SSA12

Science Session with Keynote: Informatics (Artificial Intelligence: Cutting Edge Artificial Intelligence)

Sunday, Dec. 1 10:45AM - 12:15PM Room: E450A

AI IN

AMA PRA Category 1 Credits ™: 1.50
ARRT Category A+ Credit: 1.75

FDA Discussions may include off-label uses.

Participants
George L. Shih, MD, New York, NY (*Moderator*) Consultant, MD.ai, Inc; Stockholder, MD.ai, Inc;
Marc Zins, MD, Paris Cedex 14, France (*Moderator*) Nothing to Disclose
Ciprian N. Ionita, PhD, Buffalo, NY (*Moderator*) Grant, Canon Medical Systems Corporation;
Ian Pan, MA , Providence, RI (*Moderator*) Consultant, MD.ai

Sub-Events

SSA12-01 Informatics Keynote Speaker: The French Radiology AI Data Hub

Sunday, Dec. 1 10:45AM - 10:55AM Room: E450A

Participants
Marc Zins, MD, Paris Cedex 14, France (*Presenter*) Nothing to Disclose

SSA12-02 FalcoNet-GMC: A 3D Convolutional Neural Network Module for Instance Segmentation and Quantification of Distant Recurrence from Gynecological Cancers

Sunday, Dec. 1 10:55AM - 11:05AM Room: E450A

Participants
Shih-Chun Cheng, Taoyuan, Taiwan (*Abstract Co-Author*) Nothing to Disclose
Pieh-Hsu Wang, MD, Taoyuan, Taiwan (*Abstract Co-Author*) Nothing to Disclose
Yi-Chin Tu, Taipei, Taiwan (*Abstract Co-Author*) Nothing to Disclose
Gigin Lin, MD, Taoyuan, Taiwan (*Presenter*) Nothing to Disclose

For information about this presentation, contact:

giginlin@cgmh.org.tw

CONCLUSION

A multifunctional web-based auxiliary system for distant recurrence from gynecologic cancer will enhance the early detection for salvage treatment, with better segmentation by compartment weight maps.

Background

Early detection of distant recurrence in the lung and thoracic lymph nodes is critical for successful salvage treatment for gynecological cancers. We introduce a novel web-based auxiliary system for ovarian and cervical cancer spread to chest, featuring: (1) A novel 3D convolutional neural network (CNN) named CompartNets for delineation of lung/lymph node metastases; (2) A 3D CNN for segmentation of lung, trachea, esophagus, heart, aorta, and spinal cord; (3) A 3D-based radiomic quantification module, VQM (Volumetric Quantification Master).

Evaluation

Contrast-enhanced CT of 40 ovarian cancers and 60 cervical cancers with distant recurrence were recruited as training dataset. Two board-certified radiologists manually delineated the lesion contours as ground truth. A 3D ESPNet model was trained for organ segmentation to generate compartment weight maps. Second, two 3D RetinaESPNets were pretrained on The Cancer Imaging Archive (TCIA). Transfer learning was conducted to detect distant recurrence. Independent testing was carried out in randomly selected 8 ovarian cancer and 12 cervical cancer patients. The segmentation of CompartNet was compared with pure 3D RetinaESPNet without compartment maps and pure 3D ESPNet without detection. The recall/precision reached 97%/93% for pulmonary recurrence and 91%/87% for nodal recurrence, respectively. For segmentation of lung/lymph node recurrence, the intersection over union (IoU) score of CompartNets reached 0.93/0.93, compared with 0.89/0.88 of pure RetinaESPNets and 0.77/0.77 of pure ESPNet. The mean IoU of organ segmentation was 0.93.

Discussion

The IoUs of CompartNets are improved compared with pure RetinaESPNets, owing to loss weighting of normal compartments, which are indecipherable within tumor bounding box. With organ segmentation and lesion-based VQM, our system can differentiate locations of metastases between mediastinum, lung, and chest wall.

SSA12-03 Automated Detection of Vertebral Fractures in CT Using 3D Convolutional Neural Networks

Sunday, Dec. 1 11:05AM - 11:15AM Room: E450A

Participants

Joeri Nicolaes, Anderlecht, Belgium (*Presenter*) Computer Scientist, UCB Pharma; Stockholder, UCB Pharma
David Robben, Leuven, Belgium (*Abstract Co-Author*) Employee, icoMetrix NV
Guido E. Wilms, MD, Leuven, Belgium (*Abstract Co-Author*) Nothing to Disclose
Dirk Vandermeulen, MSc, Leuven, Belgium (*Abstract Co-Author*) Nothing to Disclose
Cesar Libanati, Anderlecht, Belgium (*Abstract Co-Author*) Employee, UCB SA; Stockholder, UCB SA
Marc DeBois, Anderlecht , Belgium (*Abstract Co-Author*) Nothing to Disclose
Steven Raeymaeckers, Jette, Belgium (*Abstract Co-Author*) Nothing to Disclose

For information about this presentation, contact:

Joeri.Nicolaes@ucb.com

joeri.nicolaes@ucb.com

CONCLUSION

Our method achieves an AUC of 0.95±0.02 outperforming Valentinitsch et al. We also illustrate that our method achieves higher recall (0.905) on the operating point reported by Bar et al. The results of our 5-fold cross-validation experiment demonstrate that our 3D data-driven method compares favourably to state-of-the-art using 2.5D learned features and 3D engineered features. The small sample size and use of cross-validation are limitations of this proof-of-concept. This will be adressed in a larger follow up study, currently ongoing.

Background

We present a data-driven approach to automatically detect vertebral fractures in spine-containing CT images. Inspired by radiology practice, existing methods are based on 2D and 2.5D features but we present, to the best of our knowledge, the first method learning 3D features for detecting vertebral fractures.

Evaluation

For this study, we build a training database of 90 de-identified CT image series. These images were acquired on three different scanners (Siemens, Philips and General Electric; 120 kVp tube voltage; maximum in-plane spacing and slice thickness are respectively 0.92mm x 0.92mm and 1.5mm) and contain 90 patients scanned for various indications (average age: 81 years, range: 70 - 101 years, 64% female patients, 12% negative cases). We present a two-staged vertebra fracture detection method that first predicts a class probability for every voxel using a 3D CNN and secondly aggregates this information to a patient-level fracture prediction.

Discussion

We performed a stratified 5-fold cross-validation to estimate the expected performance of our 3D method. For each run, we selected 15% of the images in the training folds as validation samples to determine when to stop training based on validation performance. We report the ROC curve because this metric describes model performance independently of the class distribution and is well suited to compare results from different test sets. Since our method involves two hyperparameters that can be chosen to deliver distinct classifiers, we build the ROC curve using the convex hull representing the optimal classifiers from a group of potential classifiers.

SSA12-04 Universal High Performance Pelvic/Hip Fracture Detection on Pelvic Radiographs of Trauma Patients Using Cascaded Deep Networks

Sunday, Dec. 1 11:15AM - 11:25AM Room: E450A

Participants
Chi-Tung Cheng, MD, Taoyuan City, Taiwan (*Presenter*) Nothing to Disclose
Chien-Hung Liao, MD, Taoyuan City, Taiwan (*Abstract Co-Author*) Nothing to Disclose
Yirui Wang, MS, Rockville, MD (*Abstract Co-Author*) Nothing to Disclose
Shun Miao, Bethesda, MD (*Abstract Co-Author*) Nothing to Disclose
Dakai Jin, MS, Bethesda, MD (*Abstract Co-Author*) Nothing to Disclose
Le Lu, Bethesda, MD (*Abstract Co-Author*) Nothing to Disclose
Chih Chen Chang, MD, Taoyuan, Taiwan (*Abstract Co-Author*) Nothing to Disclose
Jing Xiao, Shenzhen, China (*Abstract Co-Author*) Nothing to Disclose

For information about this presentation, contact:

atong89130@gmail.com

PURPOSE

Detecting fracture from pelvic radiographs is an important yet challenging task because of the high variety of possible fracture types in hip and pelvis. Existing detection methods only detect hip fracture and do not address the more complex pelvic fracture. We propose a universal fracture detector via two-stage cascaded deep neural network that is capable of handling both pelvic and hip fractures.

METHOD AND MATERIALS

Our method is developed using 4,410 pelvic radiographs (1,975 hip fractures, 801 pelvic fractures and 1,630 images without fractures) with only image level fracture labels. The first stage deep network mines the potential fracture regions of interest (ROIs) from the whole image globally, which are then examined locally by the second network to classify the fracture and type (i.e., pelvic/hip fracture) per local ROI. A fracture probability heatmap is produced to indicate potential fracture sites. We recruit 17 primary physicians (emergency physician and surgeon) and 6 consulting physicians (orthopedic specialist and radiologist) to review an independent test dataset of 150 pelvic radiographs (50 hip fractures, 50 pelvic fractures, and 50 without fractures), and compare the detection results from the physicians with the proposed model.

RESULTS

Our two-stage 'global-to-local' cascaded deep neural network achieves an accuracy of 0.907 in the independent testing dataset, which is comparable with the primary physicians (0.882, IQR[0.863-0.923]), but slightly lower than consulting physicians. The model sensitivity and specificity are 0.962 and 0.938 for the hip fracture, 0.842 and 0.953 for the pelvic fracture. For all physicians, the model may avoid 2.17% missed hip fracture, and 7.74% missed pelvic fracture. For the primary physicians, the model may avoid 2.82% missed hip fracture, and 9.18% missed pelvic fracture. The fracture heatmaps consistently show correct fracture sites for true positive detection cases.

CONCLUSION

We propose an algorithm to detect pelvic and hip fractures from pelvic radiographs. It consistently and significantly outperforms previous automated methods and is a promising tool to avoid misdiagnosis by physicians.

CLINICAL RELEVANCE/APPLICATION

Our method provides accurate diagnosis of both hip and pelvic fractures in pelvic radiographies. It also produces fracture location heatmap to assist physicians to identify potential fracture sites.

SSA12-05 Defacing Neuroimages

Sunday, Dec. 1 11:25AM - 11:35AM Room: E450A

Participants
Daisy T. Kase, Sao Paulo, Brazil (*Abstract Co-Author*) Nothing to Disclose
Osvaldo Landi Junior, MD, Sao Paulo, Brazil (*Abstract Co-Author*) Nothing to Disclose
Marcelo Arcuri, MD, Sao Paulo, Brazil (*Abstract Co-Author*) Nothing to Disclose
Felipe C. Kitamura, MD, MSC, Sao Paulo, Brazil (*Presenter*) Consultant, MD.ai, Inc
Ian Pan, MA , Providence, RI (*Abstract Co-Author*) Consultant, MD.ai
Neil Tenenholtz, PhD, Boston, MA (*Abstract Co-Author*) Employee, Microsoft Corporation
George L. Shih, MD, New York, NY (*Abstract Co-Author*) Consultant, MD.ai, Inc; Stockholder, MD.ai, Inc;
Leon Chen, New York , NY (*Abstract Co-Author*) Nothing to Disclose
Anouk Stein, MD, Paradise Valley, AZ (*Abstract Co-Author*) Consultant, MD.ai, Inc; Stockholder, MD.ai, Inc
Nitamar Abdala, MD, PhD, Mogi Das Cruzes, Brazil (*Abstract Co-Author*) Nothing to Disclose

For information about this presentation, contact:

osv.landi@gmail.com

CONCLUSION

The present model was trained to deface brain CTs and also worked well for FLAIR images. The image binarization preprocessing step shows a promising approach of "train one, earn many", that could potentially be extended to all MRI sequences, not only FLAIR. This work demonstrates the use of AI to protect patient privacy, allowing the use of head CTs in open databases for worldwide collaboration in machine learning projects.

Background

The Health Insurance Portability and Accountability Act (HIPAA) defines 18 identifiers as protected health information that need to be removed from healthcare exams, including 'full face photographic images and any comparable images'. This raises the concern about the possibility of patient identification by 3D rendering of head CTs or brain MRIs. There is a need for sharing imaging data for open collaboration while ensuring the patient's privacy. In this context, image de-identification has become a necessity. We propose a two-step deep learning model to automatically deface head CTs and brain MRIs.

Evaluation

This study was approved by our institutional review board, and written informed consent was waived. A total of 1123 axial brain CT studies were anonymized. Manual segmentation of the face was done using bounding boxes in each slice using md.ai (md.ai, New York). Masks were generated from the bounding boxes and the CTs werebinarized. The first model consists of a binary classification (NASNet mobile) that predicts if that slice contains a face or not. The second step is a Unet trained to segment the face only in the slices that had faces. The final performance was evaluated with AUC, Dice Similarity Coefficient and visual inspection. The same model trained on binarized CTs was tested on FLAIR (630 studies) and in an external batch of CTs (500 studies).

Discussion

Step one resulted in a model with an AUC of 0.999 in the test set. Step two resulted in a Dice Coefficient score of 0.97/0.93/0.91 in the train/validation/test sets, respectively. Visual inspection of the head CTs from the test set and the external batch showed 100% defacing and on FLAIR resulted in 99.5% defacing.

SSA12-06 Automated Detection and Delineation of Hepatocellular Carcinoma on Multiphasic Contrast-Enhanced MRI Using Deep Learning

Sunday, Dec. 1 11:35AM - 11:45AM Room: E450A

Participants
Khaled Bousabarah, MSc, Duesseldorf, Germany (*Presenter*) Nothing to Disclose
Brian S. Letzen, MD, Orange, CT (*Abstract Co-Author*) Nothing to Disclose
Jonathan Tefera, New Haven, CT (*Abstract Co-Author*) Nothing to Disclose
Isabel T. Schobert, BS, New Haven, CT (*Abstract Co-Author*) Nothing to Disclose
Lynn J. Savic, MD, New Haven, CT (*Abstract Co-Author*) Nothing to Disclose
Todd Schlachter, MD, New Haven, CT (*Abstract Co-Author*) Research Grant, Guerbet SA
Julius Chapiro, MD, New Haven, CT (*Abstract Co-Author*) Research Grant, Guerbet SA; Consultant, Guerbet SA; Research Grant, Koninklijke Philips NV; Consultant, Koninklijke Philips NV; Research Grant, Boston Scientific Corporation;
Ming de Lin, PhD, North Haven , CT (*Abstract Co-Author*) Employee, Visage Imaging, Inc; Former Employee, Koninklijke Philips NV

For information about this presentation, contact:

julius.chapiro@yale.edu

PURPOSE

The Liver Imaging Reporting and Data System (LI-RADS) uses multiphasic contrast-enhanced (CE) imaging for diagnosis of hepatocellular carcinoma (HCC). In order to make the workflow more efficient and to provide a first benchmark for this modality, a deep learning algorithm was trained to segment the liver and HCC based on CE-MRI.

METHOD AND MATERIALS

A single deep convolutional neural network (DCNN) for liver segmentation and HCC delineation was trained on late arterial (25-30s), portal venous (60-70s) and delayed phase (3 min) CE-MRI. The U-Net was chosen as the DCNN's architecture and recent optimizations (residual blocks, Leaky ReLUs, instance normalization) were adopted. The network was presented with stacks of adjacent axial slices across the three phases. The U-Net was trained (70%), validated (15%) and tested (15%) on a dataset consisting of 174 patients with 231 lesions. Manual 3D segmentations of the liver and HCC made by a board-certified radiologist served as ground truth. The dice similarity coefficient (DSC) was measured between the manual and automated methods. In addition to the U-Net, a random forests classifier employing radiomic features (RF) and thresholding (TR) the mean activation of a segmentation were used to reduce the false positive rate (FPR).

RESULTS

The algorithm detected 73% and 75% of HCC on validation and test sets, respectively, using a DSC criterion between the individual lesion and corresponding segmentation of >0.2. The FPR on the validation set were 2.81, 0.77, and 0.85 for the U-Net, U-Net+RF, and U-Net+TR, respectively. A combination of all methods (U-Net+RF+TR) further improved the FPR to 0.62 and on the test set, it was 0.75. Mean DSC/case was 0.49 and 0.48 on validation/test. Mean DSC between detected lesions and corresponding segmentation was 0.64/0.68. Liver segmentations had a mean DSC of 0.91/0.91.

CONCLUSION

Our results are comparable to studies using monophasic CT by Vorontsov et al., and Chlebus et al. They achieved a higher DSC per case (0.66/0.58) whereas our model was more sensitive (0.66/0.57) and could be used to identify regions of interest which an experienced radiologist could either discard or flag for further inspection.

CLINICAL RELEVANCE/APPLICATION

DCNNs are capable of supporting radiologists by segmenting the liver and identifying potential HCCs automatically. This could enable a more workflow efficient and clinically realistic implementation of LI-RADS.

SSA12-07 Constructing a Platform Based on Deep Learning Model to Mimic the Self-Organization Process of CT Images Order for Automatically Recognizing Human Anatomy

Sunday, Dec. 1 11:45AM - 11:55AM Room: E450A

Participants
Feng-Mao Lin, New Taipei , Taiwan (*Presenter*) Nothing to Disclose
Chi-Wen Chen, New Taipei, Taiwan (*Abstract Co-Author*) Nothing to Disclose
Wei-Da Huang, New Taipei, Taiwan (*Abstract Co-Author*) Nothing to Disclose
Liangtsan Wu, Milwaukee, WI (*Abstract Co-Author*) Nothing to Disclose
Anthony Costa, PhD, New York, NY (*Abstract Co-Author*) Nothing to Disclose
Eric K. Oermann, MD, New York, NY (*Abstract Co-Author*) Nothing to Disclose
Weichung Wang, PhD, Taipei, Taiwan (*Abstract Co-Author*) Nothing to Disclose

For information about this presentation, contact:

kiralintw@gmail.com

PURPOSE

To demonstrate the ability of a deep learning application to automatically identify computed tomography (CT) slice regions by major Human anatomy. This application will be deployed in National Health Insurance of Taiwan (NHI) to classify the around 458 million CT images in 2018.

METHOD AND MATERIALS

954 and 4095 CT series were selected for training and testing correspondingly from NHI and TCIA. The voxel spacing must > 0.6 mm, and series must > 40 slices. Each image was standardized to 128^2 pixels. The AlexNet and ResNet was trained with grey scale images and the 3 color images (bone, liquid and air), correspondingly. The loss function is identical to Ke Yan, and et al. in 2018 and guides slice scores increased by slice order. Linear regression was used to adjust slice score of a series which the r-square < 0.8. The series was split into 4 parts and new slice score was estimated from two of the best parts. Manually annotated lung boundary was used to find the cutoff for measuring sensitivity and specificity.

RESULTS

The AlexNet and ResNet was trained for 2 days. The r-square of linear regression was to measure the linearity between slice score and its order. The amount of series with r-square < 0.8 was reduced from 4.1% to 1.7% in AlexNet and 6.8% to 2.2% in ResNet by using our error correction approach. Fig. 1 depicted the images with similar slice score having similar body part. Based on lung boundary, the score variant of lower boundary was larger than upper boundary. The cutoff was selected based on the highest value of specificity*sensitivity. ResNet had the best prediction performance in training data and validation data (Spec. > 0.94, Sens. > 0.9). AlexNet provided the best prediction performance in NHI validation data (Spec. > 0.91 and Sens. > 0.94). The error correction slightly improved the sensitivity and specificity. The specificity and sensitivity were both larger than 0.9 in NHI validation data by using AlexNet and ResNet.

CONCLUSION

First, the preprocessing process could accelerate training process and reach lower loss by using ResNet and AlexNet is efficient in during the prediction. Fig 2. showed our error correction process sucessfully adjusting slice score to corresponding body part. Since the organ boundary was varied from person to person, this approach is good for large part Identification. Although we found ResNet and error correction could provide good prediction quality with small training data, the model proposed by Ke Yan, and et al. in 2018 trained with large training data is one of the state of art methods.

CLINICAL RELEVANCE/APPLICATION

NHI collected around 458 million medical CT images in 2018. Our application will deploy in one of the largest medical databases in the world. Precisely retrieve the certain images of Human Anatomy could accelerate related application development and reduce the storage usage.

SSA12-08 An Ensemble of Models with a Multi-Threshold Approach to Improve Chest X-Ray Predictions

Sunday, Dec. 1 11:55AM - 12:05PM Room: E450A

Participants
Jessica d. de Oliveira, MSc, Sao Paulo , Brazil (*Presenter*) Employee, NeuralMed
Maria Fernanda B. Wanderley, DSc, Sao Paulo, Brazil (*Abstract Co-Author*) Employee, Neuralmed
Vitor De Mario, Sao Paulo , Brazil (*Abstract Co-Author*) Employee, NeuralMed
Andre C. Castilla, MD,PhD, Sao Paulo, Brazil (*Abstract Co-Author*) Stockholder, Neuralmed
Anthony Eigier, BA, Sao Paulo, Brazil (*Abstract Co-Author*) Stockholder, NeuralMed

For information about this presentation, contact:

info@neuralmed.ai

PURPOSE

Our main goal is to assess if deep learning can decrease the list of exams that radiologists need to read, with minimal loss of critical cases. We propose an ensemble with a multi-threshold approach, focusing on the detection of general opacities.

METHOD AND MATERIALS

We use four public datasets: JSTR, OpenI, Shenzen, and Chest-Xray14. After removing some lateral and low quality images, the total amount of images were 117,094 images. We cut the images surrounding the lung mask predicted with a trained U-net, applied a Limited Adaptive Histogram Equalization (CLAHE), resized to 384x384 and normalized based on the mean and standard deviation of images in the ImageNet. Then we developed three models: M1: a binary classifier to detect if an image has some finding or if it is normal M2: a multilabel trained with all images to predict five classes: mass/nodule, edema, atelectasis, alveolar opacity, and non-opacity. M3: a multilabel to predict the same five classes, but without the normal images in the training set. All of them use Inception V4. The ensemble was created using a weighted average in the form: (4*ym1 + 3*ym2 + 3*ym3)/10. We calculate the AUC of ROC Curve and choose two best cut-points using Youden's index.

RESULTS

The mean F1 Score of our model is 0.478 among all classes with an AUC of 0.90 for mass/nodule, 0.86 for edema, 0.85 for atelectasis, 0.86 for alveolar opacity and 0.93 for nonopacity. Analyzing the predictions, we saw that normal images had lower values, the target classes had high values, and in the middle values were images of other pathologies. This justifies the use of two thresholds. With the two thresholds, the general quality of our model is improved. We correctly classified more than 70% of all normal images with just 5% of False Negative Rate (FNR) and the average True Positive Rate (TPR) is 44% in the target classes.

CONCLUSION

The image preprocessing along with the use of ensemble and multi thresholds techniques produced a model with greater certainty and better results.

CLINICAL RELEVANCE/APPLICATION

We can accelerate the radiologist's work by detecting 70% of normal images, decreasing the number of images analyzed and suggesting the pathology according to what was predicted.

SSA12-09 CT Organ Segmentation: Use of Variational Autoencoders to Detect Incorrect Segmentations in a Large Dataset (> 12,000 CT Scans)

Sunday, Dec. 1 12:05PM - 12:15PM Room: E450A

Participants
Veit Sandfort, MD, Bethesda, MD (*Presenter*) Nothing to Disclose
Ke Yan, Bethesda, MD (*Abstract Co-Author*) Nothing to Disclose
Peter Graffy, Madison, WI (*Abstract Co-Author*) Nothing to Disclose
Perry J. Pickhardt, MD, Madison, WI (*Abstract Co-Author*) Stockholder, SHINE Medical Technologies, Inc; Stockholder, Elucent Medical; Advisor, Bracco Group;
Ronald M. Summers, MD,PhD, Bethesda, MD (*Abstract Co-Author*) Royalties, iCAD, Inc; Royalties, Koninklijke Philips NV; Royalties, ScanMed, LLC; Royalties, Ping An Insurance Company of China, Ltd; Research support, Ping An Insurance Company of China, Ltd; Research support, NVIDIA Corporation; ; ; ;

PURPOSE

Organ segmentation on CT using neural networks is highly effective but training labels are expensive and insufficient training data impairs performance. Typically, organ segmentation datasets have <400 CTs. Many pathologies and variations are not captured in such small training data. This may explain the gap between theoretical and real-world performance. Images which differ from training data can cause surprisingly severe failures of the algorithms.We hypothesize that failed segmentations can be detected using variationalautoencoders (VAE) without supervision.

METHOD AND MATERIALS

The Segmentation Decathlon data and internal data were used for training (n of 131,41 and 56 for liver, spleen and kidney). CT colonography scans from a large cohort (n=12495) were used for training and testing the autoencoder. A modified 3DUnet was trained on the labeled data. Organ segmentations were performed on 12495 CTs. For each organ a 3D variational autoencoder was trained on all segmentations in an unsupervised fashion. Then, the organ segmentations were passed through the variational autoencoder and the reconstruction error (Dice score) was measured. Organ segmentations (n=2510x3) were visually assessed for significant error by a physician. ROC curves and AUC for detection of failed segmentations were calculated.

RESULTS

The reconstruction errors of the autoencoder were 0.87, 0.76 and 0.81, for liver, spleen and kidney, respectively. Of the reviewed segmentations, 1.6-4.9% showed significant errors. The variational autoencoder reconstruction error was highly effective in detecting problematic segmentations, evidenced by AUCs of 0.94, 0.87 and 0.9 for liver, spleen and kidney (for ROC curve and an example see figure, note that the erroneous area [arrow] is not present in the autoencoder output).

CONCLUSION

The use of deep learning based segmentation in medical imaging has been increasing rapidly. These algorithms are powerful but not very robust in regard to cases with unexpected characteristics and this may cause catastrophic failure of the algorithm.We show that our method can detect failed segmentations effectively (AUCs 0.87-0.94).This is useful for continuous quality monitoring and for active learning.

CLINICAL RELEVANCE/APPLICATION

Deep learning methods are vulnerable to failure when confronted with unexpected cases. This is a critical issue for clinical uses. Our method has the potential to detect such failure without supervision.

Printed on: 01/18/21