Supplementary Online Content

Mak RH, Endres MG, Paik JH, et al. Use of crowd innovation to develop an artificial intelligence–based solution for radiation therapy targeting [published online April 18, 2019]. *JAMA Oncol*. doi:10.1001/jamaoncol.2019.0159

eTable 1. Contest Prize Distributions, with payouts based on ranked performance

eTable 2. Common evaluation metrics for comparing segmentations

eTable 3. Algorithm, ensemble, and benchmark performance averaged over all scans in the validation, holdout and external data set

eTable 4. Primary methods, and utilization of supplemental data for training and/or inference

eTable 5. Algorithm, ensemble, and benchmark performance averaged over all scans in the validation and holdout data sets

eTable 6. Algorithm, ensemble, and benchmark performance averaged over all scans in the external data set

eFigure 1. Additional Examples of Human Expert versus Automated Segmentations

eFigure 2. Lung Tumor Volume Distribution in This Study Versus A Publicly-Available Dataset

eFigure 3. Distribution of Segmentation Score (S-score) as a function of V/V0

eFigure 4. Evolution of Winning AI Algorithms During the Multi-Phase Contest

eFigure 5. Performance of Ensemble AI Algorithms in Clinical Sub-Groups

eFigure 6. Performance of Top Segmentation Algorithms from The Contest on an External Dataset

eReferences.

eAcknowledgements.

This supplementary material has been provided by the authors to give readers additional information about their work.

A. ONLINE-ONLY METHODS:

Patient/CT Imaging and Annotation Data Sets

The study was conducted under an IRB-approved retrospective protocol with a waiver of consent (Dana-Farber/Harvard Cancer Center protocol 11-286). The data set consisted of fully anonymized computerized tomography (CT) scans (512x512 pixel, 16-bit grayscale image slices) which were clinically utilized for radiation treatment planning in 461 patients diagnosed with non-small cell lung cancer (NSCLC) from 2001 to 2014. All CT scans were taken using GE Lightspeed QX/i (4%), Lightspeed RT (36%) and Lightspeed RT16 (60%) equipment with slice thicknesses 2.5mm (96%), 3.75(2%) and 5mm (2%). The scans comprised 77,942 images in total; the median number of image slices per scan was 157 (range: 36-371).

The primary tumor segmentations were re-drawn by a single radiation oncologist with expertise in lung cancer treatment (R.H.M), and incorporated both analysis of CT image and the following available clinical data: 1) other image modalities such as 18F-fluorodeoxyglucose positron emission tomography (PET), and 2) clinical reports including staging procedures and pathology. Approximately 10% of the scan slices (8144 images) had tumor present. The remaining regions of interest (ROIs) corresponding to other normal structures drawn at the time of radiation treatment planning (e.g. body, bronchi, esophagus, heart, lung, trachea; each were available for subsets of the scans) were produced using a combination of manual and/or automated segmentation tools.

Training, Validation and Holdout Test Data Sets

The curated data set was randomly divided into three disjoint subsets: a training set comprising 229 patients, a validation set comprising 96 patients and a holdout test set comprising 136 patients. The

contestants were provided with the anonymized training set, which included 229 CT scans and associated primary tumor segmentation data, the additional segmentation data for ROIs corresponding to other normal structures, DICOM metadata (including pixel spacing, slice thickness and patient orientation, CT scanner type and settings), and supplementary clinical information indicating whether or not intravenous contrast agent was used for the scan. They had access to CT imaging and annotation data, and associated DICOM metadata in both the original anonymized DICOM data format and more familiar data formats (lossless PNGs, and field separated text files). In addition, the contestants were provided the 96 validation set CT scans and accompanying supplementary clinical information, but not the associated primary tumor segmentation, nor the additional segmentation data for normal structure ROIs**.** The holdout test data set was withheld from the contestants during all phases.

In phase 2 and phase 3, contestants were provided boundary-avoiding seed points within each tumor ROI as additional training and validation data (seed points were withheld for the 136 patients in the holdout test set). The seed points were generated randomly based on the existing primary tumor segmentation data.

Evaluation Metric for Scoring Segmentations Generated by Contestants' Algorithms

The contestant's algorithms were scored by comparing the volumetric segmentation produced by each algorithm on a given patient's scan (including all CT slices/images in the scan) against the expert's segmentation on that same scan. A wide variety of metrics are available for evaluating the volumetric segmentation quality with respect to reference data produced either manually (e.g. radiologist or radiation oncologist) or automatically by a gold standard algorithm. Commonly used metrics are provided for reference in eTable 2, along with the metric developed for this study (S-score). Metrics such as the Dice Coefficient (i.e., F_1 score), and Jaccard Index (J) are insensitive to the absolute error in tumor volume segmentation (i.e., these metrics are invariant

under scaling of the ground truth and predicted tumor volumes). From a clinical standpoint, however, the absolute error in segmentation is particularly relevant for radiation treatment of the patient, since leaving a portion of the tumor out of the segmentation will result in under-dosing of the tumor and potentially lead to the bad outcome of local tumor recurrence.

Given a truth ROI (**T**) which is the manually generated volumetric segmentation by the human expert, and a predicted ROI (**P**) automatically generated by the contestant's algorithm, the true positive (TP), false positive (FP) and false negative (FN) region volumes are defined by TP = $|\mathbf{T} \cap \mathbf{P}|$, FP = $|\mathbf{P} \setminus \mathbf{T}|$ and FN = $|\mathbf{T} \setminus \mathbf{P}|$ respectively. The total tumor volume for a given scan is V = $|\mathbf{T}|$. For this work, an evaluation metric S was defined in terms of a non-negative error volume E = V FN / TP + FP, which vanishes when **T** = **P** and diverges when $\mathbf{T} \cap \mathbf{P} = \emptyset$. An additional scale parameter $V_0 = (4\pi/3)\ (30\text{mm})^3$ is introduced in order to explicitly break scale invariance of the metric (which must be a dimensionless function of TP, FP, FN and scale breaking parameter V_0), allowing for greater penalty for larger tumors compared to smaller tumors, for a fixed relative error. The dependence of the S-score on V/V_0 for various choices of E/V are shown in eFigure 3, and compared with the per-scan tumor volume cumulative distribution function.

Contest Design and Execution:

The contest was hosted on the Topcoder.com (Wipro, Bengalaru, India) online algorithm contest platform that has a community of ~1,000,000 programmers and data scientists competing regularly. The competitors were tasked with producing an automatic tumor delineation algorithm that parallels the lung tumor delineation accuracy of an "expert clinician", while exceeding the expert in processing speed and delineation consistency. The contest offered a prize pool of $50,000 for the first two phases of competition, which were open to the entire Topcoder community. Top ranking contestants from Phase 1 received prizes of: $10,000, $7,000, $5,000, $4,000, $3,000, $2,000, and $1,000, plus $3,000 in mid-

contest bonus prizes and top-ranking participants of Phase 2 of the contest received prizes of: $7,000, $4,500, $2,000, $1,000 and $500 (see eTable 1). For these first two phases, the rankings were determined by the highest average S-score based on the segmentations generated by the independent application of the submitted algorithms on the holdout test set by the study team. A third invitational phase, which was restricted to the five top contestants from phase 2 and an additional participant invited from a separate contest, focused on refinement of the top algorithms from phase 2. In that phase, each contestant was awarded $500 for submitting proposals which outlined a strategy for further algorithmic improvement, and each were awarded $100 for submitting a solution, plus $15 for every 0.001 improvement beyond the winning solution score of 0.574 from phase 2 (payouts were rounded up to the nearest multiple of $25). The first two phases of the contest each ran for three weeks (phase 1: 2/7/2017-3/1/2017; phase 2: 3/28/2017-4/18/2017), and the third phase ran for five weeks (7/27/2017-8/31/2017, including a one week hiatus following the first week), with a total elapsed contest time of seven months. Contestants were unaware of the phased nature of the contest, or the launch date of each subsequent phase.

At the beginning of the competition, participants were provided an introductory video explaining the background and concepts behind manual delineation of lung tumors (https://youtu.be/An-YDBjFDV8). During each phase, participants had access to the training (CT and segmentation data) and validation (CT data only) data sets (in phase 2 and phase 3, the data sets were expanded to include seed points randomly generated within the tumor region to indicate the location of each tumor as an additional input) and were permitted to submit predicted volumetric segmentations on the latter for online evaluation. Participants received real-time evaluation of their submissions, which were published on an online leaderboard, and could make multiple submissions during each contest phase with modifications in response to that feedback. To aid off-line evaluation of their algorithms during development, a proprietary visualization tool was created and provided to contestants to

automatically score their segmentations and make visual and quantitative comparisons by scan and slice against the segmentations produced by the human (such comparisons were only possible on the provided training data). The source codes from each phase were released to participants in each subsequent phase of competition.

At the end of each phase, participants were invited to submit their code and trained model parameters for final validation and evaluation by the study team on the independent holdout data set, which the contestants did not have access to (see eTable 5 and 6 for full performance details). Note that by providing an online leaderboard during competition and ability to query the validation data set multiple times, there is a possibility for over-fitting algorithms to the validation data set. Thus, a crucial part of reliably assessing the true performance of the algorithms submitted required final evaluation of the algorithms in each phase of the contest on the holdout dataset, which was independent and not accessible to the contestants. The study team including the medical expert reviewed performance both in terms of the performance metrics and on an individual scan and slice-by-slice basis to identify and classify sources of systematic error. These analyses between phases were utilized to revise the contest design and objectives in each subsequent phase, but the conclusions from these internal reviews were not explicitly disclosed to the contestants to minimize biasing solutions generated in subsequent phases of the contest.

During the third invitational phase, the 5 winning participants from phase 2 were invited to collaborate with each other and receive direct input, feedback and suggestions from the human expert and other study investigators. All communications were channeled through an open online forum to ensure that all participants had access to the same information and insights. Through the clinical expert review of the segmentations generated by the winning algorithms in phase 2, a pattern of performance deficiencies was observed including very small tumors and tumors adjacent to soft tissue structures and the study investigators believed that these deficiencies could be

addressed collaboratively. It was observed, for example, that the winning phase 2 solution failed to make predictions on about 8% of the scans in the validation data set (6% of the scans in the holdout data set) and that it did not take advantage of the full 16-bit grayscale information of the images (i.e., images were read and preprocessed as 8-bit), thus potentially omitting important textural features associated with the tumor regions. Thus, clear avenues for performance improvements existed and the contestants were given a goal to address deficits observed in the phase 2 solutions and improve the average S-score by ≥9% (the projected score of the top solution from phase 2 if no tumors were missed) and to focus their attention primarily on improving the algorithms previously submitted during phase 2. To facilitate improvement of the algorithms, each contestant was given access to compute resources on Amazon Web Services (AWS) upon request.

Ensemble Models:

Ensemble predictions were constructed by considering all 2^N-1 possible subsets of solutions (excluding the empty set) formed from the top (N=5) solutions from phase 2 and phase 3, respectively. For every such subset comprising M<=N solutions, M predictions were formed by considering the intersection of predicted tumor regions on each CT scan, for all possible nontrivial agreement rate thresholds (i.e., greater than or equal to p-fold agreement for p=1, …, M). In total, 80 non-trivial ensemble models were possible when considering the top five solutions from phase 2 and phase 3 of the contest. The best ensemble model for each phase was selected based on the model performance on the validation set using the S-score as the evaluation metric. Finally, the selected ensemble model was evaluated on the holdout data set.

Benchmarks: Intra and Inter-Observer Variation

To develop benchmarks of intra-observer variation, the clinical expert from this study (R.H.M.) performed the same segmentation task (blinded from prior segmentations) on a random subset of 42 scans (21 drawn from the validation set and 21 drawn from the holdout set) from the contest 3 months after the initial segmentations were performed. For quantification of inter-observer variation, the clinical expert (R.H.M.) segmented the lung tumors on CT scans of 21 patients (https://xnat.bmia.nl/app/template/XDATScreen_report_xnat_projectData.vm/search_element/xnat:projectData/search_field/xnat:projectData.ID/search_value/stwstrategymmd) from an external, previously published, publicly-available data set[1], which included lung tumor segmentations from five separate expert radiation oncologists who were not involved in this current study. Although the external data set was publicly available, based on a review of the submitted codes and associated solution descriptions provided by the contestants, the external data set was not used by any of the winning competitors for the purpose of training or evaluating solutions during the competition.

Benchmarks: Post-Contest External Validation

One deficiency of our challenge design is that the winning contestants received multiple, albeit limited exposure to the holdout data set during the contest (i.e., they received a final score for their algorithm's performance once after phase 1 and once after phase 2). Although extremely infrequent, and limited in terms of the information gained, this raised the prospect of overfitting on the holdout data set. The same concerns may be raised in relation to the information sharing between the study investigators and the contestants that is inherent to competition redesigns between phases. For our study, the likelihood of making the same decisions regarding competition redesign on the basis of algorithm performance in the validation set alone is high (i.e., the sources of under-performance in both phases were clear from analysis of the validation set), and thus the potential leakage of information is likely negligible. However, one cannot preclude with certainty the

possibility for overfitting on the hold-out data set through this process without more complex contest designs that better preserve the integrity of the holdout data set[2].

In light of these concerns, we validated the phase 3 performance results by evaluated all solutions subsequent to the contest on an independent external, publicly-available data set comprising CT scans from 21 patients with lung cancer (the same was used to produce our inter-observer benchmarks) which was segmented by the same human expert (R.H.M.) as the contest data set[1]. The average performance of the phase 3 algorithms on this external data set (eTable 6) is largely consistent (i.e., within 1-2 standard deviations) with that on holdout data set. The average performance across all algorithms appears to systematically exceed those obtained for the holdout data set, which may be indicative of small characteristic differences between the holdout and external data sets.

Using the same external data set, we investigated the performance of the contest algorithms, treating each observer's segmentations as the ground truth (eFigure 6). The results show that the algorithms generally perform better when compared to the ground truth segmentations produced by the expert in this study (R.H.M.), compared to those produced by the other five radiation oncologists. Although not conclusive, these findings suggest that the algorithms, trained on contours produced by the expert in this study (R.H.M.), may in fact have learned the contouring preferences or "style" of the radiation oncologist. These findings may be viewed in multiple ways. On the one hand, the inter-observer variability between radiation oncologists suggest that there is no consensus on what is ground truth, thus the algorithms produced by training on a single radiation oncologist's segmentations result in a potentially biased algorithm. On the other hand, the results suggest that the algorithms can offer a means for delivering the skill set and knowledge of a particular radiation oncologist to others. Finally, such algorithms may be regarded as a personalized tool, which conforms with the preferences of its user after an initial training on a user-defined training data set.

Benchmarks: Commercially-Available Software

A semi-automated segmentation tool from a commercially-available radiation therapy image viewing and segmentation software (MIM Maestro, MIM Software, Inc. Cleveland, OH) was selected as a representative existing clinical tool as a benchmark to compare the algorithms against. The specific tool (Region Grow) requires the end user to select a seed point and then utilizes tissue density in Hounsfield Units (HU) to grow the segmented volume. User controlled inputs include window & level, tendril diameter, and definition of a region to grow into, which was either manually optimized by the user (Commercial Software + Expert) or not manually optimized (Commercial software) to generate the two benchmarks.

Benchmarks: Efficiency Gains

The mean time required for an expert radiologist (R.H.M) to produce tumor contours was approximately eight minutes per scan, based on a random sampling of 21 scans each from the validation and holdout data sets, with the slowest cases taking as long as 23 minutes.

The top five phase 3 contest solutions were profiled on an Amazon Web Services p2.xlarge elastic compute cloud instance (Intel Xeon E5-2686 v4 processor and one NVIDIA Tesla K80 GPU). Timing measurements were obtained for each algorithm by running inference on the entire holdout data set using either a single CPU core (random forest-based algorithms) or a single GPU (CNN-based algorithms implemented with open source deep-learning frameworks) with NVIDIA CUDA 9 and cuDNN 5 GPU-accelerated deep learning libraries. The fastest algorithm produced segmentations at a mean rate of approximately 15 seconds per scan; the slowest produced contours at a mean rate of approximately two minutes per scan. The ensemble model constructed from all five algorithms required approximately five minutes per scan, based on the cumulative cost of all five algorithms.

Of note, during the contest the algorithms were only evaluated in terms of the quality of their predictions and not efficiency. This decision was motivated by the observation that efficiency gains achieved by an automated algorithm are only realized in practice if the segmentations produced require little subsequent manual modification by the radiation oncologist.

B. ONLINE-ONLY TABLES:

eTable 1: Contest Prize Distributions, with payouts based on ranked performance. Phase 1 included two mid-contest bonuses of $1500 awarded to the best solution after the first and second weeks of competition; payouts in phase 3 comprised $500 to each participant for proposals, $100 to each for submitting a solution, and an additional payout based on the performance of their solution.

Place	Phase 1	Phase 2	Phase 3
1	$10,000	$7,000	$500+$100+$750
2	$7,000	$4,500	$500+$100+$600
3	$5,000	$2,000	$500+$100+$200
4	$4,000	$1,000	$500+$100+$0
5	$3,000	$500	$500+$100+$0
6	$2,000		
7	$1,000		
Bonuses	$3,000		
Total	$35,000	$15,000	$4,550

eTable 2: Common evaluation metrics for comparing segmentations (TP, FP, FN, E, V and V_0 are further defined in the ONLINE-ONLY METHODS main text).

Metric Name	Symbol	Definition
Precision	P	TP/(TP+FP)
Recall	R	TP/(TP+FN)
Dice Coefficient	F_1	2 TP/(2TP+FP+FN)
Jaccard Index	J	TP/(TP+FP+FN)
Segmentation Score	S	$\exp\left[-\frac{E}{2V}\left(1+\left(\frac{V}{V_0}\right)^{1/3}\right)\right]$

Abbreviations: TP: true positive; FP: false positive; FN: false negative; E: Non-negative error volume defined as V FN / TP + FP; V: tumor volume; and V_0: scale parameter = $(4\pi/3)$ $(30mm)^3$

eTable 3: Algorithm, ensemble, and benchmark performance averaged over all scans in the validation, holdout and external data set. Inter-observer performance represents the average of all pair-wise comparisons between six different observers (15 pairs for the symmetric metrics Dice Coefficient and Jaccard Index, and 30 pairs for the asymmetric Segmentation-score). Standard errors on averages are provided in parentheses (numbers in parenthesis represent the one standard deviation variation in the least significant digit) ; inter-observer variation in parentheses represent the standard deviation of the distribution of scores. Intra-observer variation was estimated on a random sample of 21 scans from each data set. Precision and recall performance results are provided in Online-Only Methods eTables 4 and 5.

		Validation (96 scans)			Holdout (136 scans)			External (21 scans)		
Phase	Algorithm	Dice	Jaccard	S-score	Dice	Jaccard	S-score	Dice	Jaccard	S-score
1	1	0.52(4)	0.43(3)	0.42(3)	0.49(3)	0.39(3)	0.38(3)	-	-	-
	2	0.46(3)	0.35(3)	0.31(3)	0.47(2)	0.35(2)	0.31(2)	-	-	-
	3	0.44(3)	0.33(3)	0.29(3)	0.43(3)	0.32(2)	0.29(2)	-	-	-
	4	0.39(4)	0.30(3)	0.27(3)	0.35(3)	0.28(2)	0.26(3)	-	-	-
	5	0.37(3)	0.27(3)	0.23(3)	0.36(3)	0.26(2)	0.22(2)	-	-	-
	6	0.29(3)	0.20(2)	0.14(2)	0.29(2)	0.20(2)	0.15(2)	-	-	-
	7	0.32(3)	0.23(2)	0.17(2)	0.30(2)	0.20(2)	0.15(2)	-	-	-
2	1	0.66(3)	0.54(3)	0.53(3)	0.69(2)	0.57(2)	0.57(2)	-	-	-
	2	0.71(2)	0.57(2)	0.56(2)	0.70(2)	0.57(2)	0.57(2)	-	-	-
	3	0.67(3)	0.55(2)	0.54(3)	0.68(2)	0.56(2)	0.56(2)	-	-	-
	4	0.67(2)	0.54(2)	0.51(3)	0.70(2)	0.56(2)	0.55(2)	-	-	-
	5	0.69(2)	0.55(2)	0.53(2)	0.68(2)	0.54(2)	0.53(2)	-	-	-
	Ensemble	0.77(1)	0.64(2)	0.64(2)	0.78(1)	0.64(1)	0.65(2)	-	-	-

3	1	0.76(2)	0.64(2)	0.63(2)	0.75(1)	0.62(2)	0.62(2)	0.81(3)	0.71(4)	0.70(5)
	2	0.71(2)	0.59(2)	0.59(3)	0.72(2)	0.61(2)	0.61(2)	0.78(4)	0.67(4)	0.67(5)
	3	0.73(2)	0.60(2)	0.60(2)	0.72(1)	0.58(2)	0.59(2)	0.79(4)	0.68(5)	0.68(6)
	4	0.64(3)	0.52(3)	0.50(3)	0.63(2)	0.50(2)	0.48(3)	0.73(4)	0.60(4)	0.58(5)
	5	0.56(4)	0.47(3)	0.46(4)	0.58(3)	0.48(3)	0.48(3)	0.72(6)	0.62(6)	0.62(7)
	Ensemble	0.77(2)	0.65(2)	0.65(2)	0.79(1)	0.67(1)	0.68(2)	0.82(4)	0.72(4)	0.72(5)
Benchmarks	Comm	0.33(3)	0.23(2)	0.17(3)	0.37(3)	0.27(2)	0.22(2)	-	-	-
	Comm+Exp	0.59(3)	0.46(3)	0.42(3)	0.67(2)	0.53(2)	0.51(2)	-	-	-
	Inter-obs	-	-	-	-	-	-	0.80(4)	0.68(5)	0.66(6)
	Intra-obs	0.83(2)	0.71(3)	0.73(3)	0.87(1)	0.77(1)	0.79(1)	-	-	-

Abbreviations: Dice: Dice Coefficient; Jaccard: Jaccard Index; S-score: Contest-specific Segmentation-score metric; Comm, Comm+Exp: Commercially available software without and with human expert intervention; Inter-obs: Inter-observer variation; Intra-obs: Intra-observer variation

eTable 4: Primary methods (model or algorithmic approach), and utilization of supplemental data for training (all allowed) and/or inference (only tumor seed points and contrast allowed). Convolutional neural network-based solutions involve custom-designed architectures unless otherwise noted.

Phase	Place	Algorithm	Seed	Contrast	Tumors	Organs
1	1	RF	No	Yes	Yes	No
	2	2D CNN	No	No	Yes	Yes
	3	2D CNN	No	No	Yes	Yes
	4	CG	No	No	No	No
	5	RF	No	Yes	Yes	No
	6	2D CNN (Overfeat)	No	No	Yes	No
	7	2D CNN (SegNet)	No	No	Yes	No
2	1	2D CNN (SegNet)	Yes	No	Yes	No
	2	RF	Yes	Yes	Yes	No
	3	2D CNN (U-Net)	Yes	No	Yes	No
	4	CG	Yes	No	No	No
	5	RF	Yes	Yes	Yes	No
3	1	RF & 2D CNN (SegNet)	Yes	Yes	Yes	No
	2	2D CNN	Yes	No	Yes	No
	3	RF	Yes	Yes	Yes	No
	4	2D CNN (U-Net)	Yes	No	Yes	No
	5	3D CNN (U-Net)	Yes	No	Yes	No

Abbreviations: Convolutional neural network (CNN), Random Forest (RF), Cluster Crowth (CG), two-dimensional (2D), three-dimensional (3D)

eTable 5: Algorithm, ensemble, and benchmark performance averaged over all scans in the validation and holdout data sets. Standard errors on averages are provided in parentheses (numbers in parenthesis represent the one standard deviation variation in the least significant digit). Intra-observer variation was estimated on a random sample of 21 scans from each data set.

		Validation (96 scans)					Holdout (136 scan)				
Phase	Algorithm	<P>	<R>	<F_1>	<J>	<S>	<P>	<R>	<F_1>	<J>	<S>
1	1	0.55(4)	0.54(4)	0.52(4)	0.43(3)	0.42(3)	0.52(3)	0.54(3)	0.49(3)	0.39(3)	0.38
	2	0.50(3)	0.53(3)	0.46(3)	0.35(3)	0.31(3)	0.55(3)	0.51(3)	0.47(2)	0.35(2)	0.31
	3	0.50(4)	0.46(4)	0.44(3)	0.33(3)	0.29(3)	0.49(3)	0.44(3)	0.43(3)	0.32(2)	0.29
	4	0.47(4)	0.38(4)	0.39(4)	0.30(3)	0.27(3)	0.41(3)	0.35(3)	0.35(3)	0.28(2)	0.26
	5	0.44(4)	0.39(3)	0.37(3)	0.27(3)	0.23(3)	0.43(3)	0.39(3)	0.36(3)	0.26(2)	0.22
	6	0.33(3)	0.35(4)	0.29(3)	0.20(2)	0.14(2)	0.32(2)	0.35(3)	0.29(2)	0.20(2)	0.15
	7	0.34(3)	0.38(3)	0.32(3)	0.23(2)	0.17(2)	0.30(2)	0.39(3)	0.30(2)	0.20(2)	0.15
2	1	0.78(3)	0.62(3)	0.66(3)	0.54(3)	0.53(3)	0.75(2)	0.68(2)	0.69(2)	0.57(2)	0.57
	2	0.76(2)	0.73(2)	0.71(2)	0.57(2)	0.56(2)	0.73(2)	0.75(2)	0.70(2)	0.57(2)	0.57
	3	0.71(3)	0.78(2)	0.67(3)	0.55(2)	0.54(3)	0.69(2)	0.80(2)	0.68(2)	0.56(2)	0.56
	4	0.70(2)	0.73(2)	0.67(2)	0.54(2)	0.51(3)	0.71(2)	0.76(2)	0.70(2)	0.56(2)	0.55
	5	0.70(2)	0.76(2)	0.69(2)	0.55(2)	0.53(2)	0.66(2)	0.79(1)	0.68(2)	0.54(2)	0.53
	Ensemble	0.90(1)	0.93(1)	0.77(1)	0.64(2)	0.64(2)	0.88(2)	0.94(1)	0.78(1)	0.64(1)	0.65
3	1	0.77(2)	0.81(2)	0.76(2)	0.64(2)	0.63(2)	0.73(2)	0.82(1)	0.75(1)	0.62(2)	0.62
	2	0.77(2)	0.70(3)	0.71(2)	0.59(2)	0.59(3)	0.76(2)	0.73(2)	0.72(2)	0.61(2)	0.61
	3	0.73(2)	0.80(2)	0.73(2)	0.60(2)	0.60(2)	0.73(2)	0.77(1)	0.72(1)	0.58(2)	0.59
	4	0.65(3)	0.76(3)	0.64(3)	0.52(3)	0.50(3)	0.58(2)	0.85(1)	0.63(2)	0.50(2)	0.48
	5	0.66(4)	0.54(4)	0.56(4)	0.47(3)	0.46(4)	0.69(3)	0.56(3)	0.58(3)	0.48(3)	0.48
	Ensemble	0.83(3)	0.92(1)	0.77(2)	0.65(2)	0.65(2)	0.85(2)	0.94(1)	0.79(1)	0.67(1)	0.68

Benchmarks	Comm	0.38(4)	0.72(3)	0.33(3)	0.23(2)	0.17(3)	0.48(4)	0.68(2)	0.37(3)	0.27(2)	0.22
	Comm+Exp	0.74(3)	0.63(2)	0.59(3)	0.46(3)	0.42(3)	0.82(2)	0.64(2)	0.67(2)	0.53(2)	0.51
	Inter-obs	-	-	-	-	-	-	-	-	-	-
	Intra-obs	0.88(2)	0.79(3)	0.83(2)	0.71(3)	0.73(3)	0.90(2)	0.85(2)	0.87(1)	0.77(1)	0.79

Abbreviations: P: Precision; R: Recall, F_1: Dice Coefficient; J: Jaccard Index; S: Contest-specific Segmentation-score; Comm, Comm+Exp: Commercially available software without and with human expert intervention; Intra-obs: Intra-observer variation

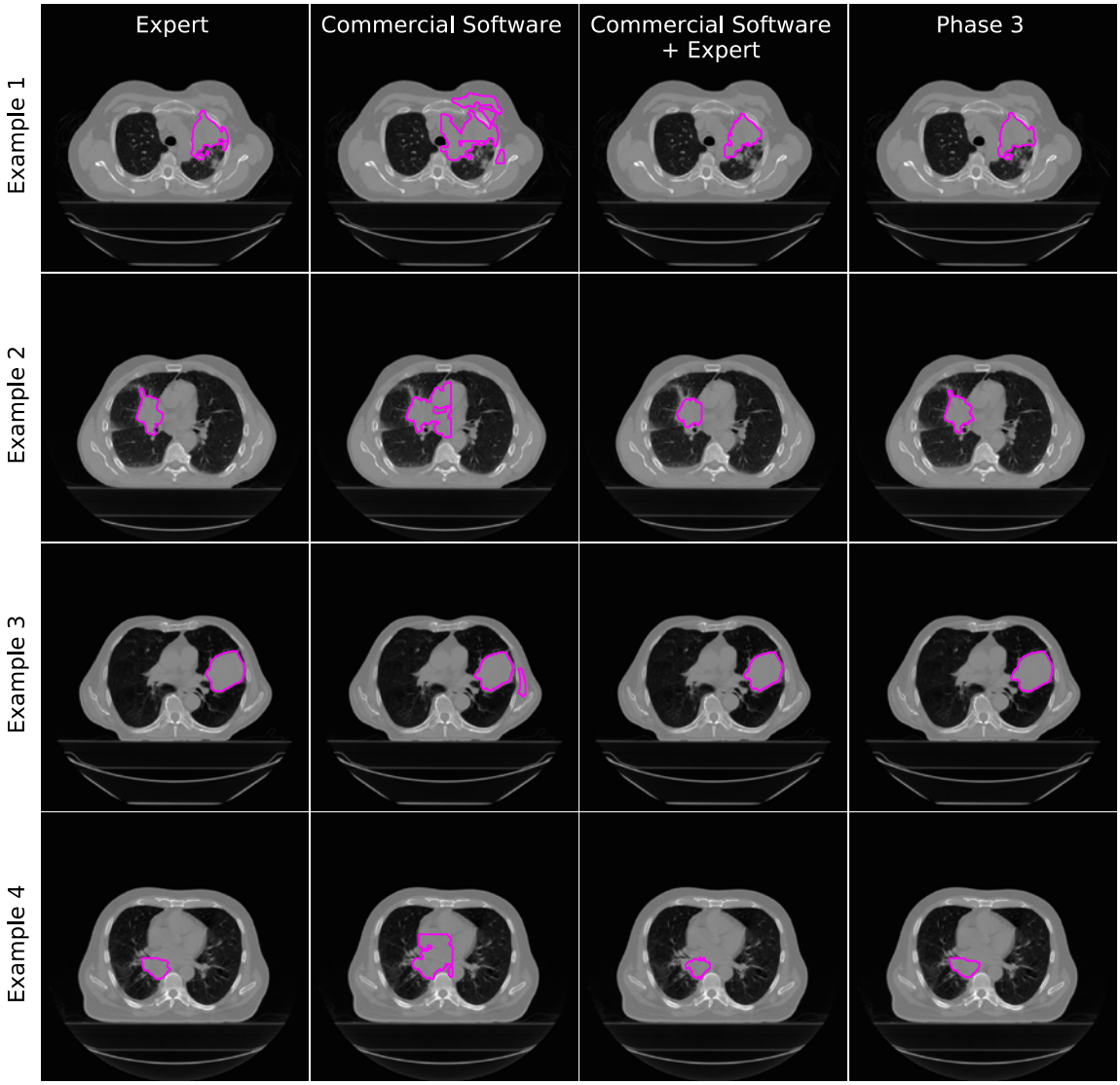
eTable 6: Algorithm, ensemble, and benchmark performance averaged over all scans in the external data set. Inter-observer performance represents the average of all pair-wise comparisons between six different observers (15 pairs for the symmetric metrics Dice Coefficient and Jaccard Index, and 30 pairs for the asymmetric Segmentation-score). Standard errors on averages are provided in parentheses (numbers in parenthesis represent the one standard deviation variation in the least significant digit); inter-observer variation in parentheses represent the standard deviation of the distribution of scores. Note that the well-known algebraic relationship between the metrics P, R, F_1 and J are only satisfied on a per scan basis, and are not satisfied in terms of their averages.

		External (21 scans)				
Phase	Algorithm	<P>	<R>	<F_1>	<J>	<S>
3	1	0.76(4)	0.92(2)	0.81(3)	0.71(4)	0.70(5)
	2	0.74(4)	0.89(3)	0.78(4)	0.67(4)	0.67(5)
	3	0.80(4)	0.84(4)	0.79(4)	0.68(5)	0.68(6)
	4	0.66(4)	0.90(4)	0.73(4)	0.60(4)	0.58(5)
	5	0.77(6)	0.71(7)	0.72(6)	0.62(6)	0.62(7)
	Ensemble	0.79(4)	0.91(3)	0.82(4)	0.72(4)	0.72(5)
Benchmarks	Comm	-	-	-	-	-
	Comm+Exp	-	-	-	-	-
	Inter-obs	0.8(1)	0.8(1)	0.80(4)	0.68(5)	0.66(6)
	Intra-obs	-	-	-	-	-

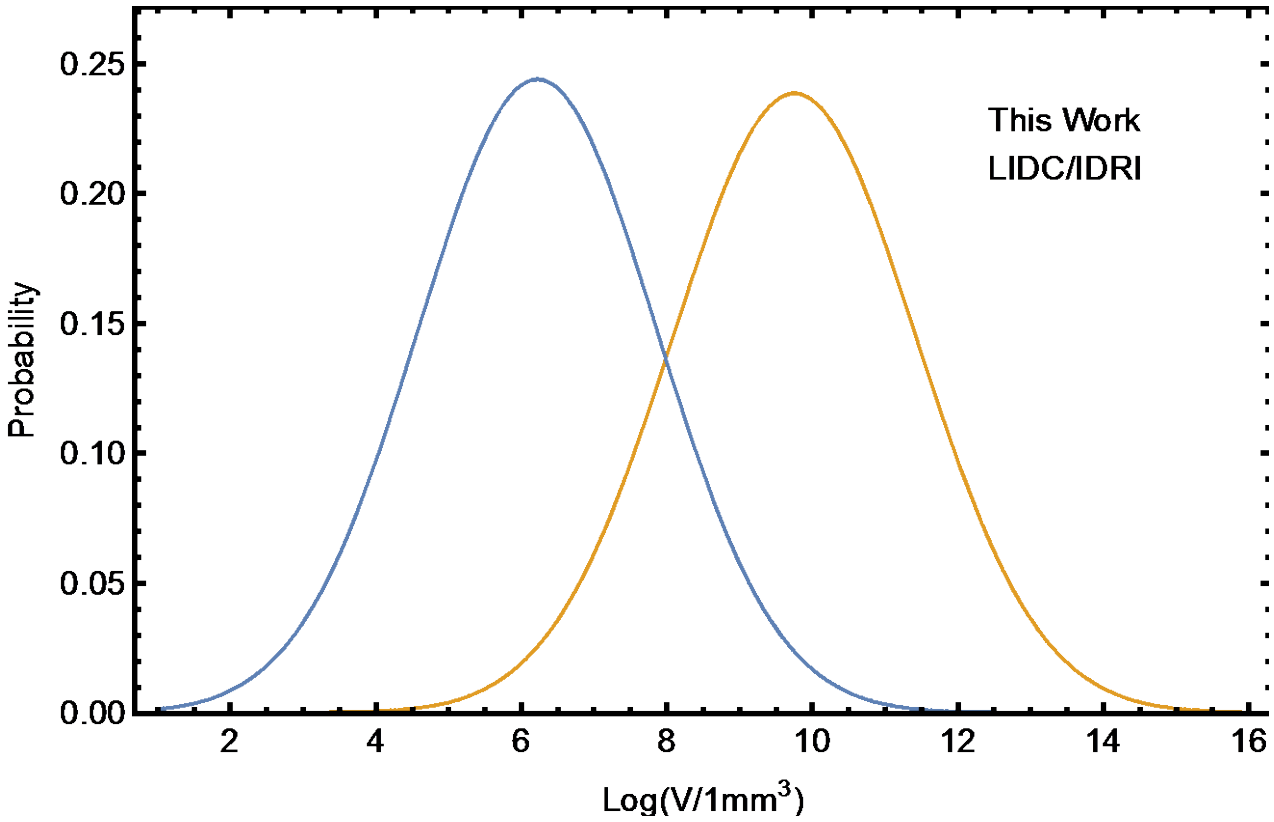
Abbreviations: P: Precision; R: Recall, F_1: Dice Coefficient; J: Jaccard Index; S: Contest-specific performance metric; Comm, Comm+Exp: Commercially available software without and with human expert intervention; Intra-obs: Intra-observer variation

C. ONLINE-ONLY FIGURES:

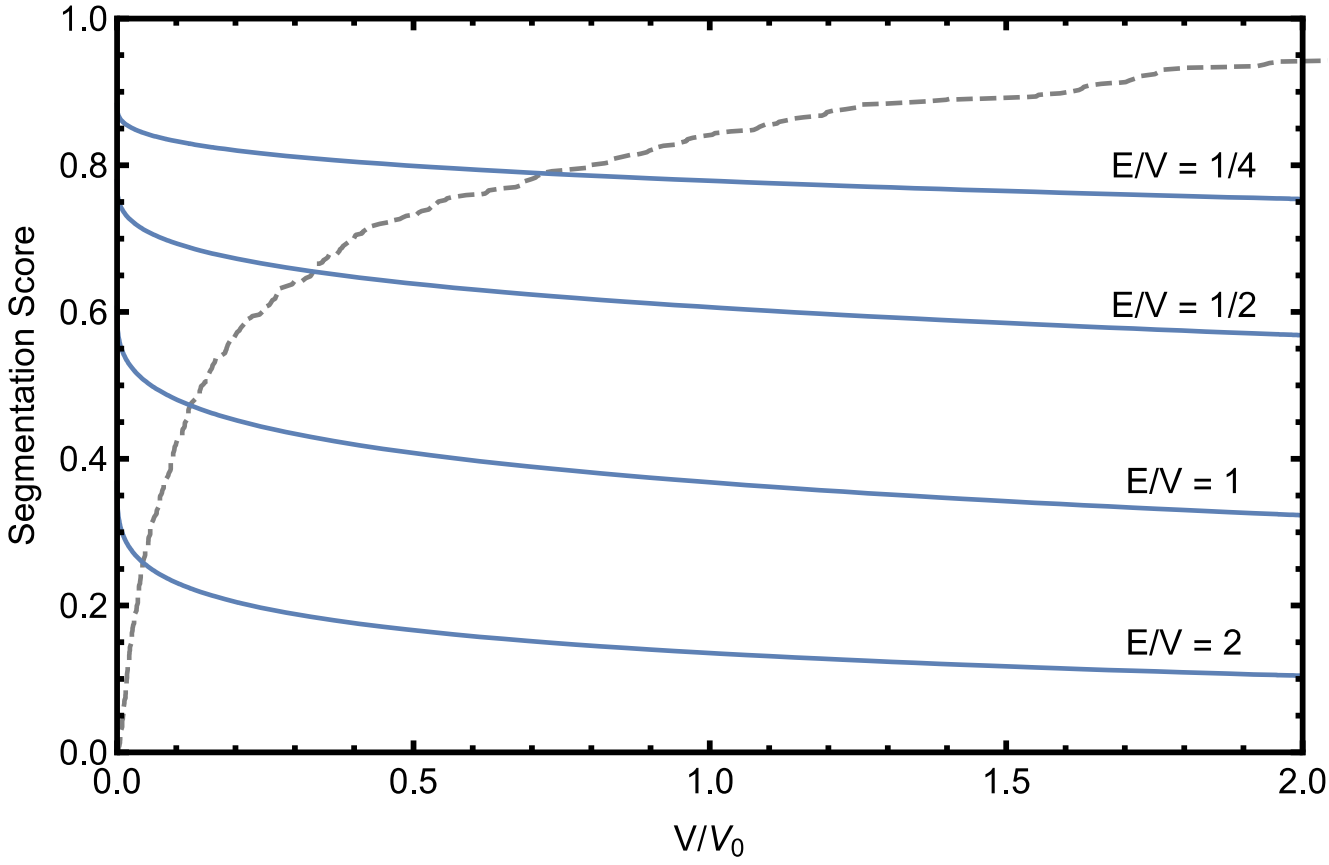
eFigure 1: Additional Examples of Human Expert versus Automated Segmentations. Four examples of a human expert segmentation (Human Expert) in comparison to automated segmentation from commercially available region growing-based segmentation technique before (Commerical Software + Expert) and after (Commercial Software + Expert) human intervention, and automated segmentations from the top algorithm from phase 3 of the contest (Phase 3).



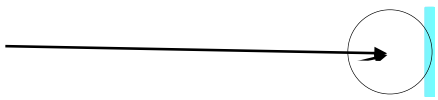
eFigure 2: Lung Tumor Volume Distribution in This Study Versus A Publicly-Available Dataset. A comparison of the per-scan lung tumor volume (V) distribution of 461 patients analyzed in this study and 1010 patients (only 868 contained nodules with sizes exceeding 3mm, as estimated by at least one observer, are shown) in the LIDC/IDRI dataset[3]. Lines indicate fits to a normal distribution, with a mean (standard deviation) were equal to 9.76 (1.67), and 6.22 (1.64), for This Work and LIDC/IDRI, respectively.



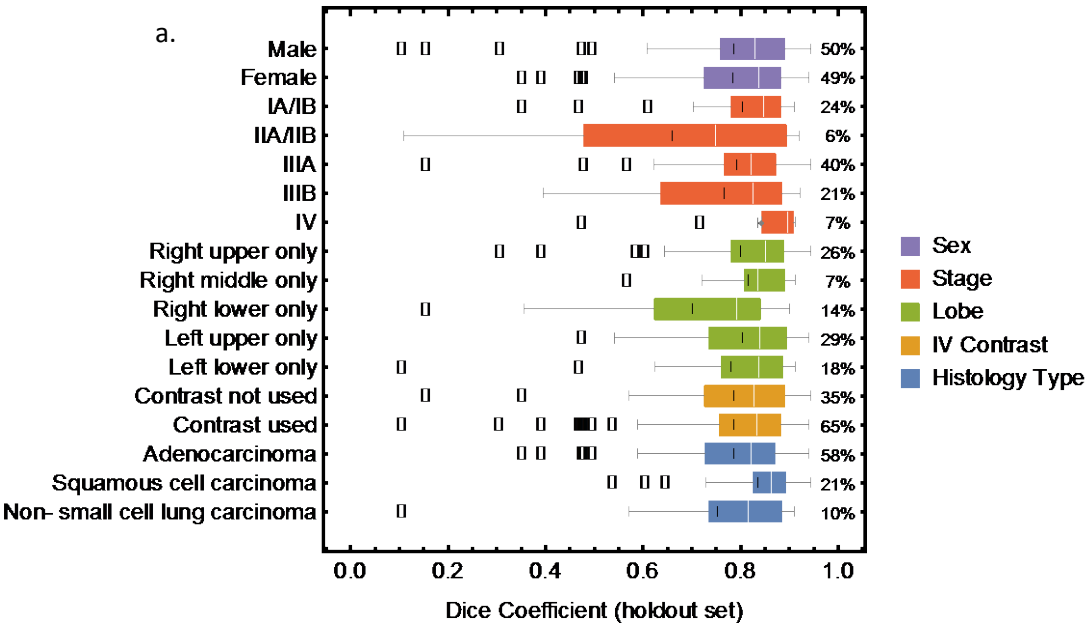
eFigure 3: Distribution of Segmentation Score (S-score) as a function of V/V_0, for various choices of E/V (solid lines). Dashed line indicates the fraction of scans with tumor volumes less than V/V_0.



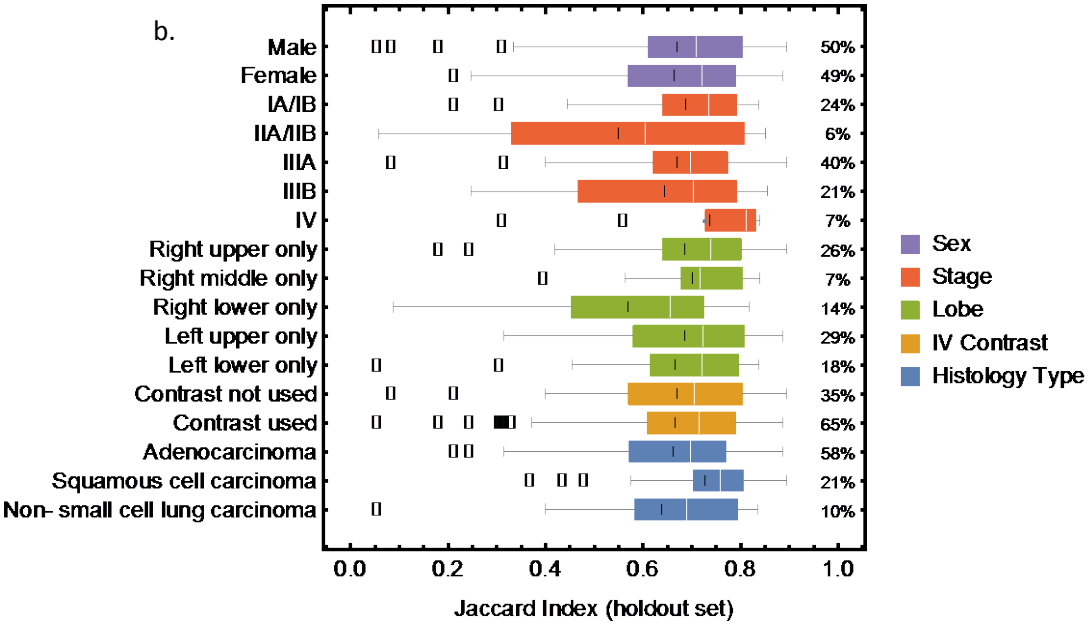
eFigure 4: Evolution of Winning AI Algorithms During the Multi-Phase Contest. Squares (circles) indicate winning solutions produced by new (previous) participants, and colors indicate the three main classes of algorithms/models employed (random forest, convolutional neural networks and cluster growth). Lines indicate the lineage of winning solutions (i.e., cases where winning solutions build upon winning solutions from previous contest phases; independent solutions are identified as those that did not build upon previous phase solutions).



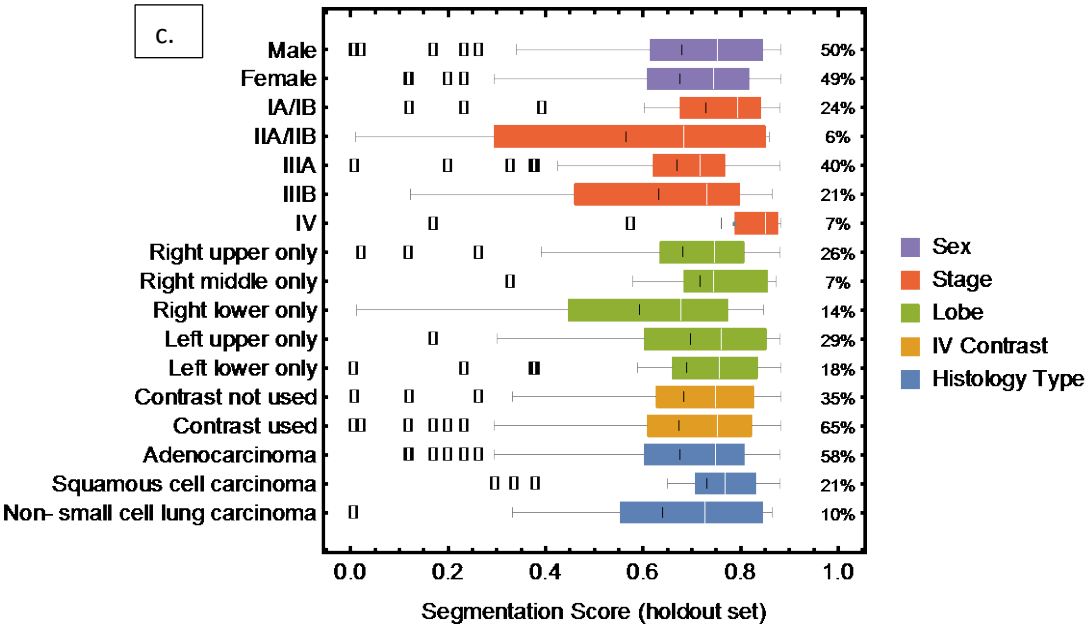
eFigure 5: Performance of Ensemble AI Algorithms in Clinical Sub-Groups. A comparison of ensemble algorithm performance based on metrics (a) Dice Coefficient (F_1), (b) Jaccard Index (J) and (c) contest-specific Segmentation score (S-score) for on a variety of clinical sub-group stratifications of the holdout dataset. Boxes span from the 0.25 to the 0.75 quantile; white vertical line within each box indicates the median; black vertical lines indicate the mean; whiskers indicate the span of the data set, excluding outliers; outliers are defined as points beyond 3/2 the inter-quantile range starting from the edge of the box. Percentages indicate the proportion of the holdout data set associated with each stratum.



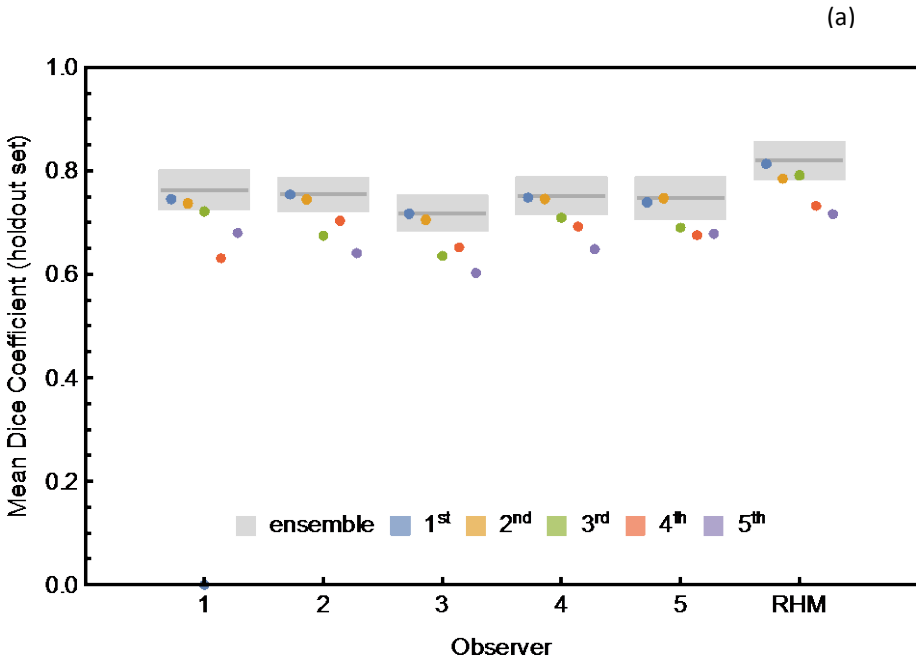
b.



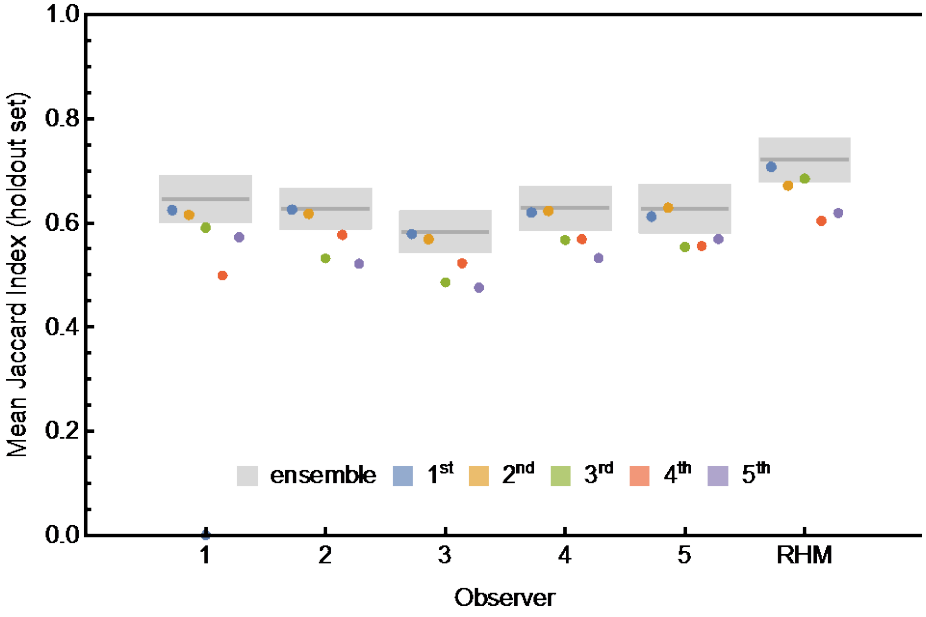
c.



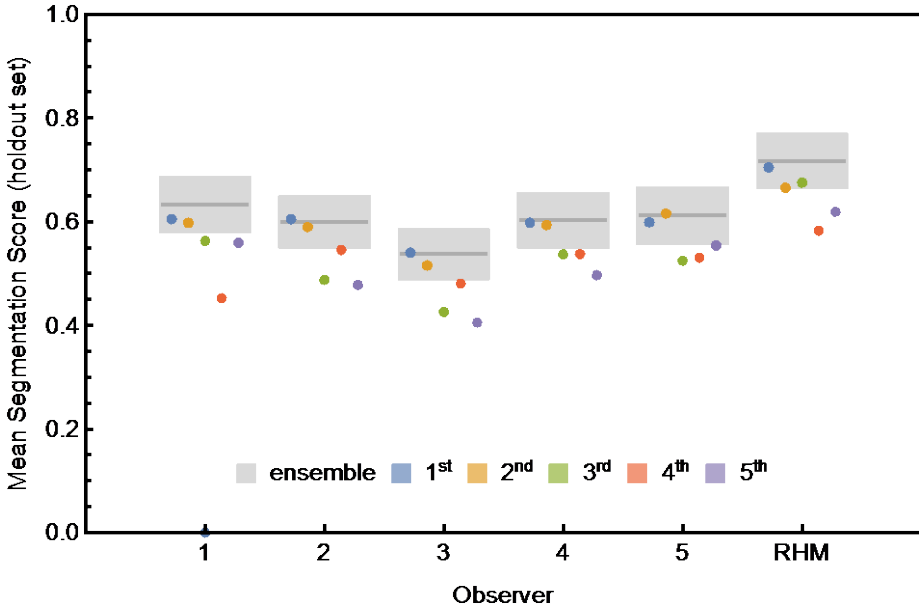
eFigure 6: Performance of Top Segmentation Algorithms from The Contest on an External Dataset with External Experts' Manual Segmentations Acting as Ground Truth. A comparison of performance based on metrics (a) Dice Coefficient (F_1), (b) Jaccard Index (J), and (c) contest-specific Segmentation score (S-score) of top five algorithms from phase 3 and the phase 3 ensemble model on an external data set of 21 scans annotated by six different observers (including R.H.M.). Performance results were computing using each observer's contours as the ground truth. Lines and points indicate mean performance and boxes indicate standard errors in the estimate.



(b)



(c)



D. ONLINE-ONLY REFERENCES:

1. Aerts HJ, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun.* 2014;5:4006.
2. Dwork C, Feldman V, Hardt M, Pitassi T, Reingold O, Roth A. STATISTICS. The reusable holdout: Preserving validity in adaptive data analysis. *Science (New York, N.Y.* 2015/8/7 2015;349(6248):636-638.
3. Armato SG, McLennan G, Bidaut L, et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans. *Medical physics.* 01/24

09/08/received

11/16/revised

11/20/accepted 2011;38(2):915-931.

E. Online-Only Acknowledgements:

The authors would like to acknowledge support from Dana-Farber Cancer Institute, Eric and Wendy Schmidt Foundation, Harvard Business School Division of Research and Faculty Development, Harvard Business School Kraft Precision Medicine Accelerator, Harvard Catalyst, Harvard Medical School, Laura and John Arnold Foundation, and NASA Center of Excellence for Collaborative Innovation.

The authors would also like to acknowledge the invaluable input of the TopCoder team and participants of the contest including the following contest winners who shared their identities:

Marek Cygan, PhD

John Gardner, PhD

Wladimir Leite

Peter Novotný, PhD

Roman Shvetsov

Thomio Watanabe, MS