more obviously scarring in orthogonal planes?” The full results of this analysis are shown in Supplementary Table 9, where each disagreement with a reader shows up once in the table for every time one of the hypotheses was labeled as true.

Additionally, to better characterize and analyze model behavior, attribution regions for 12 cases in the tune set were examined through focused questions. These regions were computed using integrated gradients to show positive and negative classification influences[53]. A series of questions concerning the model’s region of focus for the global and local views were given. All readers unanimously agreed that both positive and negative attributions focused on the nodules in all cancer positive cases. In 40% of the negative cases, the readers noted that parenchymal vasculature was highlighted. In 86% of the cancer positive cases, the readers noted that the full volume model focused on the same nodule as the two-stage model. Finally, in characterizing the region on the nodule examined, the strongest agreement was that for 4 of 7 of the cancer positive cases the readers agreed that the negative attributions were examining the edges of the nodule. Extended Data Figures 6a and 6b give examples of these cases.

Review of Results

The final manuscript draft was evaluated using the Radiomics Quality Score system (Radiomics, Maastricht, Netherlands) prior to submission receiving a score of 92%.

Subset Analysis

We computed the sensitivity and specificity of the model's risk buckets and the average readers risk buckets on subsets based on nodule properties, lung cancer staging, and nodule size. This information was collected in the NLST trial. For some subsets, such as cancer staging, there were only cancer positive examples in the subset and therefore we only computed sensitivity. Full results are shown in Supplementary Tables 5 and 6.