Appendix

Mask-RCNN Training

The Mask-RCNN model was adopted from a TensorFlow implementation (GitHub – ahmedfgad/Mask-RCNN-TF2: Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow 2.0). A dedicated Anaconda environment was established for model training and evaluation. As this implementation of Mask-RCNN requires a JSON file for the ground truth annotation, the segmentation mask images were converted to a JSON file. Both the downsampled radiograph tag image files and the JSON file were input for model training.

The initial learning rate was set at 0.001 for stage 1 and stage 2, then decreased by a factor of 10 in stage 3. The number of epochs for training for stage 1, stage 2, and stage 3 were 80, 40, and 80, respectively, with no early stopping. Training was performed on a single NVIDIA GTX Titan Xp (12GB VRAM) with a batch size of 2. This was the largest batch size that could be used for training due to memory constraints. A second round of training using the weights from the best model (highest validation accuracy) from the first 200 epochs was performed. The learning rate was decayed by a factor of 10 for each successive stage in the second round of training. No significant improvement in performance was observed.

Mask-RCNN Predictions

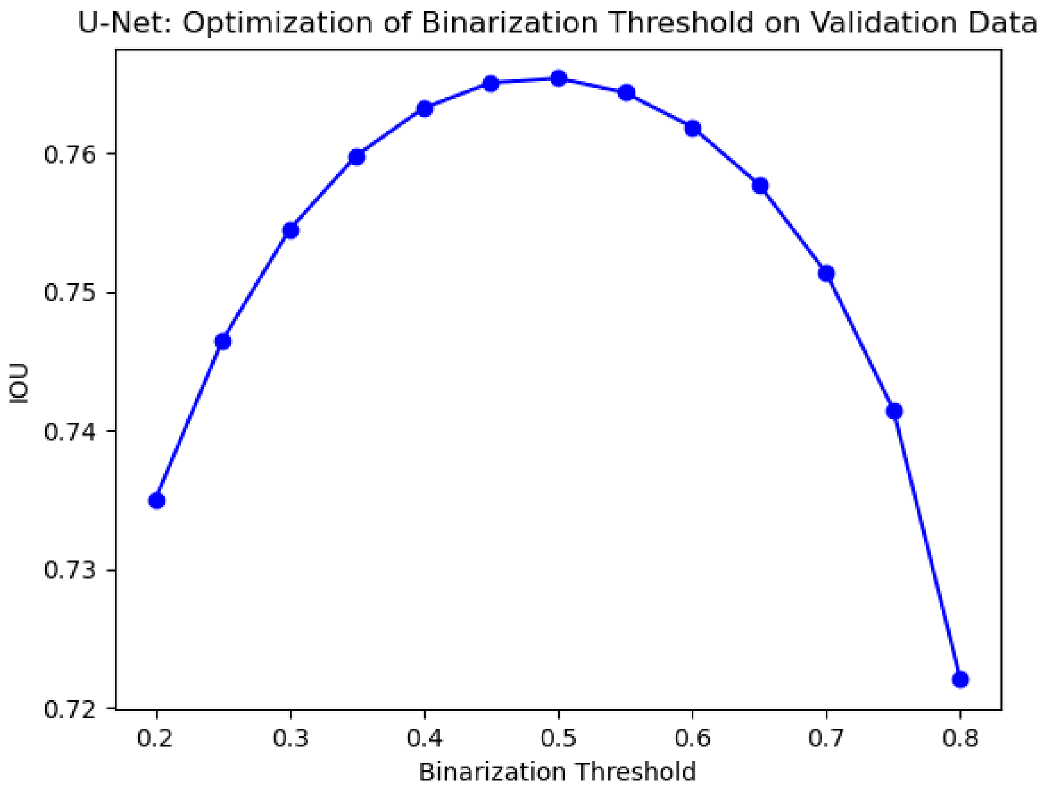
Using the best performing model, downsampled radiograph images from the held-out test set were input for inferencing. Outputs of the model include the bounding box coordinates for each detected VB, a single channel in a multidimensional array for each segmentation mask, and a corresponding score, or the confidence probability for each predicted class. The score is calculated on two tasks: 1) if it classifies a segmentation mask to the correct class and 2) the intersection over union (IoU) regression of the predicted mask to the ground truth mask. With these outputs, radiograph images overlayed with the segmentation masks and bounding boxes were generated in the original DICOM image. Average inference time per image was 0.214 seconds over the held-out test set.

U-Net Training

A U-Net model was adapted from the existing Keras implementation here: https://github.com/zhixuhao/unet. Many variations were explored, but the only model variation that produced meaningful improvements was training to perform a basic binary semantic segmentation task (i.e. background is one class, and any vertebral body is another class), and using contrast limited adaptive histogram equalization (skimage v0.18.3 implementation) with default parameters as a preprocessing step. The Keras (v2.3.1) framework was used to set up data ingestion, prediction, and inference. The model was trained using the ADAM optimizer with the initial learning rate set to 0.0001 for a maximum of 100 epochs, with a batch size of 4. Training was performed on a single NVIDIA GTX Titan Xp (12GB RAM). Loss function was the sum of cross-entropy and a soft-Dice loss function described here: https://www.jeremyjordan.me/semantic-segmentation/. The EarlyStopping callback was used with a patience parameter set to 4. The model checkpoint callback was also used. Total training time was 3:43:56.7 (HH:MM:SS) for 11 epochs before early stopping. Final training metrics are shown in the Supplemental Table 1.

Loss	0.0409
Accuracy	0.9972
Cross Entropy	0.0081
Validation Loss	0.0651
Validation Accuracy	0.9970
Validation Cross Entropy	0.0089

Supplemental Table 1 Training metrics for U-Net.



Supplemental Figure S1 Sensitivity analysis of binarization threshold on validation data.

U-Net Predictions

Average inference time for the held-out test set was 0.052s per image. Jaccard scores were calculated for interim evaluation of segmentation prior to postprocessing. Average Jaccard Score (IoU) over the held-out test set is 0.716 per image. Other performance metrics are discussed in the main section of the paper.

To calculate centroid coordinates for each VB instance, the scikit-image Python library (scikit-image: Image processing in Python — scikit-image) was employed. For Mask-RCNN, as each predicted segmentation mask is represented as a unique channel in an array, or instance, each channel was converted from a Boolean (TRUE/FALSE) to an integer (1/0). Centroid coordinates were calculated for each channel by labeling image regions (Label image regions — skimage v0.19.2 docs (scikit-image.org)). For each labeled image region, the image moment is determined by a scikit-image function, from which centroids are calculated as the first moment of area. For example, each prediction mask from a given radiograph would be labeled as an image region and have its image moment be determined. The centroids are then scaled to the dimensions of the original DICOM radiograph. These coordinates were saved to a data frame along with the image ID, which was used to evaluate model performance against ground truth centroids, as well as combining predictions from U-Net.

For U-Net, output from inference is a grayscale image, where each pixel is an estimated probability of the pixel overlapping a vertebral body. This mask was binarized using a threshold of 0.61, and patches of connected positively identified pixels were analyzed separately using the scikit-image library. This threshold was tuned by sensitivity analysis of the IoU on the validation set. A suboptimal value (as measured by IoU) was used because using centroids makes the analysis robust to under-segmentation, and by using a higher value, connections between adjacent patches were minimized. Patches were filtered by size using the following method: the threshold was set per-image using one-third the area of the second largest patch. This method was employed to limit errors from outliers in area, as when patches overlapping two VBs were connected erroneously. Centroids of the filtered patches were then reported as results. The centroids were saved to a data frame with the image ID.

Intersection over Union (IoU) and Dice Coefficient Calculations

Both U-Net and Mask-RCNN IoU and Dice coefficient were calculated in the same manner, independently. The ground truth segmentation masks in the MrOS held-out test set were saved as separate, single channel 512x512 arrays to match the downsampled tag image files. Mask-RCNN outputs separate, single channel arrays for each detected VB. IoUs were calculated on a per VB basis for each radiograph. If a ground truth mask was not detected by a model, the IoU for that mask was registered as a zero. IoU scores were calculated in this manner were averaged on a per radiograph basis, resulting in a final average IoU score for each model. Dice coefficients were calculated in the same manner.

Categorizing Detections

VB predictions were categorized by gross location on the radiograph compared to ground truth annotations. This was done to count true positives, false positives, and false negatives. Predictions that were above and below the topmost and bottommost ground truth were respectively labeled “top” and “bottom”. Predictions were also filtered by horizontal coordinate before being evaluated as TP/FP/FN. The average horizontal coordinate of all VBs per radiograph was found. Predicted VB centroids far enough from this centerline were labeled as “off-column” and excluded from further analysis. The distance threshold was based on the average endplate width of the predicted centroids. Predicted centroids farther than one half the average endplate width in horizontal distance were considered “off-column”. Any remaining predictions were labeled as “gap”, corresponding to those that were false negative detections between the topmost and bottommost ground truth annotations. Details of the definitions for classifying the gross position of predicted VB centroids are included in Supplemental Table 2.

Subcategory	***Definition***
Top	VB centroid *y* coordinate < minimum GT centroid *y* coordinate that is not off-column
Bottom	VB centroid *y* coordinate > maximum GT centroid *y* coordinate that is not off-column
Off-column	VB centroid ≥ 1/2 average endplate widths from GT *x* coordinate mean
Gap	If not any of above subcategories

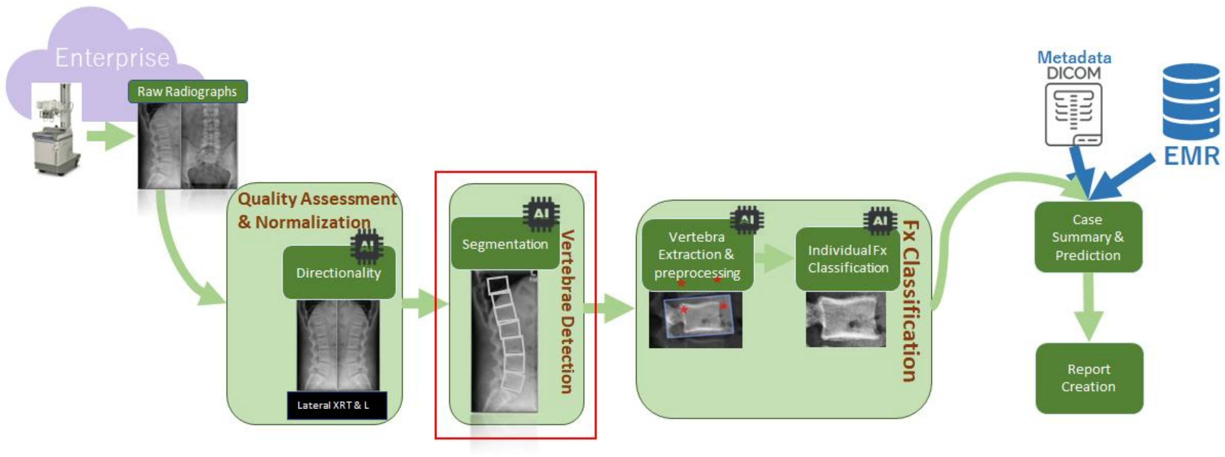
Supplemental Table 2 Definitions for Classifying Gross Position. *GT = ground truth. VB centroid refers to the centroid of a predicted VB, either from U-Net or Mask-RCNN, as both were evaluated in the same manner. All definitions are with respect to x (horizontal)- and y (vertical)-axes of images when referring to minimum or maximum.*

Ensembling Predictions

The ensemble model is a rule-based method as follows. Centroids from U-Net and Mask-RCNN predictions were pooled with the intent to maximize the number of correctly detected VBs in an image from both models: the distance between each U-Net and Mask-RCNN predicted centroid was calculated. First the minimum distance for each centroid was compared to a threshold, half the average endplate width per image. If this distance was less than this threshold, the corresponding centroids were considered “matching”, or the same centroid. To pool the matches and non-matches together, the centroids not labeled as matches from the U-Net predictions were concatenated on a per-radiograph basis to all the Mask-RCNN predictions. Thus, the final pool of centroids included the centroids labeled as matches and non-matches from the Mask-RCNN predictions plus those labeled as non-matches from the U-Net predictions. In this fashion we aggregated the centroids detected by both models without repeats from those detected by only one model.

A Schema for the Full Pipeline

Supplemental Figure S2 describes the proposed fracture detection pipeline in further detail and shows how this work fits into the workflow. Images from the clinical data warehouse would be standardized, then segmented, then classified as fractured or non-fractured, aggregated at the patient level, and finally a human-readable report would be generated.



Supplemental Figure S2. *The proposed imaging analysis pipeline comprises independent, sequential phases, or steps – pulling DICOM radiograph exams from* *the database, quality assessment & normalization, vertebral body detection, fracture classification, and case summary & prediction, ultimately generating a patient-level report. In this paper, we focus on vertebral body detection, outlined in red.*

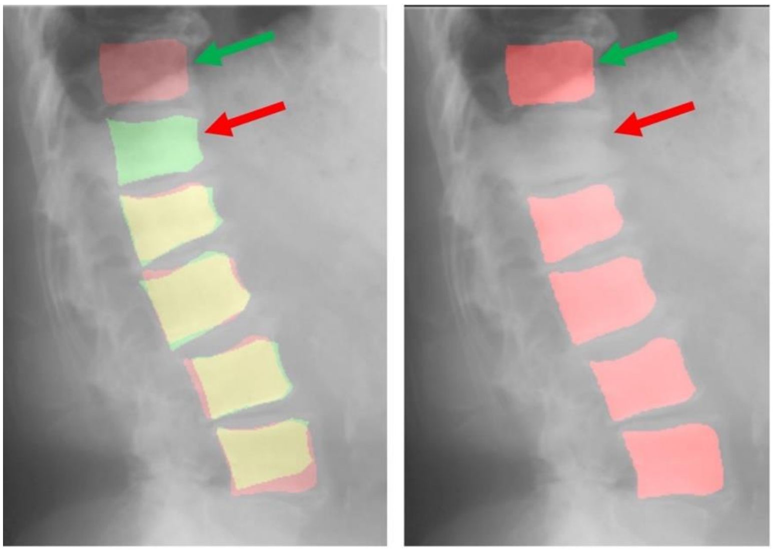
Bootstrap Methods for Confidence Intervals

To calculate confidence intervals for precision, recall, and F1 score for the train/validation/test sets separately, we performed bootstrapping with replacement on a per radiograph basis. A total of 100 bootstrap iterations were run, each with 200 radiographs. The bootstrap analysis for the subset of fractured VBs used 100 iterations and 20 samples owing to the smaller size of that subset. The confidence intervals were calculated as mean ± 2*standard error for precision, recall, and F1 score, respectively.

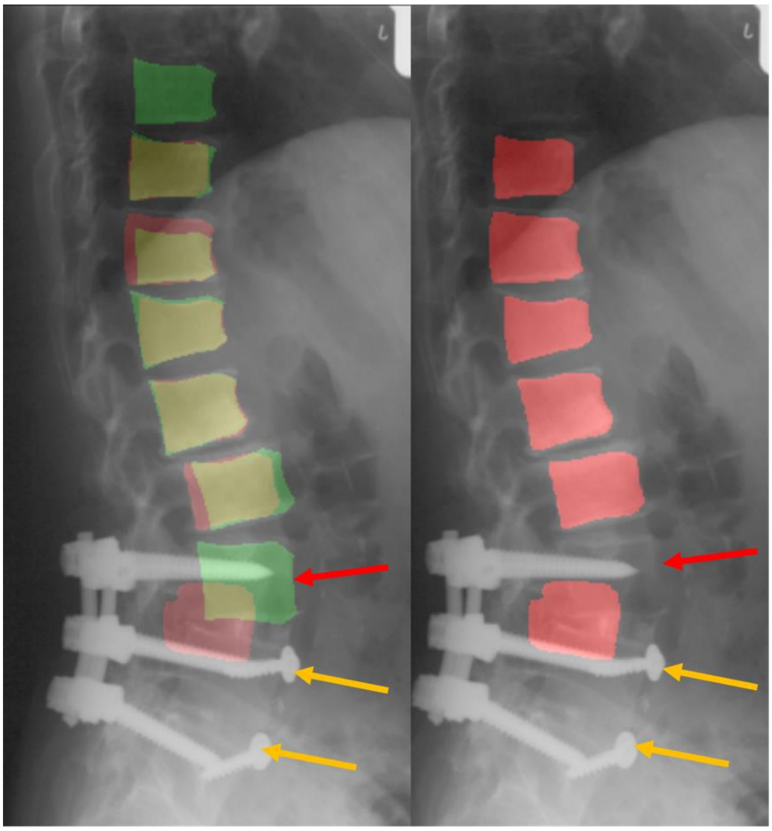
	Total			***Fractured Vertebrae***			***Non-Fractured Vertebrae***		
False Negative	U-Net	Mask-RCNN	Ensemble	U-Net	Mask-RCNN	Ensemble	U-Net	Mask-RCNN	Ensemble
Top	224	180	146	0	3	0	224	177	146
Gap	24	14	9	4	0	0	20	14	9
Bottom	243	220	201	4	2	2	239	218	199
Off	0	0	0	0	0	0	0	0	0
Total	491	414	356	8	5	2	483	409	354
False Positive									
Top	1	3	4	0	0	0	1	3	4
Gap	2	0	2	0	0	0	2	0	2
Bottom	1	2	2	0	0	0	1	2	2
Off	0	16	16	0	0	0	0	16	16
Total	4	21	24	0	0	0	4	21	24

Supplemental Table 3 Tabulation of False Positives and False Negatives by Gross Position. *Note the gross majority of false negatives are categorized as Top or Bottom, at the extrema of the spinal column.*

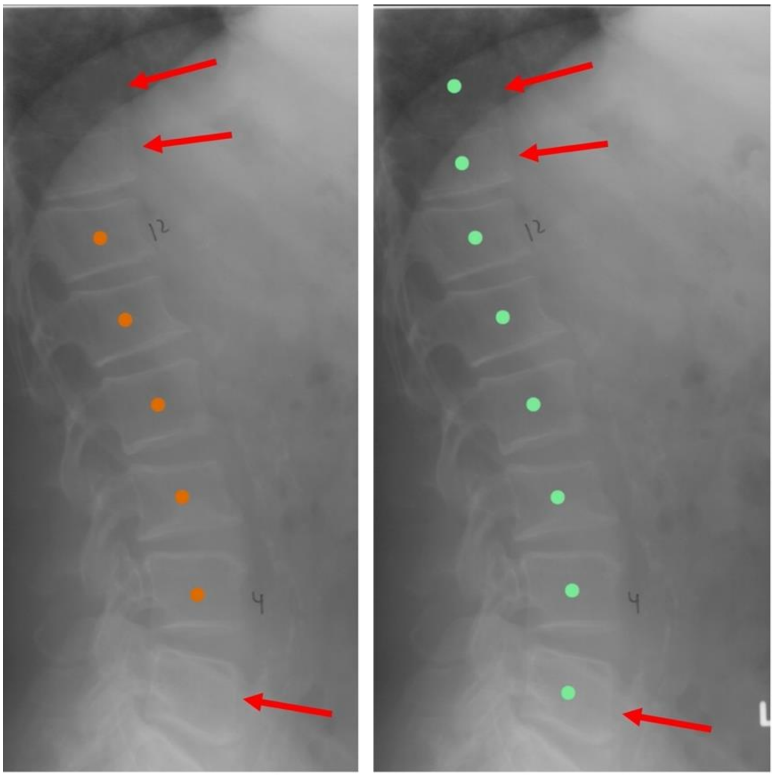
S3A) False negative (missed) detection in the middle of the inferred spinal column



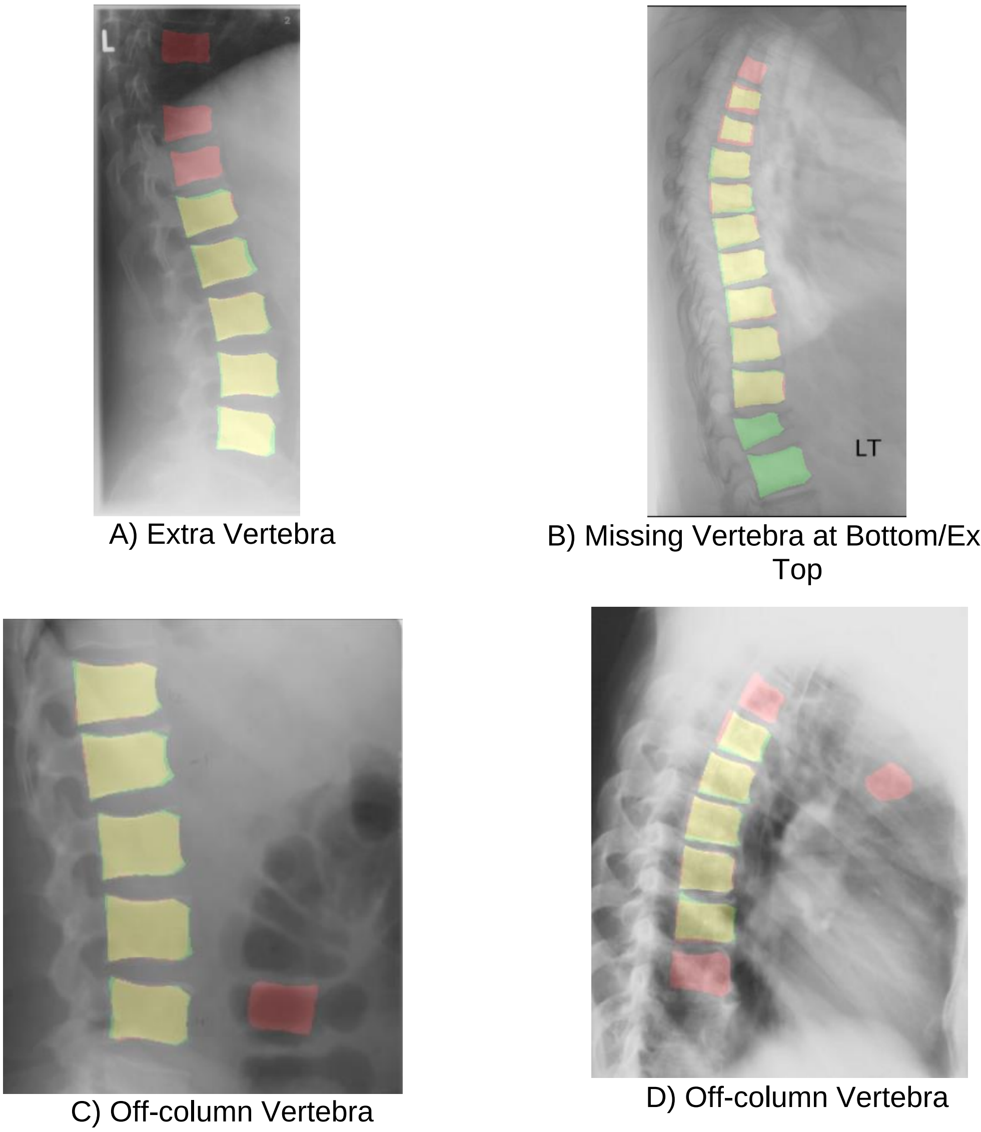
S3B) False negative (missed) detections of VBs with hardware



S3C) False negative (missed) detections at superior and inferior ends of radiograph



Supplemental Figure S3: ***S3A)*** *LEFT: Ground truth (green) masks are overlayed on Mask-RCNN prediction (red) masks. Overlaps are shown in yellow. RIGHT: Predicted (red) masks from Mask-RCNN. Example of a false negative (missed) detection: a ground truth annotated VB is not detected by Mask-RCNN, denoted by the red arrow. Example of a VB that was not annotated on ground truth is detected by Mask-RCNN, denoted by the green arrow.* ***S3B:*** *LEFT: Ground truth (green) masks are overlayed on Mask-RCNN predicted (red) masks. Overlaps are shown in yellow. RIGHT: Predicted (red) masks from Mask-RCNN. Example of VB with hardware that was not detected by Mask-RCNN, denoted by the red arrow. VBs with hardware that were not annotated on the ground truth radiograph is denoted by the gold arrow.* ***S3C:*** *LEFT: Ensemble centroids. RIGHT: Ground truth centroids from subset of complete annotations. Example of false negative (missed) detections denoted by the red arrows. For all figures, black boxes were overlayed to hide patient identifiable information.*



A) Extra Vertebra

B) Missing Vertebra at Bottom/Extra at Top

C) Off-column Vertebra

D) Off-column Vertebra

<u>Supplemental Figure S4.</u> *On qualitative review of vertebral body segmentation, several systematic sources of error were identified. It was not uncommon for VBs to not be segmented in the region of the diaphragm demonstrated in (A) which could be due to the strong contrast gradient, above, and below the diaphragm. Since the training data did not always catch all the vertebra on and off the image and left out vertebra at the top and bottom of the radiograph, this seems to carry forward into the inferences from these models, shown in (B and D). Very rarely, the model would identify a vertebral body outside of the spinal column, and it seemed to occur in areas where there is bowel gas (C) and in areas in the anterior middle mediastinum between ribs, almost as if it thought the edge of the ribs represented the superior and inferior endplate of a vertebral body. For all images, ground truth (green) masks are overlayed on predicted (red) masks. Overlaps are shown in yellow.*