Supplemental Information:

Supplemental Tables:

Supplemental Table 1:

Label	Measurement	Modality	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
(a) Comparison of patient characteristics between training and testing sets	p-value from 2-sample t-test ($\alpha = 0.05$, $p > 0.05$ indicates no signif. diff b/w groups on that given characteristic)	MR	Age: p=0.09 Sex: p=0.18 T-Score: p=0.18	Age: p=0.13 Sex: p=0.10 T-Score: p=0.19	Age: p=0.17 Sex: p=0.18 T-Score: p=0.13	Age: p=0.17 Sex: p=0.13 T-Score: p=0.13	Age: p=0.19 Sex: p=0.14 T-Score: p=0.18
		CT	Age: p=0.11 Sex: p=0.09 T-Score: p=0.11	Age: p=0.18 Sex: p=0.16 T-Score: p=0.19	Age: p=0.09 Sex: p=0.05 T-Score: p=0.06	Age: p=0.18 Sex: p=0.14 T-Score: p=0.08	Age: p=0.13 Sex: p=0.15 T-Score: p=0.14
		X-ray	Age: p=0.16 Sex: p=0.10 T-Score: N/A	Age: p=0.14 Sex: p=0.08 T-Score: N/A	Age: p=0.11 Sex: p=0.07 T-Score: N/A	Age: p=0.12 Sex: p=0.19 T-Score: N/A	Age: p=0.14 Sex: p=0.20 T-Score: N/A
(b) Slice Selection Accuracy	Accuracy	MR	86.6%	86.2%	86.7%	85.3%	87.5%
		CT	85.2%	88.9%	85.2%	92.6%	81.5%
		X-ray	-	-	-	-	-
	Avg Selected Slice Distance from Ground Truth (95% CI; p-value; range)	MR	-0.04 (-0.02 –0.10; p-value = 0.20; range = -1 to 1)	-0.04 (-0.09 – 0.01; p-value = 0.10; range = -1 to 2)	0.04 (-0.01 – 0.10; p-value = 0.10; range = 0 to 2)	0.04 (-0.15 – 0.09; p-value = 0.14; range = -1 to 0)	0.009 (-0.04 – 0.06; p-value = 0.75; range = -1 to 1)
		CT	0.07 (-0.08 – 0.23; p-value = 0.33; range = -1 to 0)	-0.03 (-0.10 – 0.17; p-value = 0.57; range = -1 to 1)	0.19 (-0.01 – 0.38; p-value = 0.06; range = -1 to 1)	0.008 (-0.11 – 0.11; p-value = 1.00; range = -1 to 0)	-0.04 (-0.21 – 0.14; p-value = 0.66; range = -1 to 2)
		X-ray	-	-	-	-	-
(c) Detected vertebrae above IoU 0.7 Reported as: accuracy % (# of correct detection/ # of vertebrae in testing fold)	Lumbar Region	MR	99% (576/581)	98% (565/578)	99% (575/580)	97% (568/582)	97% (560/577)
		CT	95% (558/590)	98% (571/579)	96% (557/580)	97% (563/577)	97% (571/584)
		X-ray	95% (145/152)	90% (136/150)	95% (143/150)	90% (139/153)	94% (141/149)
	Thoracic Region (accuracy %)	MR	92% (260/280)	96% (270/280)	97% (271/280)	95% (271/284)	94% (261/276)
		CT	97% (271/278)	95% (267/279)	97% (272/280)	97% (270/277)	98% (275/278)
		X-ray	91% (65/71)	92% (60/65)	96% (66/69)	94% (66/70)	93% (62/66)
	Cervical Region (accuracy %)	MR	96% (653/680)	98% (676/683)	97% (659/679)	98% (672/679)	92% (621/675)
		CT	94% (645/683)	91% (617/678)	94% (635/679)	96% (656/683)	91% (617/675)
		X-ray	-	-	-	-	-
(d) Segmentation DICE Score Reported as: Median Dice Score [25^{th}-75^{th} percentile IQR]	Lumbar Region	MR-Bone	0.967 (0.940-0.977)	0.949 (0.945-0.958)	0.967 (0.958-0.973)	0.958 (0.953-0.961)	0.961 (0.957-0.961)
		MR-Disc	0.956 (0.947-0.973)	0.956 (0.945-0.965)	0.964 (0.955-0.975)	0.967 (0.950-0.972)	0.967 (0.952-0.972)
		CT	0.966 (0.944-0.976)	0.954 (0.949-0.960)	0.965 (0.958-0.971)	0.970 (0.950-0.974)	0.965 (0.956-0.974)
		X-ray	0.957 (0.926-0.960)	0.956 (0.943-0.961)	0.953 (0.945-0.961)	0.960 (0.937-0.963)	0.949 (0.927-0.959)
	Thoracic Region	MR-Bone	0.943 (0.927-0.951)	0.931 (0.926-0.940)	0.950 (0.940 – 0.957)	0.948 (0.927-0.958)	0.951 (0.931-0.955)
		MR-Disc	0.949 (0.928-0.965)	0.936 (0.921-0.949)	0.949 (0.935-0.962)	0.950 (0.921-0.961)	0.951 (0.916-0.968)
		CT	0.961 (0.938-0.974)	0.955 (0.952-0.965)	0.966 (0.955-0.976)	0.975 (0.949-0.987)	0.975 (0.949-0.989)
		X-ray	0.950 (0.944-0.958)	0.944 (0.933-0.953)	0.964 (0.948 – 0.976)	0.974 (0.944-0.984)	0.950 (0.929-0.972)
	Cervical Region	MR-Bone	0.967 (0.942-0.974)	0.969 (0.959-0.969)	0.966 (0.962-0.973)	0.970 (0.944-0.979)	0.967 (0.942-0.972)
		MR-Disc	0.974 (0.956-0.981)	0.957 (0.951-0.971)	0.968 (0.958-0.977)	0.975 (0.958-0.984)	0.970 (0.945-0.983)
		CT	0.972 (0.962-0.981)	0.966 (0.961-0.974)	0.973 (0.967-0.979)	0.970 (0.957-0.981)	0.969 (0.967-0.985)
		X-ray	-	-	-	-	-
(e) Radiomics: Reported as	Kurtosis	MR	0.004 ± 0.211 (p-value: 0.43,	0.003 ± 0.170 (p-value: 0.40,	0.001 ± 0.188 (p-value: 0.75,	0.006 ± 0.186 (p-value: 0.21,	0.008 ± 0.218 (p-value: 0.14,

mean difference ± 1 SD, p-value (result from paired t test assessing diff b/w predicted and actual measurement. $\alpha = 0.05$, favorable for null to not be rejected), and r-value (comparison between ground-truth derived and predicted segmentation derived feature values)			r=0.98)	r=0.98)	r=0.99)	r=0.96)	r=0.96)
		CT	-0.064 ± 0.546 (p-value: 0.22, r=0.97)	-0.034 ± 0.433 (p-value: 0.41, r=0.98)	0.046 ± 0.454 (p-value: 0.29, r=0.98)	-0.003 ± 0.414 (p-value: 0.94, r=0.99)	-0.057 ± 0.621 (p-value: 0.34, r=0.97)
		X-ray	0.003 ± 0.199 (p-value: 0.70, r=0.96)	0.013 ± 0.226 (p-value: 0.17, r=0.96)	0.005 ± 0.218 (p-value: 0.62, r=0.99)	0.009 ± 0.240 (p-value: 0.38, r=0.97)	0.009 ± 0.196 (p-value: 0.28, r=0.96)
	Mean of positive valued pixels	MR	0.010 ± 0.942 (p-value: 0.67, r=0.98)	0.047 ± 1.413 (p-value: 0.16, r=0.96)	0.096 ± 2.803 (p-value: 0.15, r=0.96)	0.064 ± 1.883 (p-value: 0.16, r=0.96)	0.017 ± 0.924 (p-value: 0.45, r=0.97)
		CT	0.035 ± 0.743 (p-value: 0.63, r=0.98)	-0.211 ± 1.995 (p-value: 0.27, r=0.97)	0.353 ± 7.087 (p-value: 0.60, r=0.98)	-0.012 ± 0.979 (p-value: 0.90, r=0.99)	-0.022 ± 0.793 (p-value: 0.78, r=0.98)
		X-ray	0.076 ± 1.697 (p-value: 0.30, r=0.97)	0.062 ± 1.645 (p-value: 0.38, r=0.97)	0.080 ± 3.275 (p-value: 0.57, r=0.98)	0.187 ± 5.231 (p-value: 0.41, r=0.97)	0.018 ± 2.067 (p-value: 0.84, r=0.99)
	Entropy	MR	0.001 ± 0.112 (p-value: 0.64, r=0.98)	0.001 ± 0.115 (p-value: 0.73, r=0.98)	0.003 ± 0.120 (p-value: 0.28, r=0.97)	0.002 ± 0.119 (p-value: 0.53, r=0.97)	0.002 ± 0.118 (p-value: 0.48, r=0.97)
		CT	-0.006 ± 0.114 (p-value: 0.57, r=0.97)	0.001 ± 0.129 (p-value: 0.93, r=0.99)	0.009 ± 0.128 (p-value: 0.45, r=0.97)	0.003 ± 0.134 (p-value: 0.82, r=0.98)	0.001 ± 0.121 (p-value: 0.98, r=0.99)
		X-ray	-0.002 ± 0.046 (p-value: 0.38, r=0.97)	0.002 ± 0.049 (p-value: 0.32, r=0.97)	0.000 ± 0.047 (p-value: 0.86, r=0.99)	0.001 ± 0.050 (p-value: 0.62, r=0.98)	-0.003 ± 0.046 (p-value: 0.19, r=0.96)
	Skewness	MR	0.001 ± 0.094 (p-value: 0.62, r=0.98)	-0.001 ± 0.093 (p-value: 0.91, r=0.99)	0.003 ± 0.086 (p-value: 0.18, r=0.96)	0.002 ± 0.088 (p-value: 0.31, r=0.97)	0.001 ± 0.089 (p-value: 0.54, r=0.97)
		CT	-0.019 ± 0.157 (p-value: 0.21, r=0.96)	0.002 ± 0.147 (p-value: 0.91, r=0.99)	0.009 ± 0.151 (p-value: 0.52, r=0.97)	0.010 ± 0.148 (p-value: 0.47, r=0.97)	-0.005 ± 0.152 (p-value: 0.75, r=0.98)
		X-ray	0.003 ± 0.099 (p-value: 0.46, r=0.97)	-0.001 ± 0.109 (p-value: 0.91, r=0.99)	-0.001 ± 0.102 (p-value: 0.87, r=0.99)	0.005 ± 0.106 (p-value: 0.32, r=0.97)	-0.002 ± 0.104 (p-value: 0.69, r=0.98)

Supplemental Materials/Methods:

Section 1: Data Annotation Procedures:

Segmentations were made using ITK-SNAP (v3.8.0) [12]. The slice of an imaging series that had the most visible vertebrae (and, in the case of a tie, was closest to the middle slice) was used for both the landmark annotation. Annotators (MSK research assistants with hundreds to thousands of hours of prior MSK annotation experience) were trained for five hours and had their annotations manually approved and revised by A.S. Review on completed scans were all completed by A.S. Scans were sent back for redoing or reassigned if the landmarks were low-quality and varied with regards to placement.

Section 2: Training Data Augmentation:

Training cases were further augmented using the imgaug library (v0.4.0) [17] with the following permutations sequentially applied to each case (note: ±x indicates that a value was randomly selected from the range -x to +x): rotations of ±15°, ±30% contrast, ±30% brightness, random cropping (up to 30% horizontally, up to 50% vertically), flip horizontal and/or flip vertical. For each image in the training set, five augmentations were generated (each different in rotation, contrast, brightness, crop, and orientation).

Section 3: Slice Selection Procedure:

For annotation and neural network evaluation, slices were selected for landmark annotation in the following manner (after vertebral body bounding boxes are found). Find the slice with the highest number of visible vertebrae. If there are multiple slices tied for the number of visible vertebrae, pick the slices with the largest vertebral body cross section (for neural networks, pick the slice that has the largest cumulative

area enclosed within bounding boxes). If multiple slices are tied for the size of vertebrae, find the slice number in the image series and pick the one closest to the middle slice of the series.

Section 4: Network Parameters and Design:

The Mask R-CNN baseline neural network design was changed to account for the single channel nature of medical images as opposed to the 3-channel RGB images originally meant to be used in the network (ref Supplemental Figure 1 for our full network design & explanation). Specifically, we made changes to the Mask R-CNN implementation created by Facebook AI Research (called Detectron2) which was selected due to its extensibility and fast training speed [18].Python (v3.6.9) and PyTorch (v1.6; built on CUDA v 10.1) were used to code and configure the network [19]. Baseline configuration for the Landmark networks was inherited from the COCO-Person Keypoint Detection Baseline w/Keypoint R-CNN R50-FPN model and the final layers of this network were changed to produce 6 landmarks for height calculations. Base learning rate (BASE_LR config variable) was set to 0.00010 (SGD optimizer w/momentum and Nesterov momentum enabled), with 1000 epochs for the landmark networks (experimentation with higher epochs yielded minimal benefits and minimal decrease in loss). Batch size (BATCH_SIZE_PER_IMAGE config variable) was set to 256. For all experiments (except Precision-Recall curve making), evaluation IoU threshold was set to 0.7 for evaluation of bounding box accuracy. Code for the network available upon request prior to publication and will be available upon publication in an open-source + documented GitHub repository.

Section 5: Network Training:
Network training was carried out on Google Colab, which allocates GPU usage based on global availability and thus training times may vary. Available GPU compute power include NVIDIA K80s and can be run for up to 12 hours for free, which is plenty of time for our experiments. On average, once imaging data was loaded into memory (also free up to 12 GB), each network took 26 ± 2 minutes to run for landmark networks. High variation in the amount of time it took to train networks was due to the variable availability of GPU compute time; however, no network training session took more than 30 minutes to run.

Section 6: Full Network Overview:
For the following description, please refer to Supplemental Figure 1. The Feature Generation Network (FGN) comprises of four stages. Before being input to the network, an image is first rescaled in dimensions and the range of image values is rescaled from 0 to 255. Additionally, the 1 channel grayscale image is converted to a 3 channel RGB image (since the pretrained ResNet was trained on RGB images). After the image is input into the network, several convolutional operations are performed that generates five intermediate representations for the neural network to learn on. Ultimately, each stage progressively reduces the height and width of the input image and increases the number of features. At the end of this part of the network, we have 5 feature maps, each with 256 features, each at different scales (200x200, 100x100, 50x50, 25x25, 13x13). These outputs will be used at subsequent stages of the network to detect objects at multiple scales, making the network scale-invariant.

The Region Recognition Network (RRN) performs a 3x3 convolution on each of the feature maps from the FGN. From there, two separate convolutions are also performed. The first convolution reduces the dimensionality to 3 channels and a loss function is applied to get this section of the network to produce a probability map (called an "objectness map") that is of value 1 where a vertebral body is present and 0 for background. The loss function for this (annotated as "loss_rpn_cls") helps the network optimize a probability map to locate objects, using objectness maps calculated from keypoint annotations. A separate convolution is also applied after the previous 256 intermediate convolution that outputs a 3x4 channel vector (corresponding to the four points of a bounding box). The loss function for this (annotated as "loss_rpn_reg") helps the network optimize candidate bounding box coordinates, using bounding boxes calculated from keypoint annotations. The top 1000 bounding boxes are isolated from this section of the network (based on proportion of candidate bounding box covering areas where probability map = 1).

Outputs from this section of the network are then combined with prior feature maps via the ROIPool and ROIAlign algorithms. These feature maps are then flattened to a 256x7x7 = 12544x1 size vector and put through two fully connected layers that each have 1024 outputs. From there, these 1024 outputs are then connected to another fully connected layer with 2 outputs (corresponding to the two potential classes for a bounding box: background or vertebral body; loss function loss_cls), a fully connected layer with 8 outputs (corresponding to 4 box coordinates for the two aforementioned classes; loss function loss_reg), and a landmark network.

The landmark network takes in the 1024 outputs from the initial set of fully connected layers and performs a set of convolutional operations (with 7 intermediate convolutions) and does a deconvolutional operation that outputs 6 channels (one for each vertebral body keypoint) and upsamples the final output. Keypoints are reported as the single highest pixel value in each of the six channels of the output (loss function; loss_kp).

The above description renames some functions from the Detectron2 framework for ease of understanding. Original names are as follows: Feature Generation Network = Feature Pyramid Network; Region Recognition Network = Region Proposal Network; Landmark Network = Keypoint RCNN; stage 1, 2, 3, 4 = res2, res3, res4, res5. Lateral outputs from FGN referred to as p2, p3, p4, p5, p6. Loss function annotations source code can be found in Detectron2 code repository on Github.

Section 7: Network Evaluation:
For all methods, evaluation was carried out on slices with the most visible vertebrae in imaging study. AUCs, Confusion Matrix creation, and Lumbar Lordosis correlations were calculated using SciPy (v1.5.2; calculations available in code repository) [20]. Figures were constructed using Matplotlib (v3.1.2) [21].