Received: 27 October 2020 | Revised: 22 January 2021 | Accepted: 16 February 2021

DOI: 10.1002/mrm.28768

FULL PAPER

Magnetic Resonance in Medicine

Automated renal segmentation in healthy and chronic kidney disease subjects using a convolutional neural network

Alexander J. Daniel | **Charlotte E. Buchanan** | **Thomas Allcock** | **Daniel Scerri** | **Eleanor F. Cox** | **Benjamin L. Prestwich** | **Susan T. Francis**

Sir Peter Mansfield Imaging Centre, University of Nottingham, Nottingham, United Kingdom

Correspondence
Susan T. Francis, Sir Peter Mansfield Imaging Centre, University Park, University of Nottingham, Nottingham, NG7 2RD, United Kingdom.
Email: susan.francis@nottingham.ac.uk

Funding information
This work was co-funded by the Animal Free Research UK and the Medical Research Council grant (MR/R02264X/1). A.J.D. is supported by a studentship from the Oxford Nottingham Biomedical Imaging Centre for Doctoral Training funded by the Engineering and Physical Sciences Research Council (EPSRC) and Medical Research Council (MRC) (EP/L016052/1)

Purpose: Total kidney volume (TKV) is an important measure in renal disease detection and monitoring. We developed a fully automated method to segment the kidneys from T_2-weighted MRI to calculate TKV of healthy control (HC) and chronic kidney disease (CKD) patients.

Methods: This automated method uses machine learning, specifically a 2D convolutional neural network (CNN), to accurately segment the left and right kidneys from T_2-weighted MRI data. The data set consisted of 30 HC subjects and 30 CKD patients. The model was trained on 50 manually defined HC and CKD kidney segmentations. The model was subsequently evaluated on 50 test data sets, comprising data from 5 HCs and 5 CKD patients each scanned 5 times in a scan session to enable comparison of the precision of the CNN and manual segmentation of kidneys.

Results: The unseen test data processed by the 2D CNN had a mean Dice score of 0.93 ± 0.01. The difference between manual and automatically computed TKV was 1.2 ± 16.2 mL with a mean surface distance of 0.65 ± 0.21 mm. The variance in TKV measurements from repeat acquisitions on the same subject was significantly lower using the automated method compared to manual segmentation of the kidneys.

Conclusion: The 2D CNN method provides fully automated segmentation of the left and right kidney and calculation of TKV in <10 s on a standard office computer, allowing high data throughput and is a freely available executable.

KEYWORDS

convolutional neural network, kidney, machine learning, magnetic resonance imaging, segmentation

1 | INTRODUCTION

Segmentation of the kidneys from MRI is a time consuming aspect of many renal MRI studies.[1-3] Total kidney volume (TKV) gives insight into renal function and is therefore used as a measured parameter for a variety of renal pathologies. The use of TKV is an active area of ongoing research for autosomal dominant polycystic kidney disease (ADPKD), which is characterized by an increase in TKV as a result of cyst formation. Disease progression can be monitored by

recording TKV, with higher rates of TKV increase being associated with a more rapid decrease in renal function.[4-6] Measurements of TKV in chronic kidney disease (CKD) subjects have shown a significant correlation with glomerular filtration rate,[7] the primary measure of CKD severity,[8] with more generally a decrease in TKV associated with a decrease in renal function.[9] When studying pathologies, which commonly lead to a change in kidney function, total kidney perfusion is often measured, this metric relies on an accurate measurement of renal blood flow and kidney volume of each kidney, and allows investigators to ascertain if the blood flow is preserved as the organ changes in size or if tissue perfusion is impaired. In addition to TKV measurements, renal segmentation is an important first step in many other processing pipelines, for example, increasing the accuracy of cortical-medullary segmentations or reducing computation times by only fitting quantitative maps for voxels within the kidney.

The gold standards of kidney segmentation are manual (region-of-interest) ROI boundary tracing[10] or stereology[11] by experienced and skilled experts, with blood vessels in the kidney and the hilum excluded. These manual processes are highly time consuming (taking ~15-30 min per subject[12-14]) and can be biased by investigator judgement because of the similar signal intensities between the kidneys and surrounding organs, anatomical differences between subjects, cysts, and image artefacts. Consequently, the resulting kidney ROIs produced are subject to intra- and inter-expert variability as a result of the varying expertise levels; experts may segment a specific image differently when performed more than once, or different experts may segment the same image differently. These factors mean that the development of a faster and ideally fully automated method of renal segmentation is highly desirable. However, the same factors that make manual segmentation difficult can also limit fully automated methods, for example, the signal intensity of the kidneys closely matches that of other abdominal structures such as the spleen.

A number of automated methods have been proposed with varied success.[12] Some simply assume the kidney is an ellipse and calculate the volume from measurements of the pole-to-pole distance[15,16] or include a correction factor to reduce overestimations.[17] Unfortunately, these techniques produce a large confidence interval and still require human intervention to define the pole-to-pole length, a process that can produce inconsistencies between readers and takes a reasonable amount of time (~5 min).[18] Other semi-automated methods use classical image processing techniques such as thresholding,[19] water-shedding,[20] level sets,[14,21] and spatial prior probability mapping.[22] These methods can either be inaccurate, over-segmenting the kidneys, or include a number of parameters that need to be manually adjusted and are computationally intensive. Further, the fact that each technique is highly optimized for a specific data set means that it needs to be re-written to be applied to different pathology, which is another time consuming and highly skilled process.

Machine learning methods have the potential to automatically detect different patterns from data given to a model that has been trained. Deep learning is a class of machine learning algorithms that can model high-level information in an image using several processing layers of transformations. This uses an architecture of multi-level linear and non-linear operations, described by layers, to learn complex functions that can represent high-level detail to map the input data to the output segmentations directly. As more data becomes available the algorithm can become more accurate and generalized, without a need to rewrite the underlying methods, therefore making it a good choice for long-term development.

In recent years, deep learning-based methods have been applied to the segmentation of medical images, especially successful has been the U-Net.[23] This modified fully convolutional neural network (CNN) architecture uses a number of convolution, pooling, and upsampling layers to detect features in the input data at multiple resolutions. The convolution layers convolve a learnable kernel with the input data to generate spatial feature maps that are passed to subsequent layers in the network. By adjusting the kernels, the resulting feature maps can be optimized to detect the location of the kidneys. Pooling layers are used to downsample the data and allow some convolution kernels to become tuned to approximate features, this also reduces the tendency of the network to over-fit the training data. When the data has been fully downsampled, upsampling layers are used to increase the resolution of the feature maps back to that of the original data while more convolution layers also learn the precise location of the kidneys. Parameters are adjusted by comparing the output from the network to a known ground truth. CNN methods have been applied to segmentation in other areas of medical imaging,[24-27] for example, to prostate segmentation of MRI images,[28] liver segmentation of x-ray CT images[29] and segmentation of polycystic kidneys.[30-32] However, to date, these methods have not been successfully applied to CKD and healthy kidney segmentation from MR images.

Here a single 2D U-Net model CNN is used for the segmentation of the kidneys in both healthy control (HC) participants and CKD patients using T_2-weighted MR images. Automatically generated kidney masks are compared with manual masks defined by experts and assessed for similarity using multiple voxel and surface based metrics and total segmented volume. A subset of subjects was scanned multiple times to assess the repeatability of the segmentations.

2 | METHODS

The study was approved by the University of Nottingham Medical School Research Ethics Committee (H14082014

and E14032013), and East Midlands Research Ethics committee REC reference: 17/LO/2036 and 15/EM/0274.

2.1 | MRI data acquisition

All kidney MRI scans were acquired on a 3T Philips Ingenia system (Philips Medical Systems, Best, The Netherlands) using a 2D T_2-weighted half-Fourier single-shot turbo spin echo (HASTE) sequence optimized to achieve the maximum contrast between the kidneys and surrounding tissue TE = 60 ms, TR = 1300-1800 ms, SENSE factor = 2.5, refocus angle 120°, bandwidth, 792 Hz, FOV = 350 × 350 mm^2, voxel size = 1.5 × 1.5 × 5 mm^3 and a slice gap of 0.5 mm with approximately 13 coronal slices, enough to image the entire kidney,[33,34] in a single 17- to 23-s breath-hold.

The data set consisted of 60 subjects, 30 HC (10 female, 20 male) with a mean age of 26 ± 11 (19-77) years and 30 CKD patients (6 female, 24 male) with a mean age of 59 ± 14 (19-80) years and mean CKD stage 3.5 ± 1.2 (1-5). Ten of the subjects (5 HCs and 5 CKD patients) were scanned 5 times in the same scan session for use as test data. In each test data scan session, subjects were repositioned between each acquisition (removed from the scanner, asked to sit up and move on the bed), additionally the scanner operator attempted to vary the acquisition geometry between each scan while still acquiring full kidney coverage. These repeated test data sets allow the consistency of the networks ability to measure TKV to be assessed.

In total, 649 2D image slices from the 50 subjects in the training data and 650 2D image slices from the 10 subjects in the test data were collected. A summary of the data collected is provided in Supporting Information Table S1 and Supporting Information Figure S1.

2.2 | Manual segmentation

The manual binary mask of the kidneys of each subject were generated by 1 of 3 observers (A, B, and C who had been trained on kidney segmentation and had an average of 2 years of experience), with each observer segmenting data from both the training and testing data sets. Kidney boundaries were manually traced using freely available software (MRIcron) and any area of non-renal parenchyma, such as the renal hilum and cysts, were excluded from the manual definition. Binary masks of the kidney were generated, and the volume of each kidney was computed from the product of the number of voxels in each kidney mask and the voxel volume. Separate kidney volume for the left and right kidneys was determined and summed to compute TKV. All measurements were performed by observers blinded for patient number and previous TKV measurements.

For the training phase, for each subject a manual mask was used from a single observer (randomized between observer A, B, or C). For the testing phase, all 5 scans from a given subject were segmented by a single reader with the 10 subjects being segmented by a mix of the 3 readers, that is, the test data comprised of subjects segmented by all readers but the repeat scans of each subject were segmented by the same reader. For 4 HC subjects from the test data set, manual masks were drawn by all 3 observers for all 5 repeat acquisitions to allow assessment of inter-observer variability in the manual masks. HCs were chosen for this analysis as they healthy kidneys have a more consistent morphology and therefore will give a best-case measure of observer variability and provide a comparison of the automated method to the highest standard of manual segmentation.

2.3 | Automated segmentation using convolutional neural network architecture

Voxel intensities were normalized between 0 and 255, where 0 was set to the mean voxel intensity minus 0.5 times the SD of that slice and 255 was set to the mean voxel intensity plus 4 times the SD of the volume. This empirically derived windowing led to a clear contrast between the kidneys and surrounding tissue while negating the effects of bulk signal changes between volumes. Each data set volume was then split into 2D coronal slices and resampled to a matrix size of 256 × 256. Twenty percent of slices were reserved for validation during the network optimization process, this validation data was used to monitor over-fitting and direct the optimization process between epochs. Once the data had been split into training and validation sets, the slice order was randomized within sets. Splitting the data before slice randomization limited the possibility of slices from only 1 subject being split over both the training and validation data sets. During training, data augmentation was applied. At the start of each epoch, a batch of images and their corresponding masks was selected at random from the training data and a series of random shifts (up to 25% of the image in both the horizontal and vertical direction), zooms (between 0.75 and 1.25 magnification), rotations (within a 20° range), and sheers (within a 5° range) were applied to the image/mask pair to produce different yet anatomically reasonable images. The weights of the network were then adjusted based on this augmented data before selecting a new batch of images for the next epoch. Augmenting the data reduces the tendency of a model to over-fit the training data and therefore increases accuracy when the model is applied to unseen images.

The U-Net consists of 2 fully CNN-like structures that are cascaded in the form of an encoder-decoder (auto-encoder) structure. The encoder is used for feature

extraction and the decoder is used for feature mapping to the original input resolution. A summary of the network architecture is shown in Figure 1. The convolution layers use a set of small parameterized filters, referred to as kernels, to perform convolution operations to produce different feature maps of their input. Here, each convolution and deconvolution layer uses a 3 × 3 kernel. Activation layers use a rectified linear unit (ReLU). Following convolution at each resolution, max pooling with a stride 2 is used on the encoding half of the network.

The network was implemented using Keras (v2.2.4)[35] with a TensorFlow backend (v1.13.1)[36] in Python 3.6.9. All training was carried out on an NVIDIA Titan Xp graphics processing unit (GPU) (3840 CUDA cores, 12 GB GDDR5X). The network uses a Dice score loss function, $D(A,B) = \frac{2|A \cap B|}{|A|+|B|} = \frac{2TP}{2TP+FP+FN}$, where TP is true-positive, FP is false-positive, FN is false-negative, a value of 1 implies complete overlap between the automated mask, and the manual mask whereas 0 implies no overlap. This function is ideal for renal segmentation because it does not weight true negatives, which represent the majority of voxels input to the network and therefore means that although the network is training, it does not become trapped in a local minimum outputting solely background voxels. Training was carried out over 150 epochs using stochastic gradient descent with an initial learning rate of 0.01 and learning rate decay of 5×10^{-7} and momentum of 0.8, these parameters help the optimizer converge quickly while also avoiding overshooting. As seen in Figure 2, after 150 epochs the validation Dice score plateaued whereas the training Dice score was still rising slightly, indicating that any further training would lead to over-fitting. Training took ~30 min.

Using the model to subsequently perform segmentation of renal masks from a given T_2-weighted volume of the test data took ~9 s on a standard office computer with no GPU, which is the type of machine end users would have access to.

2.4 | Statistical analysis

Baseline demographics are reported as mean ± SD. Inter-observer variability in manual segmentation and TKV was calculated by comparing the TKV of the manual masks each observer generated for a given volume, and also assessing the Bland–Altman and regression analysis. Intra-observer variability in manual segmentation was calculated by comparing the TKV of the 5 masks generated by an observer for a given subject. For each, the mean coefficient of variation (CoV; defined as SD/mean) and intraclass correlation (ICC) were used as measures of repeatability of TKV. Voxel-based (eg, Dice score) and surface based (eg, Hausdorff distance) metrics were also calculated between each observer.

The performance of the automated segmentation was assessed using multiple voxel and surface based similarity metrics. Performance was further assessed by determining the mean difference in TKV between the automatic and manual methods. Both actual and percentage (%) difference in TKV were evaluated. Bias (mean) obtained from the automatic and manual methods were assessed using a paired sample t test. The mean CoV and ICC were also used as measures of repeatability of the automated TKV.

3 | RESULTS

3.1 | Characteristics of the training cohort

Data were collected using a T_2-weighted HASTE sequence providing optimal contrast between the kidneys and surrounding tissue, examples shown in Figure 5, however, there is limited contrast between the left kidney and spleen because of their similar T_2-weighting. Cysts of variable size are clearly visible in the kidneys of the CKD patient. The training data comprised 25 HCs (9 female, 16 male) with a mean age of 26 ± 12 (19-77) years and 25 CKD patients (6 female,

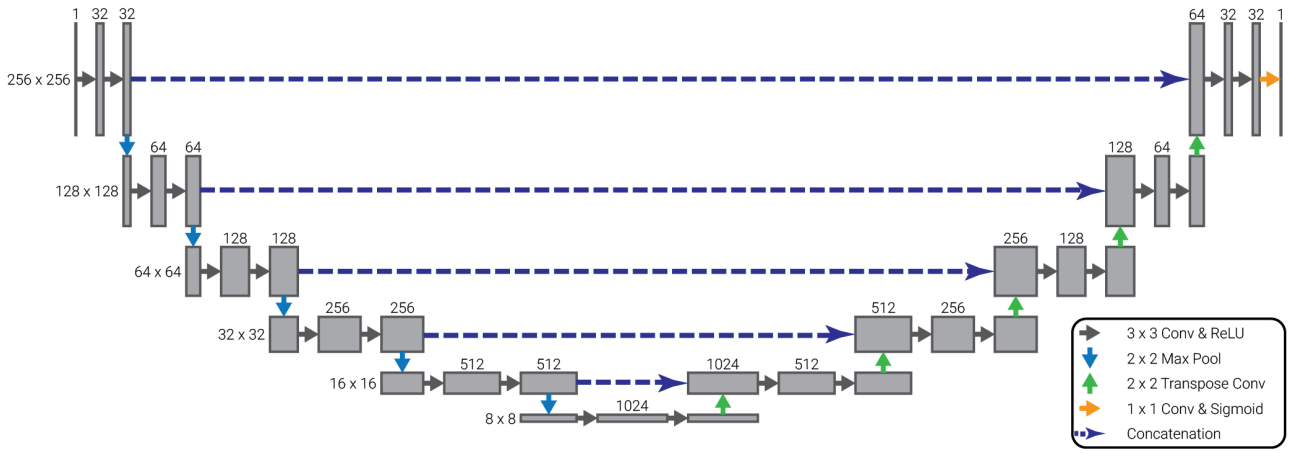

FIGURE 1 An overview of the network architecture

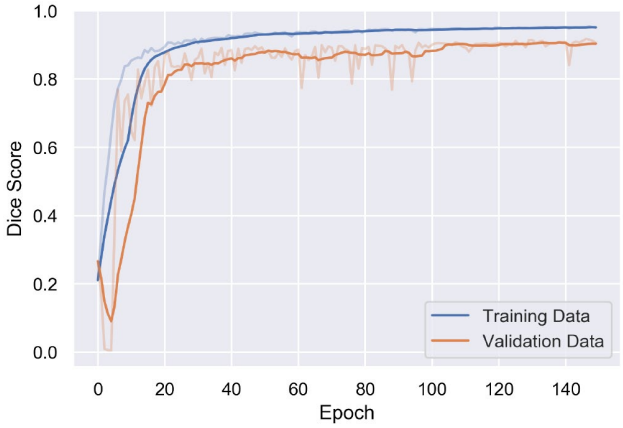

FIGURE 2 Dice score of the network for the training and validation data. Data are shown with a 10-epoch rolling average

TABLE 1 Repeatability of the manual segmentation for left, right, and TKV with coefficient of variation and intraclass correlation coefficient computed

Observer	Kidneys	CoV (%)	ICC
Intra A	Total	2.2 ± 0.7	0.939
	Left	3.2 ± 0.8	0.783
	Right	1.9 ± 0.5	0.957
Intra B	Total	1.9 ± 0.3	0.895
	Left	2.0 ± 0.5	0.807
	Right	2.4 ± 0.3	0.892
Intra C	Total	2.5 ± 0.9	0.908
	Left	2.8 ± 1.3	0.769
	Right	3.1 ± 1.9	0.940
Inter	Total	3.0 ± 1.0	0.897
	Left	4.0 ± 1.4	0.713
	Right	2.9 ± 1.0	0.910

Abbreviations: TKV, total kidney volume; CoV, coefficient of variation; ICC, intraclass correlation coefficient.
All values are quoted as mean ± SD.

19 male) with a mean age of 58 ± 15 (19-80) years and mean CKD stage 3.3 ± 1.1 (1-5). The manual TKV was 277 ± 60 mL, ranging between 145 and 422 mL. Including both HC subjects and CKD patients meant the kidneys had variable morphology (shape, size, and heterogeneous cysts) within the training data set. Supporting Information Table S1 provides the characteristics of data sets used for training and testing of the CNN, whereas Supporting Information Figure S1 shows the distribution of TKV within the training and testing data.

3.2 | Accuracy of manual segmentation

Four of the test subjects were each scanned 5 times, with the left and right kidneys in the 20 data sets each masked by Observers A, B, and C. The intra-observer and inter-observer variability for this manual segmentation was computed, as shown in Table 1 additionally, similarity metrics were used to assess the overlap between each observer's manual masks, Table 2. As a result of the large difference between in-plane and out-of-plane resolution (1.5 mm^3 vs. 5.5 mm^3) the Hausdorff distance is very susceptible to inaccuracies in the anterior–posterior direction; this metric is highly sensitive to noise and as such the 95th percentile is used to generate a more representative value. Bland–Altman plots and regression analysis of inter-observer variance in measured TKV are provided in Supporting Information Figure S2.

3.3 | Network testing

The trained network was used to predict segmentations of the 2D kidney slices and compute TKV for each of the unseen test volumes. The mean Dice score over the 50 test volumes was 0.93 ± 0.01 (0.94 ± 0.02 for HC and 0.92 ± 0.01 for CKD patients). The TKV predicted by the network was, on average, 1.2 ± 16.2 mL less than the manually segmented TKV

TABLE 2 Metrics comparing each combination of observers manual masks (A-B, A-C, and B-C)

Observer	Kidney	Dice score	Jaccard index	Average distance (mm)	Hausdorff distance (mm) (95th percentile)	Volume difference (mL)
A-B	Both	0.93 ± 0.03	0.87 ± 0.05	0.81 ± 0.58	5.59 ± 2.77	20.84 ± 9.33
	Left	0.92 ± 0.07	0.85 ± 0.10	0.94 ± 1.12	5.53 ± 3.65	13.36 ± 5.76
	Right	0.94 ± 0.01	0.88 ± 0.02	0.65 ± 0.14	4.75 ± 1.15	7.48 ± 5.63
A-C	Both	0.93 ± 0.01	0.87 ± 0.02	0.79 ± 0.18	5.83 ± 1.86	16.01 ± 8.56
	Left	0.93 ± 0.01	0.87 ± 0.02	0.84 ± 0.27	6.83 ± 3.12	6.93 ± 5.78
	Right	0.93 ± 0.01	0.87 ± 0.02	0.72 ± 0.17	4.82 ± 1.25	9.08 ± 5.41
B-C	Both	0.94 ± 0.04	0.89 ± 0.06	0.63 ± 0.62	3.59 ± 2.74	−4.83 ± 9.92
	Left	0.93 ± 0.08	0.88 ± 0.11	0.78 ± 1.22	4.31 ± 3.58	−6.44 ± 6.17
	Right	0.95 ± 0.01	0.90 ± 0.02	0.48 ± 0.14	3.39 ± 1.15	1.61 ± 6.56

All values are quoted as mean ± SD.

and therefore not significantly different ($P = .615$) (Figure 3). This accuracy was comparable for the HC and CKD cohorts, with automated CNN TKV measurements of 4.7 ± 17.7 mL greater than manual and 7.0 ± 12.4 mL less than manual, respectively. A summary of the CNN accuracy when evaluated using similarity metrics and volume difference from manual measures is shown in Table 3. Note a slightly larger discrepancy for the left compared to the right kidney.

Figure 3 shows plots of the difference in volume between manual segmentation and automated segmentation of the test data set.

In Figure 4, the TKV predicted by the CNN is plot against the manual TKV, in 90% of subjects, the SD of TKV measurements between each volume for a subject was smaller when the TKV was measured using the CNN as appose to manually. The mean CoV and ICC were 2.7% ± 0.9% and 0.979, respectively, across the 5 repeats of the manually segmented test data (using masks from observers A, B, and C), compared to a value of 1.5% ± 0.5% and 0.993, respectively, for the automatic segmentations of the 5 repeats of test data. The CNN produced a significantly lower CoV than the manual segmentations ($P = .008$).

Representative examples of the output from the network for both HC and CKD data are shown in Figure 5. The automated CNN accurately segments the kidneys, and for CKD patients, often omits cysts from the masks.

Because this is a 2D CNN, it is important to assess the accuracy across the anterior–posterior 2D slices of the kidney. This was achieved by comparing the Dice score of the CNN to the inter-reader Dice scores, Figure 6. A decrease in accuracy in the outer slices can be seen in both the CNN and manual masks.

This decrease in accuracy manifests itself on the outer slices of the volume, where the proportion of kidney per slice is smaller and as such the 2D network, with a lack of spatial context in the anterior–posterior direction, finds these outer slices more challenging. This decrease in accuracy can partly be explained by the fact that larger structures (in terms of number of voxels) will in general produce higher scores for comparable errors because the vast majority of errors are on the perimeter of the kidney in each slice, slices with fewer voxels of kidney have a smaller area to perimeter ratio.

4 | DISCUSSION

In this study, a 2D CNN has been trained to generate automatic segmentations of HC and CKD patients. Segmentations of the left and right kidneys are computed from which total kidney volume is estimated. The CNN was trained on both HC and CKD kidneys with a range of TKV (144.76-422.49 mL), which included the presence of cysts. The automated segmentation by the CNN yielded a mean Dice score of 0.93 ± 0.01 and took an average time of 9 s to measure TKV compared to 15-30 min[12] for manual segmentation. The automated CNN can be run as a self-contained package with the data and program freely available (https://github.com/alexdaniel654/Renal_Segmentor).[37] Note the software released at present can only be used to process coronal HASTE images and will not be accurate with other geometries and/or contrasts. To accurately segment other geometries and/or contrasts the network would need to be trained using a different data set, this cannot be done using the self-contained package and would necessitate the use of a GPU.

4.1 | Evaluation of methodology

The network performed with high precision on the test data with a 1.2 ± 16.2 mL, statistically insignificant, discrepancy between manual and automated TKV measurements. Table 3 shows the agreement between the CNN and manual masks

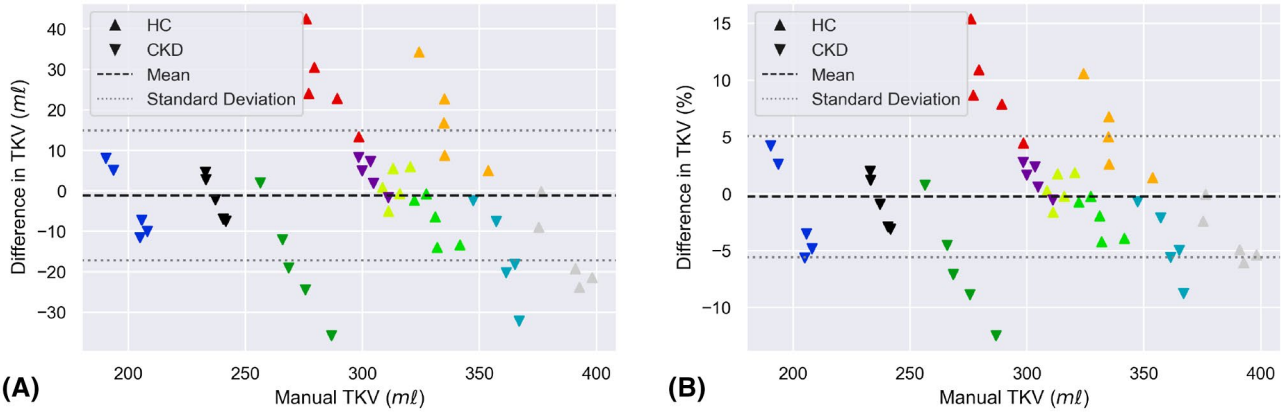

FIGURE 3 The difference between the TKV predicted by the CNN and the manually segmented true TKV. Mean and SD TKV difference are shown as dashed and dotted lines, respectively. Each subject is shown in a different color. (A) shows the absolute volume difference. (B) shows the percentage volume difference

TABLE 3 The accuracy of the CNN compared to manual segmentations using a variety of metrics stratifying the testing data by cohort and left versus right kidney

Cohort	Kidney	Dice score	Jaccard index	Sensitivity	Specificity	Precision	Accuracy	Mean Surface distance (mm)	Hausdorff distance (mm) (95th percentile)	Volume difference (*P*)
All	Total	0.93 ± 0.01	0.87 ± 0.03	0.93 ± 0.03	0.997 ± 0.001	0.93 ± 0.02	0.995 ± 0.001	0.65 ± 0.21	4.33 ± 1.64	−1.16 ± 16.23 (0.615)
	Left	0.92 ± 0.02	0.86 ± 0.04	0.91 ± 0.05	0.997 ± 0.001	0.94 ± 0.03	0.994 ± 0.002	0.76 ± 0.31	4.42 ± 1.52	−3.95 ± 12.38 (0.029)
	Right	0.94 ± 0.02	0.89 ± 0.03	0.95 ± 0.03	0.997 ± 0.001	0.93 ± 0.03	0.996 ± 0.001	0.54 ± 0.21	3.66 ± 1.76	2.79 ± 6.84 (0.006)
HC	Total	0.94 ± 0.02	0.88 ± 0.03	0.95 ± 0.05	0.997 ± 0.001	0.93 ± 0.03	0.995 ± 0.001	0.68 ± 0.27	4.50 ± 1.97	4.66 ± 17.72 (0.201)
	Left	0.93 ± 0.02	0.87 ± 0.04	0.94 ± 0.05	0.997 ± 0.001	0.93 ± 0.03	0.994 ± 0.002	0.79 ± 0.37	4.47 ± 1.81	1.91 ± 12.93 (0.467)
	Right	0.95 ± 0.02	0.90 ± 0.03	0.96 ± 0.03	0.997 ± 0.001	0.94 ± 0.02	0.996 ± 0.001	0.56 ± 0.26	3.81 ± 2.11	2.75 ± 7.70 (0.087)
CKD	Total	0.92 ± 0.01	0.86 ± 0.02	0.91 ± 0.02	0.998 ± 0.001	0.94 ± 0.02	0.995 ± 0.001	0.63 ± 0.14	4.16 ± 1.24	−6.98 ±12.38 (0.009)
	Left	0.92 ± 0.02	0.85 ± 0.03	0.89 ± 0.04	0.998 ± 0.001	0.95 ± 0.02	0.994 ± 0.002	0.73 ± 0.24	4.37 ± 1.21	−9.81 ± 8.62 (0.00001)
	Right	0.93 ± 0.01	0.88 ± 0.02	0.94 ± 0.02	0.997 ± 0.001	0.92 ± 0.03	0.996 ± 0.001	0.51 ± 0.13	3.51 ± 1.34	2.83 ± 6.02 (0.027)

All values are quoted as mean ± SD.

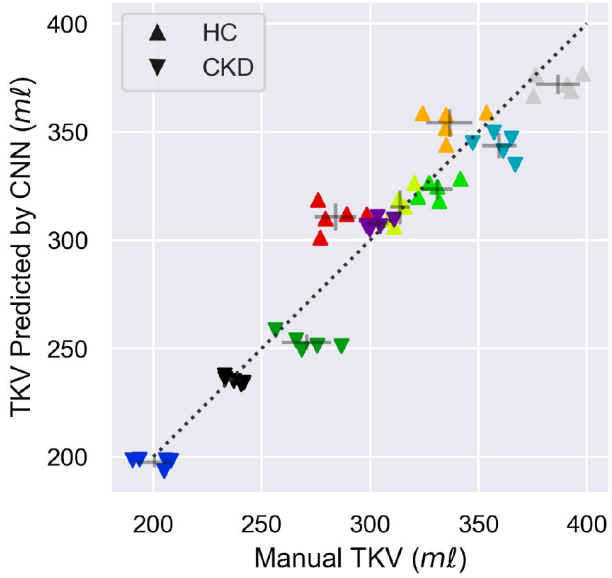

FIGURE 4 The TKV predicted by the CNN plot against the manually segmented true TKV with each subject plot in a different color. The SD measured using both methods are shown as error bars originating from the mean of each subject. The dotted line represents perfect correlation between the CNN and manual segmentation

is higher for the right than left kidney, this is in part because of the proximity and lack of contrast between the left kidney and the spleen making distinguishing this boundary difficult for the CNN. This difficulty also leads to inconsistencies in manual masks, borne out by the increased CoV and decreased ICC and similarity metrics of the left kidney when compared to the right kidney in Tables 1 and 2 assessing the variability in manual masks between observers. From Table 3, it can also be seen that the agreement between the CNN and manual masks is greater for the HC cohort than the CKD cohort, this is expected because of the increased variation in kidney morphology and the presence of cysts in the CKD cohort. Figure 3 shows that the difference between the manual TKV and CNN predicted TKV is not dependent on the true TKV, therefore, the training data are balanced and well augmented because the network is able to accurately perform over the full range of kidney size in the test data.

Here, 5 volumes of test data were collected for each subject by repositioning the subject in the scanner within an hour scan session, and therefore, any variance in measured TKV is purely because of inaccuracies in the kidney ROI definition. On assessing the correlation between manual and CNN measured TKV in Figure 4, it can be seen that, in 90% of subjects the intra-observer variance in manual TKV between the segmentation of the 5 volumes collected in each subject is larger than using the CNN to estimate TKV, as reflected by the lower CoV and increased ICC of the TKV measured using the CNN (CoV 1.5% $\pm$ 0.5%, ICC 0.993) compared to the manual measures (CoV 2.7% $\pm$ 0.9%, ICC 0.979). Because the network is trained on the kidney segmentations from 3 observers (A, B, and C), it has been optimized by inheriting the most accurate tendencies of each observer, for example, 1 observer may have been very accurate when excluding cysts but not as accurate at defining the kidney-spleen boundary. The network will have learnt to exclude cysts from this observer but to delineate between kidney and spleen from another observer. Therefore, the network can become more precise than each individual observer's manual segmentations. This increased precision can be seen in Figure 3 when compared to Figure 4 where the variance in difference in TKV is driven by the larger variance in manual TKV. The smallest TKV per subject is consistently overestimated when compared to its manual mask and vice versa the largest manual TKV per subject is often an underestimation compared to the manual TKV.

Figure 5 illustrates the masks produced by the manual segmentation and the CNN for both a HC and CKD patient. For the HC, the CNN includes more voxels around the edge of its mask than manual segmentation, and the network is more anatomically accurate, for example, where the interface between the kidney and spleen is very narrow, the CNN predicts the kidney is adjacent to the spleen whereas the observer's manual segmentation leaves a gap. The CKD data shown in Figure 5 includes a cyst in each of the kidneys. The network was trained on a combination of healthy and CKD data, with 19 of the 25 CKD training data sets containing at least 1 cyst. The CNN can be seen to segment out the cysts, despite their highly variable morphology and prevalence in the overall training data.

The amount of augmentation applied to the training data was empirically derived (random shifts up to 25% of the image in both the horizontal and vertical direction, zooms between 0.75 and 1.25 times magnification, rotations within a 20° range, and sheers within a 5° range) and led to the potential for large transforms being applied to the data and masks if the extremes of each transform were randomly selected. This large degree of augmentation was advantageous because it mirrors the large variation in acquisition planning in abdominal imaging.

A 2D CNN was used to process each 2D slice of a full volume, rather than a 3D volume. This was advantageous for the relatively small training data set the network was optimized on, because it avoids over-fitting and allows the network to easily be used on volumes of variable slice number. However, this can come at the expense of accuracy because 2D CNNs do not leverage the information from adjacent slices in the segmentation as is done in 3D CNNs, but 3D CNNs come with a computational cost as a result of the increased number of parameters used. 3D networks have successfully been implemented on neural data using patching methods where the image volume is divided up into smaller cubes[26] to reduce memory requirements and allow for differing input shapes. Although this works well in the brain, there are a number of reasons why this method may not be as successful for body

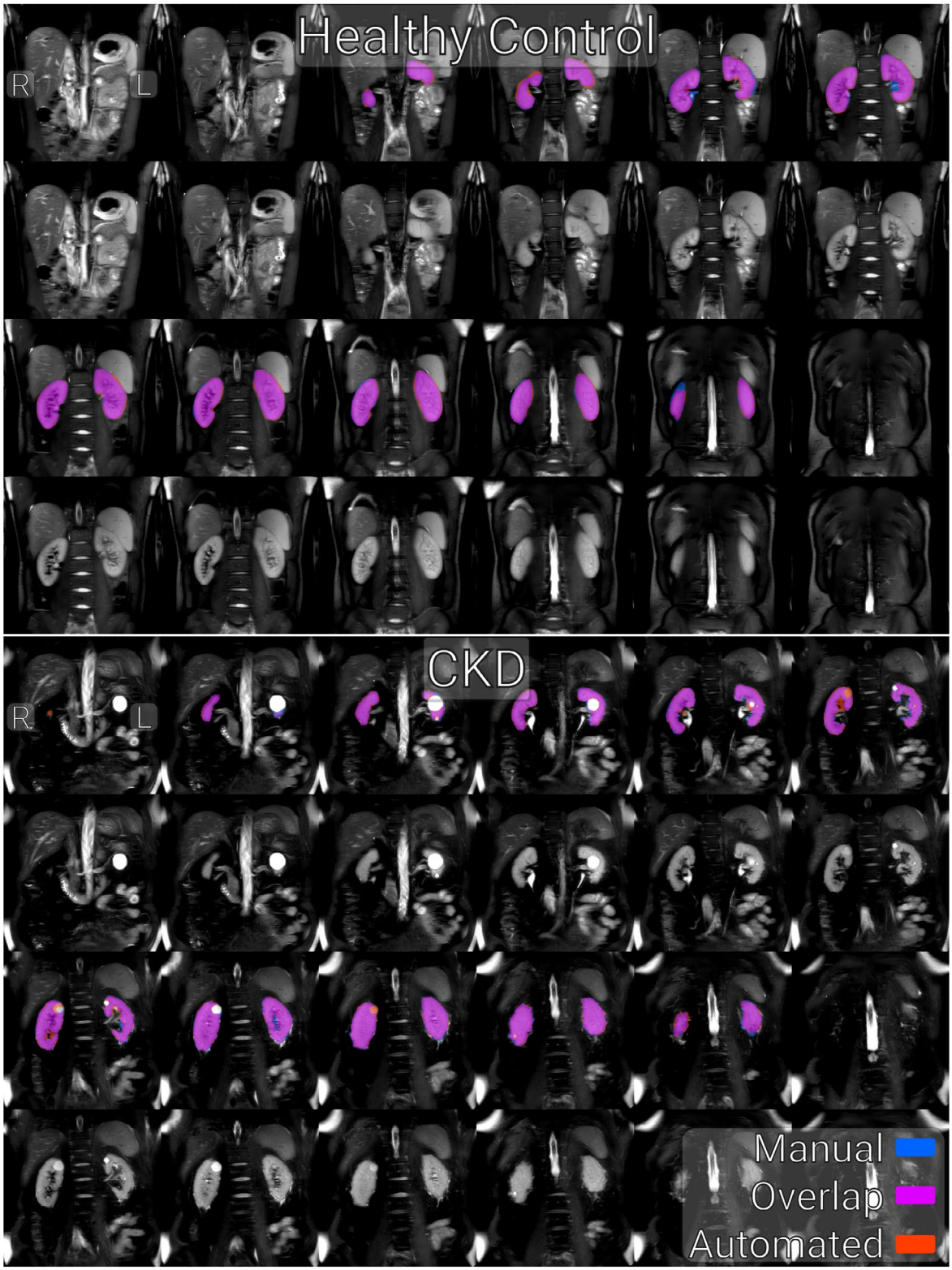

FIGURE 5 Representative raw test data and corresponding masks of a HC and CKD subject. Manually generated masks are shown in blue, automatically generated masks are shown in red and the overlap of the 2 is shown in magenta

applications. The out-of-plane resolution is significantly less than the in-plane resolution; this results in far fewer slices in 1 direction than the other 2. To avoid over-fitting for a certain number of slices, for example, training on an 11 slice image with a 11^3 patch, and subsequently the network not performing well when the patch is applied to a 16 slice image, the patch would need to be much smaller than the number of slices, therefore diminishing the benefits of the 3D

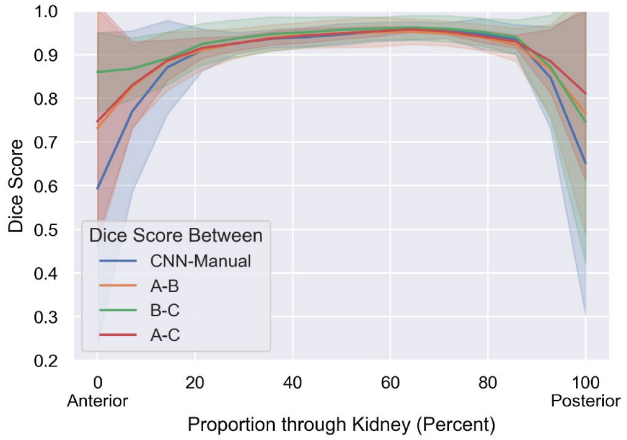

FIGURE 6 Mean Dice score for 2D slices from anterior to posterior. The shaded area represents 1 SD from the mean Dice score

methodology. Additionally, the extra memory requirements for a 3D network limit the ease of use of the software for inference on many standard office computers.

4.2 | Future directions

Future work will collect more training data to compare the 2D CNN with a 3D CNN to ascertain if the potential increase in accuracy is worth the increased hardware requirements and reduced generalizability. Here, the CNN has been developed for use on a T_2-weighted sequence and has not been validated on T_1-weighted images. This image contrast was chosen as a result of recent publications comparing T_1- and T_2-weighted images for TKV assessment reporting that T_2-weighted images provide better quality to enable TKV measurements, leading to improved reproducibility with lower intra- and inter-reader variability.[38] T_1-weighted data could be registered to the T_2-weighted data and used as an extra channel to inform the segmentation.

This network was validated on healthy subjects and CKD patients, but has not been trained and validated on subjects with ADPKD. These subjects have many more cysts in their kidneys, although the CNN was able to segment cysts encountered in the CKD cohort, it would be beneficial for future work on ADPKD to retrain the network with HC, CKD, and ADPKD data, where TKV is a recognized biomarker of disease progression.

Another common segmentation task in renal imaging is generating an ROI for the renal cortex and medulla. There are some automated methods of achieving this once a total kidney mask has been produced,[1,39] however, there has been no work on the application of deep learning to this task. In addition to the acquisition of the T_2-weighted data set used here, a T_1-weighted data set designed to optimize the contrast between cortex and medulla was also collected on each subject.[34] Using these data, it may be possible to develop this method further such that an automated mask for each tissue type is produced.

5 | CONCLUSIONS

A CNN has been successfully applied to accurately segment the kidneys from T_2-weighted renal MRI data and measure TKV in both HCs and CKD patients with higher than human precision. In the future, this will be used in clinical trials to study large numbers of CKD patients for serial measurements of TKV to monitor natural history or response to treatment.

ACKNOWLEDGMENTS

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research. This work was co-funded by the Animal Free Research (United Kingdome (UK) and the Medical Research Council (MR/R02264X/1). Animal Free Research UK is a UK medical research charity that funds and promotes non-animal techniques to replace animal experiments. A.J.D. is supported by a studentship from the Oxford Nottingham Biomedical Imaging Centre for Doctoral Training funded by the Engineering and Physical Sciences Research Council (EPSRC) and Medical Research Council (MRC) (EP/L016052/1).

CONFLICT OF INTEREST

Alexander Daniel is enrolled in a doctoral training program that MRM editor in chief, Prof. Jezzard, is director. Prof. Jezzard is also principle investigator on grant EP/L016052/1.

ORCID

Alexander J. Daniel https://orcid.org/0000-0003-2353-3283

REFERENCES

1. Cox EF, Buchanan CE, Bradley CR, et al. Multiparametric renal magnetic resonance imaging: validation, interventions, and alterations in chronic kidney disease. *Front Physiol*. 2017;8:696.
2. Cohen EI, Kelly SA, Edye M, Mitty HA, Bromberg JS. MRI estimation of total renal volume demonstrates significant association with healthy donor weight. *Eur J Radiol*. 2009;71:283-287.
3. van den Dool SW, Wasser MN, de Fijter JW, Hoekstra J, van der Geest RJ. Functional renal volume: quantitative analysis at gadolinium-enhanced MR angiography—feasibility study in healthy potential kidney donors. *Radiology*. 2005;236:189-195.
4. Chapman AB, Bost JE, Torres VE, et al. Kidney volume and functional outcomes in autosomal dominant polycystic kidney disease. *Clin J Am Soc Nephrol*. 2012;7:479-486.
5. Tangri N, Hougen I, Alam A, Perrone R, McFarlane P, Pei Y. Total kidney volume as a biomarker of disease progression in autosomal dominant polycystic kidney disease. *Can J Kidney Health Dis*. 2017;4:2054358117693355.

6. Grantham JJ, Torres VE, Chapman AB, et al. Volume progression in polycystic kidney disease. *N Engl J Med*. 2006;354:2122-2130.
7. Buchanan CE, Mahmoud H, Cox EF, et al. Quantitative assessment of renal structural and functional changes in chronic kidney disease using multi-parametric magnetic resonance imaging. *Nephrol Dial Transplant*. 2019;35:955-964.
8. Stevens LA, Coresh J, Greene T, Levey AS. Assessing kidney function—measured and estimated glomerular filtration rate. *N Engl J Med*. 2006;354:2473-2483.
9. Gong IH, Hwang J, Choi DK, et al. Relationship among total kidney volume, renal function and age. *J Urol*. 2012;187:344-349.
10. Di Leo G, Di Terlizzi F, Flor N, Morganti A, Sardanelli F. Measurement of renal volume using respiratory-gated MRI in subjects without known kidney disease: Intraobserver, interobserver, and interstudy reproducibility. *Eur J Radiol*. 2011;80:e212-e216.
11. Bae KT, Commean PK, Lee J. Volumetric measurement of renal cysts and parenchyma using MRI: phantoms and patients with polycystic kidney disease. *J Comput Assist Tomogr*. 2000;24:614-619.
12. Zöllner FG, Svarstad E, Munthe-Kaas AZ, Schad LR, Lundervold A, Rørvik J. Assessment of kidney volumes from MRI: acquisition and segmentation techniques. *Am J Roentgenol*. 2012;199:1060-1069.
13. Sharma K, Caroli A, Quach LV, et al. Kidney volume measurement methods for clinical studies on autosomal dominant polycystic kidney disease. *PLoS One*. 2017;12:e0178488.
14. Simms RJ, Doshi T, Metherall P, et al. A rapid high-performance semi-automated tool to measure total kidney volume from MRI in autosomal dominant polycystic kidney disease. *Eur Radiol*. 2019;29:4188-4197.
15. Cheong B, Muthupillai R, Rubin MF, Flamm SD. Normal values for renal length and volume as measured by magnetic resonance imaging. *Clin J Am Soc Nephrol*. 2007;2:38-45.
16. Spithoven EM, van Gastel MDA, Messchendorp AL, et al. Estimation of total kidney volume in autosomal dominant polycystic kidney disease. *Am J Kidney Dis*. 2015;66:792-801.
17. Seuss H, Janka R, Prümmer M, et al. Development and evaluation of a semi-automated segmentation tool and a modified ellipsoid formula for volumetric analysis of the kidney in non-contrast T2-weighted MR images. *J Digit Imaging*. 2017;30:244-254.
18. Magistroni R, Corsi C, Martí T, Torra R. A review of the imaging techniques for measuring kidney and cyst volume in establishing autosomal dominant polycystic kidney disease progression. *Am J Nephrol*. 2018;48:67-78.
19. Coulam CH, Bouley DM, Sommer FG. Measurement of renal volumes with contrast-enhanced MRI. *J Magn Reson Imaging*. 2002;15:174-179.
20. Karstoft K, Lødrup AB, Dissing TH, Sørensen TS, Nyengaard JR, Pedersen M. Different strategies for MRI measurements of renal cortical volume. *J Magn Reson Imaging*. 2007;26:1564-1571.
21. Gloger O, Tonnies KD, Liebscher V, Kugelmann B, Laqua R, Volzke H. Prior shape level set segmentation on multistep generated probability maps of MR datasets for fully automatic kidney parenchyma volumetry. *IEEE Trans Med Imaging*. 2012;31:312-325.
22. Kim Y, Ge Y, Tao C, et al. Automated segmentation of kidneys from MR images in patients with autosomal dominant polycystic kidney disease. *Clin J Am Soc Nephrol*. 2016;11:576-584.
23. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015. Lecture Notes in Computer Science*. Cham: Springer International Publishing; 2015. pp. 234-241.
24. Lu F, Wu F, Hu P, Peng Z, Kong D. Automatic 3D liver location and segmentation via convolutional neural network and graph cut. *Int J Comput Assist Radiol Surg*. 2017;12:171-182.
25. Sharma K, Rupprecht C, Caroli A, et al. Automatic segmentation of kidneys using deep learning for total kidney volume quantification in autosomal dominant polycystic kidney disease. *Sci Rep*. 2017;7:2049.
26. Wachinger C, Reuter M, Klein T. DeepNAT: deep convolutional neural network for segmenting neuroanatomy. *Neuroimage*. 2018;170:434-445.
27. Fu Y, Mazur TR, Wu X, et al. A novel MRI segmentation method using CNN-based correction network for MRI-guided adaptive radiotherapy. *Med Phys*. 2018;45:5129-5137.
28. Hassanzadeh T, Hamey LGC, Ho-Shon K. Convolutional neural networks for prostate magnetic resonance image segmentation. *IEEE Access*. 2019;7:36748-36760.
29. Li X, Chen H, Qi X, Dou Q, Fu C-W, Heng P-A. H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE Trans Med Imaging*. 2018;37:2663-2674.
30. Kline TL, Korfiatis P, Edwards ME, et al. Performance of an artificial multi-observer deep neural network for fully automated segmentation of polycystic kidneys. *J Digit Imaging*. 2017;30:442-448.
31. van Gastel MDA, Edwards ME, Torres VE, Erickson BJ, Gansevoort RT, Kline TL. Automatic measurement of kidney and liver volumes from MR images of patients affected by autosomal dominant polycystic kidney disease. *J Am Soc Nephrol*. 2019;30:1514-1522.
32. Shin TY, Kim H, Lee J-H, et al. Expert-level segmentation using deep learning for volumetry of polycystic kidney and liver. *Investig Clin Urol*. 2020;61:555-564.
33. Petzold K, Gansevoort RT, Ong ACM, et al. Building a network of ADPKD reference centres across Europe: the EuroCYST initiative. *Nephrol Dial Transplant*. 2014;29:iv26-iv32.
34. Will S, Martirosian P, Würslin C, Schick F. Automated segmentation and volumetric analysis of renal cortex, medulla, and pelvis based on non-contrast-enhanced T1- and T2-weighted MR images. *Magn Reson Mater Phys Biol Med*. 2014;27:445-454.
35. Chollet F. Keras. Keras; 2015. https://keras.io/getting_started/faq/#how-should-i-cite-keras. Accessed March 17, 2021.
36. Abadi M, Agarwal A, Barham P, et al. TensorFlow: large-scale machine learning on heterogeneous distributed systems. 2015;19. https://www.tensorflow.org/about/bib. Accessed March 17, 2021.
37. Daniel A. *alexdaniel654/Renal_Segmentor: v1.0.0*. Genève, Switzerland: Zenodo; 2020. https://zenodo.org/record/4068851. Accessed March 17, 2021.
38. van Gastel MDA, Messchendorp AL, Kappert P, et al. T1 vs. T2 weighted magnetic resonance imaging to assess total kidney volume in patients with autosomal dominant polycystic kidney disease. *Abdom Radiol*. 2018;43:1215-1222.
39. Morris D, Donnelly M, Gifford F, et al. Segmentation of the cortex and medulla in multiparametric magnetic resonance images of the kidney using K-means clustering. In: Proc. Intl. Soc. Mag. Reson. Med. 27. Vol. 27. Montreal; 2019. p. 1915.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

FIGURE S1 Distribution of TKV within the training and testing data

FIGURE S2 Bland–Altman plots and regression analysis of inter-observer variance in measured TKV

TABLE S1 Characteristics of data sets used for training and validation of the 2D U-Net model CNN

How to cite this article: Daniel AJ, Buchanan CE, Allcock T, et al. Automated renal segmentation in healthy and chronic kidney disease subjects using a convolutional neural network. *Magn Reson Med*. 2021;00:1–12. https://doi.org/10.1002/mrm.28768