scientific reports

Check for updates

OPEN

Kidney segmentation in neck-to-knee body MRI of 40,000 UK Biobank participants

Taro Langner[1]✉, Andreas Östling[1], Lukas Maldonis[2], Albin Karlsson[1], Daniel Olmo[1], Dag Lindgren[2], Andreas Wallin[2], Lowe Lundin[2], Robin Strand[1,3], Håkan Ahlström[1,2] & Joel Kullberg[1,2]

The UK Biobank is collecting extensive data on health-related characteristics of over half a million volunteers. The biological samples of blood and urine can provide valuable insight on kidney function, with important links to cardiovascular and metabolic health. Further information on kidney anatomy could be obtained by medical imaging. In contrast to the brain, heart, liver, and pancreas, no dedicated Magnetic Resonance Imaging (MRI) is planned for the kidneys. An image-based assessment is nonetheless feasible in the neck-to-knee body MRI intended for abdominal body composition analysis, which also covers the kidneys. In this work, a pipeline for automated segmentation of parenchymal kidney volume in UK Biobank neck-to-knee body MRI is proposed. The underlying neural network reaches a relative error of 3.8%, with Dice score 0.956 in validation on 64 subjects, close to the 2.6% and Dice score 0.962 for repeated segmentation by one human operator. The released MRI of about 40,000 subjects can be processed within one day, yielding volume measurements of left and right kidney. Algorithmic quality ratings enabled the exclusion of outliers and potential failure cases. The resulting measurements can be studied and shared for large-scale investigation of associations and longitudinal changes in parenchymal kidney volume.

The UK Biobank studies more than half a million volunteer participants, collecting health data on medical records, blood and urine samples, lifestyle, and genetics[1]. Together with the vast range of metadata and a long-term follow-up, medical images are acquired for a subgroup of 100,000 participants. Of these, 10,000 are also scheduled to attend a repeat scan at a second, later imaging visit. The protocols include Magnetic Resonance Imaging (MRI) of the brain, heart, pancreas and liver, but also neck-to-knee body MRI[2] which combines vast amounts of anatomical information in a single comprehensive 6-minute scan, which covers all tissue of the kidneys in overlapping imaging stations. The human kidney plays a vital role in the filtration of blood, secretion of hormones, and regulation of blood pressure. Its shape and function are impacted by genetic factors, but also underlie natural variation based on sex, body size, and age[3]. In addition to congenital anomalies such as renal fusion, or horseshoe kidneys[4], and autosomal dominant polycystic kidney disease (ADPKD)[5], morphological changes with associated medical complications can arise from factors such as chronic kidney disease, hypertension[6], and diabetes[7]. Kidney volume as a biomarker is therefore of clinical interest for diagnostics, monitoring of disease progression, and medical hypothesis testing. With the extensive image data available in the UK Biobank, non-invasive, image-based assessments of kidney volume could provide a substantial sample size of these measurements.

In clinical practice, kidney volume is often approximated with a rotational ellipsoid model based on kidney width, depth, and length as obtained by sonography[8]. However, validation with water displacement methods has shown that ellipsoid models can underestimate kidney volume by up to 29%[9], or 25% even with MRI[8]. As an alternative, measurements by voxel count, or disc-summation, can be obtained by delineation of three-dimensional voxels which correspond to kidney tissue in volumetric medical images. When obtained from MRI, these segmentation-based measurements have been found to show no significant deviation from those determined by water displacement[8]. When applied to the UK Biobank, manual segmentation is no longer feasible, however, as even a typical processing time of ten minutes per subject would amount to tens of thousands of man-hours for the UK Biobank cohort as a whole. A rich body of literature has been devoted to computer-aided

[1]Department of Surgical Sciences, Uppsala University, 751 85 Uppsala, Sweden. [2]Antaros Medical AB, BioVenture Hub, 431 53 Mölndal, Sweden. [3]Department of Information Technology, Uppsala University, 751 85 Uppsala, Sweden. ✉email: taro.langner@surgsci.uu.se
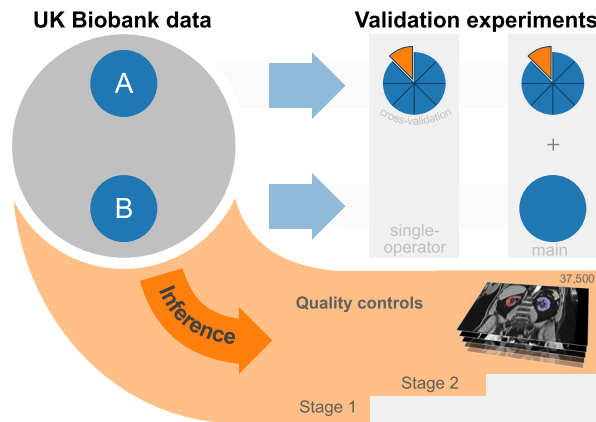
nature research

1

Figure 1. Among all UK Biobank subjects two subsets, A and B, were manually segmented. A neural network was evaluated in two cross-validation experiments on these and applied for inference to all remaining subjects. After excluding 5% of results in two quality control stages, about 37,500 measurements remain as final result.

segmentation techniques of the kidneys and other visceral organs in volumetric medical imaging data. For the kidney in particular, various approaches have been proposed predominantly for image data from Computed Tomography (CT), including techniques based on statistical shape models and region growing[10], graph cuts[11], and deformable boundaries[12]. Contemporary benchmark challenges are increasingly dominated by machine learning techniques such as deep learning with convolutional neural networks, as seen in the MICCAI *2019 Kidney and Kidney Tumor Segmentation (KiTS19)*[13] and with CT image data, in which similar approaches have also been proposed for measurements of total volume in subjects with ADPKD[14]. Fully convolutional networks for semantic image segmentation[15] range from architectures such as the U-Net with 2D data[16] to 2.5D[17] and 3D techniques[18], which are able to learn the task of segmenting specific image structures from reference data in training. In the UK Biobank, related approaches have already been applied for segmentation of cardiac MRI of up to 5000 subjects[19,20], but also to the liver[21] and pancreas[22]. Large-scale segmentation of this data has been conducted with other methods as well, such as sparse active shape models on up to 20,000 subjects[23].

The purpose of this work is to propose, validate, and apply a segmentation pipeline for automated quantification of parenchymal kidney tissue in UK Biobank neck-to-knee body MRI. A neural network based on a 2.5D U-Net variation is evaluated in cross-validation and applied for inference to all 40,000 subjects with available MRI data, resulting in measurements of healthy tissue volume of the left and right kidney, as well as their mutual distance. Potential failure cases and other outliers are identified with algorithmic quality ratings, with a large number of anomalies such as renal fusion and polycystic cases being highlighted for scrutiny. We are not aware of any existing kidney volume measurements within the UK Biobank, or any other study with MRI-based measurements of kidney volume at a comparable sample scale. The obtained values and code samples can be shared as return data and made available for further research.

Methods

A neural network was trained for semantic segmentation of two-dimensional, axial slices of the UK Biobank neck-to-knee body MRI. Manually created reference segmentations of parenchymal kidney volume in 122 subjects were used to train and validate the network, and to make design choices regarding data pre-processing and hyperparameter selection. The resulting network configuration was then embedded in a processing pipeline and applied in inference to the entire cohort, with algorithmic quality ratings flagging suspected failure cases for exclusion. A schematic overview over the experiments is shown in Fig. 1.

UK biobank data. At the time of writing, UK Biobank neck-to-knee body MRI of 40,264 participants has been released. Subjects were originally recruited by letter from the National Health Service and scanned at three different imaging centres in Great Britain with a Siemens Aera 1.5T device, using a dual-echo protocol that acquired overlapping images in six stations covering the body from neck to knee within about 6 minutes with TR = 6.69, TE = 2.39/4.77 ms, and flip angle 10deg[2]. The reconstructed volumetric station images encode voxel-wise intensity values with a separate water and fat signal (UK Biobank field 20201-2.0). The head, arms, and lower legs extend outside of the field of view and are often distorted near the image borders.

The kidneys are typically located in the second and third imaging station, each of which were acquired in a 17s breath-hold with typical dimensions of (224 × 174 × 44) voxels of (2.232 × 2.232 × 4.5) mm. In this work, those subjects with image artefacts in these two stations, such as water-fat swaps, background noise, metal objects, but also non-standard poses, misalignment in the scanner, and corrupted data were excluded after visual inspection of mean intensity projections[24], leaving 39,560 subjects. At scan time, these men and women (52% female) were 44 to 82 (mean 64) years old, with BMI 14 to 62 (mean 27) kg/m^2 and a 95% majority of self-reported White British ethnicity.
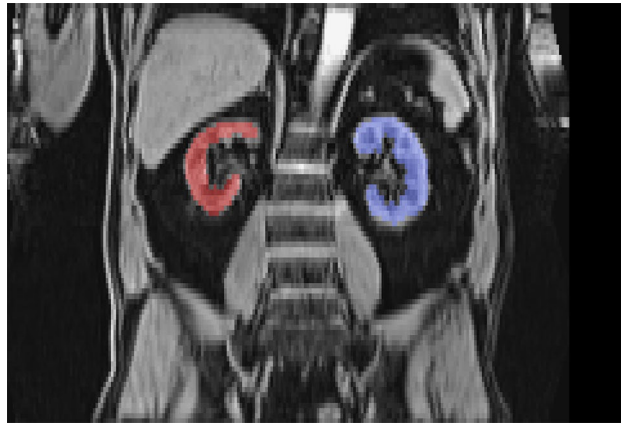
Figure 2. Segmented parenchymal tissue of right (red) and left (blue) kidney in MRI of a male subject.

Reference segmentations. Three operators created reference segmentations by marking all voxels corresponding to healthy, parenchymal kidney tissue in the water signal of the second and third imaging station. The segmented tissue corresponds to the combined cortex and medulla, both of which appear with high MRI water signal intensity. Based on their lower signal intensities cysts, the calyces, ureters, and major vessels were excluded. These reference segmentations were used to train and validate the neural network. For the final measurements, the left and right kidney were separated by subsequent post-processing with an algorithmic connected component analysis. An example for a segmented MRI slice is seen in Fig. 2.

Dataset *A* consists of 64 subjects selected by random sampling stratified by age, gender, and weight[25], whose water signal images were manually segmented in the software SmartPaint[26] by an image analyst with three years of working experience. The consistency of these segmentations was evaluated on a subset of 5 subjects, which were segmented repeatedly by the operator for a blinded assessment of intra-operator variability.

Dataset *B* contains another 64 subjects, with no overlap to dataset *A*. Of these, 33 cases were segmented by the second and 31 cases by the third operator, both novice image analysts working also with SmartPaint. Instead of using a fully manual procedure, this dataset was segmented by correcting the proposals generated by a preliminary inference network trained on dataset *A*. The 64 candidates were selected among the most challenging cases based on an algorithmic quality rating for segmentation smoothness which is described in more detail further below. As a result, subjects with morphological anomalies are over-represented in this dataset, with several pathological cases that are challenging to segment even for human operators. To determine inter-operator variability, these two operators also segmented the same subset of 5 subjects from dataset *A* for which the intra-operator variability was previously determined.

Neural network configuration. A fully convolutional neural network was trained for semantic image segmentation of axial slices. The underlying architecture is a 2.5D variation of the U-Net[16] with a VGG11 encoder pre-trained on ImageNet[27], extended with ResNet-style short skip connections. The 2.5D input formatting combines three adjacent slices to form one sample, providing additional volumetric information to the network. Similar techniques with five slices[17] ranked among the most successful contributions for segmentation of liver tissue in the 2017 MICCAI Liver Tumor Segmentation (LiTS) Benchmark Challenge[28].

The network assigns pixel-wise labels to two-dimensional, axial slices. Each 2.5D input sample is formed by a stack of three adjacent slices from a water signal station. In addition to the target slice itself, one additional slice is extracted from above and below, using a periodic border condition. At this stage, no image fusion is performed and during training some of the outermost station slices were excluded due to excessive artefacts, such as signal loss and folding. The intensity values of each axial slice were then normalized after clipping the brightest one percent of values for stability. For an evenly divisible size, the slices were then symmetrically zero-padded to form a stack of $224 \times 192 \times 3$ pixels. Each of these stacks forms one input sample for the network, which predicts a two-dimensional segmentation for the central slice in the format of 224×192 pixels. The network architecture was trained for 80,000 iterations with a pixel-wise cross-entropy loss, batch size one, and online augmentation with randomized, elastic deformations[25]. Using the Adam optimizer, a learning rate of 0.0001 is maintained until iteration 60,000 and then lowered by a factor of 10 for improved stability. After reverting the slice padding, the segmented slices with pixel-wise labels can be stacked to obtain voxel-wise labels for an entire input station. A GitHub implementation is linked in the Supplementary Material.

Training data. Three experiments were conducted with this neural network configuration. For training and validation of the network, both dataset *A* and *B* were available, with image data of 64 subjects each. The samples of dataset *B*, however, were selected among the most challenging, including some pathological cases and are largely based on refined segmentations originally proposed by the network itself. Using these samples for validation would yield results that are not representative for the UK Biobank cohort as a whole and dataset *B* was therefore never used for validation. Six of its 64 cases were furthermore excluded due to excessive morphological

anomalies, tumours, cysts and congenital renal fusion where both kidneys are interconnected and form a single structure. Thus, 58 cases of dataset *B* remained for further use.

A single-operator validation was performed by conducting a classical 8-fold cross-validation on the 64 cases of dataset *A*. This dataset was consequently split on subject level into 8 subsets of even size, for each of which in turn segmentations were predicted by a network instance trained on data of all remaining 7 subsets. Each network instance was thereby trained on data of 56 subjects, corresponding to about 4,250 labelled slices. This single-operator cross-validation aims to quantify how well the operator of dataset *A* can be emulated on a representative sample of the UK Biobank.

Secondly, the main validation experiment quantifies the benefit of access to both datasets ($A \cup B$) by repeating the cross-validation with the exact same splits, but with samples of dataset *B* added for training only. In this way, the network instance for each split used the same validation subjects as before, but was trained on both the remaining 56 cases of dataset *A* and all well-formatted 58 cases of dataset *B* combined, for a total of 114 subjects, or 8,650 labelled slices for training. The network thereby learns a compromise in segmentation style between all operators, with validation results that are representative for the actual inference pipeline.

Finally, the network was applied for inference itself to all those subjects with no reference data. It was trained on a combined dataset of the 64 cases of datasets *A* and the well-formatted 58 cases of dataset *B*, for a total of 122 cases with about 9,250 labelled slices in total.

Measurements. The second and third stations of a given subject were labelled by the neural network and subsequently fused into a single, combined volume for both the water signal and voxel-wise labels each, by resampling to a common voxel grid and interpolation along the overlapping areas. A kidney volume measurement was obtained from these fused segmentation images by summing up the number of voxels labelled as kidney tissue, scaled with the physical voxel dimensions. Post-processing extracted the two largest connected components individually, which are assumed to be the left and right kidney, identified by the relative position of their centres of mass. The latter also enables a measurement of the relative position and euclidean distance between both kidneys.

Validation metrics. When validating the network output against known reference segmentations, the segmentation quality was evaluated with the Sørensen–Dice coefficient, or Dice score, and Jaccard index. To avoid averaging with empty imaging stations, these metrics were only calculated after fusing the image stations for a given subject. All measurements were likewise only derived after image fusion and evaluated with several complementary metrics. Averaging the absolute differences between predicted value and reference for all subjects yields a mean absolute error (MAE). In addition to this value, a relative error measurement is reported for a better sense of scale. Dividing the absolute differences on a per-case basis by the true measurement value, estimated here as the mean between prediction and reference, before averaging, results in a symmetric mean absolute percentage error (SMAPE). For a single example case with true volume of 250 cm^3, an absolute difference of 25 cm^3 would thereby amount to a SMAPE of 10%. Instead of estimating the true value as the mean of prediction and reference, an alternative would be to simply use the reference value directly. However, the known high variation between references created by different operators suggests that the chosen symmetrical definition may be more robust. In addition to these metrics, the quality of fit between predicted values and reference can be quantified with the coefficient of determination (R^2), whereas error bounds are estimated by the 95% limits of agreement (LoA).

Algorithmic quality controls. When applying the network in inference to those cases with no existing reference measurements, the aforementioned validation metrics can not be calculated. Exhaustive quality control by manual inspection is likewise hardly feasible at this scale. The evaluation during inference is therefore based on algorithmic quality ratings as simple indicators for outliers and potential failure cases. While ratings of high quality provide no guarantee for correct results, low ratings can help to identify those cases that are likely to contain anomalies or potential segmentation failures. The distribution of ratings were examined in two separate control stages, after each of which the most severe outliers were flagged for exclusion (see supplementary material for details).

The first stage of quality controls evaluates the image quality with an *image fusion rating*, *segmentation fusion rating*, and *location rating*. Even a hypothetically perfect segmentation can result in faulty measurements if parts of the kidney are not contained in the image at all, or occur multiple times due to motion. The agreement between both stations in their overlapping area is therefore examined, both for the MRI water signal and their segmented labels. Large differences indicate bad anatomical alignment between the imaging stations, leading to low quality ratings. Additionally, the relative offset along the longitudinal axis between the centre line of the fused subject volume and the centre of mass of all segmented voxels is penalized. Low values for this rating indicate that the kidneys are located at the top or bottom edge, possibly extending beyond the field of view.

The second quality control stage rates the segmentation quality with a *segmentation smoothness rating* and *scrap volume rating*, examining the smoothness of the segmented volume along the longitudinal axis and the share of voxels which are not part of either of the two largest connected components. The slice-wise segmentation by the 2.5D neural network may encounter failure cases where entirely disconnected islands of tissue are spuriously segmented or excluded due to their position and local appearance. Low ratings indicate atypical shapes for the segmented kidney tissue which may require further scrutiny.

	N	Dice	MAE	SMAPE	R^2	LoA
Network validation						
Main result	64	0.956	$10\,cm^3$	3.8%	0.950	$(-26\ to\ 14\,cm^3)$
Single-operator	64	0.956	$9\,cm^3$	3.4%	0.950	$(-22\ to\ 23\,cm^3)$
Human variability						
Intra-operator	5	0.962	$6\,cm^3$	2.6%	0.994	$(-4\ to\ 13\,cm^3)$
Inter-operator	5	0.920	$27\,cm^3$	10.0%	0.839	$(-59\ to\ 5\,cm^3)$

Table 1. Validation results and human variability for combined kidney volume. Validation on N subjects for the neural network and repeat segmentation by human operators. Whereas the single-operator validation is a classical cross-validation on dataset *A*, the main result was obtained by training on samples of both datasets *A* and *B*. The resulting measurements of combined kidney volume were evaluated with the mean absolute error (MAE), symmetric mean absolute percentage error (SMAPE), coefficient of determination (R^2) and 95% limits of agreement (LoA). The latter are calculated as (reference − predicted), yielding a negative shift for oversegmentation.
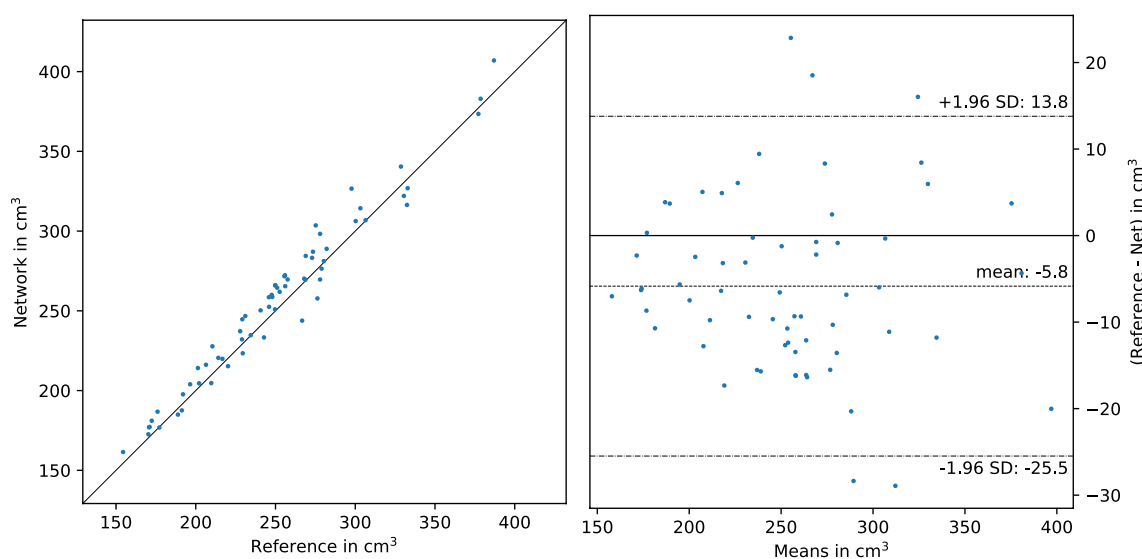

Figure 3. Main validation result for 64 subjects of dataset *B*, with images from dataset *B* added for training. The diagonal line in the scatter plot on the left represents a hypothetical perfect result, whereas the dashed lines in the Bland–Altman plot on the right give the 95% Limits of Agreement. When compared to the reference, the network appears to emulate a tendency towards oversegmentation which is also seen in Table 1 for the operators who provided reference segmentations for *B*.

Results

Network validation.   In both validation experiments, the neural network reached a Dice score of 0.956 on the 64 subjects dataset *A*. The main result, in which the 58 selected subjects of dataset *B* were added for training, measured combined kidney volume with an average error of $10\,cm^3$, or 3.8%. These values are slightly worse than those achieved by the single-operator cross-validation, and the LoA indicate systematic oversegmentation by the network relative to the operator of dataset *A*, similar to the operators who supplied the training data for dataset *B*. Table 1 and Fig. 3 show more validation metrics for these results, together with the variability between human operators for context. Additional detail is given in the Supplementary Tables 1, 2, and 3, with the corresponding Jaccard indices and individual measurements of left and right kidney for both network and operators.

Inference.   The inference network generated measurements for all those 39,432 subjects lacking reference segmentations. Only a small number of these cases exhibited disjunct or fragmented segmentations, with the *scrap volume* rating indicating that, on average, only about 1 in 900 voxels were not part of the two largest connected components segmented for the given subject (about half of a preliminary run trained on dataset *A* only). Low quality ratings are concentrated in a small subgroup of subjects, as shown in Fig. 4, which were isolated in the following quality control stages.

After examining the algorithmic ratings for image quality, it was decided to flag the top one percent of worst *location cost* and *image fusion cost* as well as the top two percent of worst *segmentation fusion cost* for exclusion in a first control stage. Due to their mutual overlap, the subjects marked in this way amount to about 3.6% of all
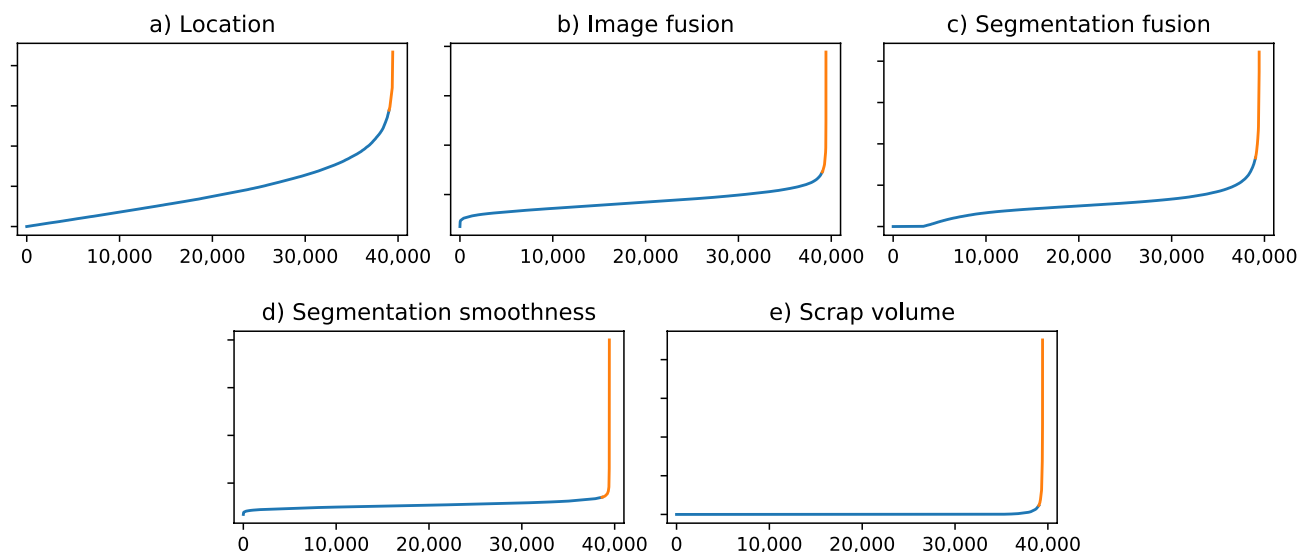
Figure 4. Distribution of algorithmic quality ratings by subject, sorted separately for each rating. High values of each cost term indicate low quality. In stage one (**a–c**) and stage two (**d,e**) of quality controls, the highlighted top one or two percent of subjects were accordingly flagged for exclusion as potential failure cases.

Property	Mean ± SD	[min, max]	Median	(10%, 90%)
Combined				
Male	282 ± 51	[91, 586]	277	(221, 348)
Female	224 ± 40	[76, 499]	220	(177, 276)
Left				
Male	143 ± 29	[0, 408]	141	(110, 178)
Female	114 ± 22	[0, 304]	112	(88, 141)
Right				
Male	139 ± 28	[0, 408]	137	(108, 173)
Female	110 ± 22	[0, 268]	108	(86, 138)

Table 2. Inferred UK Biobank parenchymal kidney volumes in cm^3. Male (N=17,846) and female (N=19,622) parenchymal kidney volumes in cm^3. SD denotes the standard deviation.

cases, many of which show signs of severe motion artefacts or misalignments of the field of view. Some of these cases were trivially re-included, having been flagged by the *location cost* for proximity to the image borders while being too small to extend beyond them. Next, the algorithmic ratings for segmentation quality were examined for the remaining subjects as the top one percent of worst *segmentation smoothness cost* and worst *scrap volume cost*, another 1.8% of subjects were flagged in this step. More cases with motion were caught at this stage, as well as genuine failure cases in which the network mistakenly segmented parts of the spleen or liver, but also cases of fragmentation caused by severe cystic formations. In total, 5% of subjects were ultimately excluded, with representative cases shown in Supplementary Fig. 1. Many of these cases contain pathological anomalies. In turn, perhaps up to a third of them could potentially be re-included without any corrections, but were not considered any further in this work.

After these exclusions, 37,468 subjects remain, with 17,846 men and 19,622 women. Disjunct scrap volume occurs in about 20% of these subjects, but amounts to only 1 in 2,200 segmented voxels on average and never exceeds a share of 2.5% for any individual. Outliers with unusually high or low volumes were inspected as potential failure cases, but were found to be plausible measurements of subjects with missing kidneys, unilateral hypertrophy/atrophy, or were associated with outliers of body size. An in-depth medical analysis of the resulting measurements is beyond the scope of this work and remains to be explored in the future. However, as a brief summary, Fig. 5 shows the distribution of measured combined kidney volume, with additional statistics given in Table 2, and further detail on the offset between kidneys in Supplementary Table 4.

Runtimes and memory requirements. Training the 2.5D U-Net on an Nvidia RTX 2080 Ti 11GB GPU for 80,000 iterations required about 30 minutes per split, or about 3.5 hours for the entire 8-fold cross-validation. The MRI data for water and fat signal was stored in DICOM format on an encrypted USB-SSD, amounting to 750 GB for 40,000 subjects. The inference pipeline loaded and processed individual scan volumes from this drive.
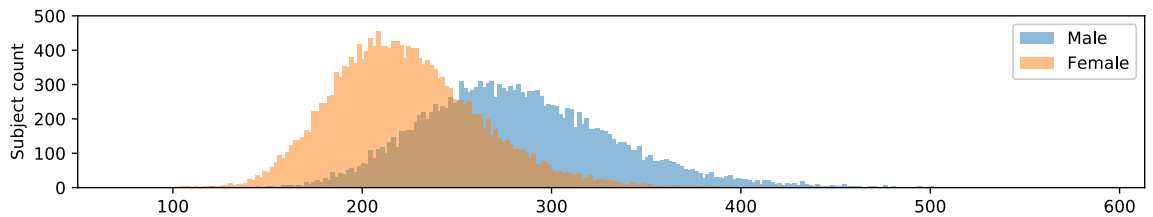
Figure 5. Inferred UK Biobank parenchymal kidney volume (left + right) in cm^3 for 17,846 male and 19,622 female subjects.

Despite an efficient GPU implementation, the image fusion formed the bulk of processing time during inference, amounting to about 11 hours for all 40,000 subjects.

Discussion

As main validation result, the proposed measurements for total kidney volume agree with the reference for a mean error of 10 cm^3, or 3.8%, and Dice score 0.956. The assessment of human performance showed slightly superior results for blinded repeat segmentation by a single operator, with mean error 6 cm^3, or 2.6%, and Dice score 0.962, whereas the variability between different human operators was more than twice as large as the network error. When applied to the entire cohort, around 37,500 subjects yielded volume measurements with no signs of potential measurement failure, whereas another 5% require further controls.

Only healthy parenchymal tissue was segmented, including cortex and medulla while excluding the renal pelvis, calyces, ureters, major vessels, and cysts. The measurements obtained by the inference therefore differ from those typically used for the tracking of conditions such as ADPKD, which may nonetheless benefit from the identification of pathological outliers in this work. These cases are highly concentrated in the 5% of subjects flagged by the algorithmic quality controls, which also helped to identify about 40 suspected cases of renal fusion. With median total kidney volumes of 277 cm^3 for men and 220 cm^3 for women, the volumes acquired by the proposed pipeline are smaller than those typically reported in the medical literature, especially when more than just parenchymal tissue is selected. In comparison, a previous study of 150 men and women reported volumes that were about 35% larger, based on a disc-summation method in MRI that excluded the renal pelvis and vasculature, with further validation by a water displacement method[9]. Another study of 1,852 men and women yielded volumes that were about 20% larger, based on a disc-summation method on manual delineations in MRI that excluded cysts and large vessels[29]. Values similar to those obtained in this work occur only in their reported lower quartile range of measurements. More similar values were obtained by previous studies that also focussed on the renal parenchyma, segmenting cortex and medulla only. For segmentations in CT of 1,344 men and women, the renal parenchyma was reported to be about 8% larger in a subgroup with similar mean age as in this work[30]. In yet another study with MRI of 50 men and women with renovascular disease, the reported volumes were only about 3% larger, based on manual segmentations and voxel count measurements in MRI, with only cortical and medullary tissue being included[31].

In terms of methodology, kidney segmentations with Dice scores of up to 0.974 have been reported in the literature for benchmark challenges involving neural networks on CT data[13,18]. Reaching comparable quality on the UK Biobank neck-to-knee body MRI may not be technically feasible, as the given images are of lower resolution and even repeat segmentation by human operators yielded lower consistency in this work. With no fixed image contrast, such as the Hounsfield units in CT, an objectively consistent placement of the kidney outline in MRI is more challenging. Nonetheless, it is possible that a 3D network architecture could reach superior performance. Future work may explore this potential, but will have to account for the massively increased runtime requirements for 3D architectures. Based on the reported runtimes[18], a 3D network may require up to an entire day for training as opposed to the 30 minutes for the 2.5 architecture used in this work, and a similar factor may apply to the inference. Competitive results have also been reported for other approaches that do not utilize neural networks. A recently published approach with appearance-guided deformable boundaries reached a mean Dice score of 0.95 with a 9.5% percentage error for total kidney volume in abdominal diffusion MRI of 72 men and women[12]. Whereas these metrics are similar to the validation results reached in this work, it is worth noting that their reported runtime would also amount to a total of almost two months for inference on 40,000 subjects as compared to less than a day required by the proposed segmentation pipeline. In an older technique based on adaptive region growing in CT of 30 subjects, comparable quality was only reached in the best case, with a mean Dice score of 0.88[10]. A more recent work with a multi-atlas technique reported a mean Dice of 0.952 for kidney segmentation in CT of 22 subjects[32]. The latter may not be directly comparable to the proposed pipeline however, as a rather convex segmentation style and about double the image resolution available in the UK Biobank was used. Another previous study on CT of ADPKD, segmented with fully convolutional neural networks similar to the one proposed in this work, reported a mean Dice score of 0.86 for three data of three different studies[14].

When training the network, dataset *B* was provided for additional guidance on the most challenging morphology. Nonetheless, severe anomalies such as renal fusion and suspected ADPKD were excluded and are thereby effectively accepted as failure cases of the proposed pipeline. This design decision was motivated by the concern that the network may learn a compromise, allowing for better results on these outliers while simultaneously performing worse on the majority of typical cases. The benefit of dataset *B* is not immediately clear from the validation results of Table 1, where the main result is actually outperformed by the cross-validation using only single-operator data. This is likely a side-effect of validating on references created by the operator of dataset *A*

only. The individual segmentation styles of the two operators of dataset *B* show a tendency towards overseg-mentation relative to the operator of dataset *A*, marking about 10% more volume. This tendency is emulated by the network trained on the combined datasets, which learned a compromise of segmentation styles. This compromise achieves a slightly lower agreement with the references of dataset *A*, but nonetheless appears more robust. Preliminary inference runs, which were trained on dataset *A* only, produced up to three times more *scrap volume* than the main configuration presented here.

The 2.5D U-Net was able to correctly segment one subject with a missing right kidney in dataset *A* dur-ing cross-validation, even though no comparable case existed in the training data. It is possible that the two-dimensional input format may have enabled the network to learn unilateral segmentation based on individual slices containing tissue of only one kidney, with the other being further above or below. Although no rigorous comparison was attempted here, the 2.5D modifications to the original U-Net[16] are estimated to accelerate train-ing by about 25% and increase the Dice score by about 0.02.

Several limitations apply. The neural networks trained in this work can only be expected to show comparable performance on future MRI with the same imaging protocol, type of MRI device, and subject demographics. When applied to data of other studies, new training data may be required. The given MRI data is arguably not optimal for kidney volumetry, being originally intended for body composition analysis[2]. With kidney tissue being typically contained in two breath-hold imaging stations, the measurement error is potentially compounded by artefacts such as motion and other factors that are not represented in the validation metrics. Even though the algorithmic quality ratings can be expected to identify the worst cases, actual correction may be possible with registration techniques for image mosaicing[33], which were not attempted here. Similarly, those cases excluded by the location rating could be trivially recovered by including the adjacent imaging stations in the pipeline. Another limitation is the degree to which the algorithmic quality ratings themselves are automated. With intui-tive, rule-based scores, they provide a high level of control over the exclusions and successfully identify the most severe outliers and failure cases. While the need for manual controls is thereby vastly reduced, the study of their distributions and choice of percentiles does require human intervention and would ideally be automated entirely. No guarantee is provided that all failure cases are exhaustively identified, or that the excluded cases are indeed inadequate. The conservative criteria chosen in this work exclude 5% of cases, which nonetheless translates to about 2,000 subjects for further inspection, many of which are presumably of acceptable quality. The current post-processing steps may furthermore yield misleading results in rare cases where severe cystic formations fragment the healthy kidney tissue such that objective delineation is no longer possible. In these cases, the two largest connected components may occur on the same side, leading to implausible distance and unilateral volume measurements.

The inference of high-quality measurements will therefore remain a continuous effort. A considerable advan-tage is posed by the high speed of the proposed pipeline, simplifying future coverage of newly released scans and repeated inference runs. With the latter, differently trained segmentation networks could potentially be applied for inference, with the model variation serving as a proxy for prediction uncertainty. Similarly, the collection of new reference segmentations could enable retraining of the network for separate segmentation of cortex, medulla, and cysts, or total volume that includes cysts to complement the measurements obtained in this work. With new post-processing steps it would also be possible to provide measurements of kidney length, width, and depth so that ellipsoid volumes could be studied. More elaborate quality controls could furthermore rely on independ-ent shape models or atlas segmentations, which have been previously used for large-scale quality controls of UK Biobank cardiovascular MRI segmentation in a variation of the concept of reverse classification accuracy[34].

While the collection of metadata and MRI acquisition by the UK Biobank are still in progress, the obtained measurements can already be provided as return data and used for further research. Whereas the currently available blood biochemistry and urine assays predate the MRI by several years, various fields on body size and composition are already available for association studies. The latter are often based on semi-automated process-ing of the same neck-to-knee body MRI as used in this work and do not yet cover all released subjects. Recent work on image-based regression with neural networks for biometry[24] can nonetheless provide accurate approxi-mations months or years before full coverage by the reference methods is achieved, so that many associations could already be studied. Likewise, genetic information is readily available and future work may also examine the repeat imaging visit as planned for another 10,000 subjects. The proposed pipeline is expected to successfully process these images without any need for changes, and the resulting measurements could enable further study of longitudinal effects and disease outcomes associated with changes in parenchymal kidney volume.

Conclusion

The proposed kidney segmentation pipeline generates fast, accurate, and objective delineations with close agree-ment to human operators. It was applied to all available UK Biobank neck-to-knee body MRI, with only 5% of results showing signs of potential measurement failure. Similar performance is expected for future UK Biobank releases, with the remaining results already forming a substantial sample of left and right parenchymal kidney volume measurements that can be shared for further medical research.

References

1. Sudlow, C. *et al.* UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Med.* **12**, e1001779. https://doi.org/10.1371/journal.pmed.1001779 (2015).

2. West, J. *et al.* Feasibility of MR-based body composition analysis in large scale population studies. *PLoS ONE***11** (2016). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5035023/.
3. Emamian, S. A., Nielsen, M. B., Pedersen, J. F. & Ytte, L. Kidney dimensions at sonography: Correlation with age, sex, and habitus in 665 adult volunteers. *AJR Am. J. Roentgenol.* **160**, 83–86 (1993).
4. Glodny, B. *et al.* Kidney fusion anomalies revisited: Clinical and radiological analysis of 209 cases of crossed fused ectopia and horseshoe kidney. *BJU Int.* **103**, 224–235 (2009).
5. Sharma, K. *et al.* Kidney volume measurement methods for clinical studies on autosomal dominant polycystic kidney disease. *PLoS One* **12**, 2 (2017).
6. Hoy, W. E. *et al.* Nephron number, glomerular volume, renal disease and hypertension. *Curr. Opin. Nephrol. Hypertens.* **17**, 258–265 (2008).
7. Rossing, P., Hougaard, P. & Parving, H.-H. Risk factors for development of incipient and overt diabetic nephropathy in type 1 diabetic patients: A 10-year prospective observational study. *Diabetes Care* **25**, 859–864 (2002).
8. Bakker, J., Olree, M., Kaatee, R., de Lange, E. E. & Beek, F. J. In vitro measurement of kidney size: Comparison of ultrasonography and MRI. *Ultrasound Med. Biol.* **24**, 683–688 (1998).
9. Cheong, B., Muthupillai, R., Rubin, M. F. & Flamm, S. D. Normal values for renal length and volume as measured by magnetic resonance imaging. *Clin. J. Am. Soc. Nephrol.* **2**, 38–45 (2007).
10. Lin, D.-T., Lei, C.-C. & Hung, S.-W. Computer-aided kidney segmentation on abdominal ct images. *IEEE Trans. Inf. Technol. Biomed.* **10**, 59–65 (2006).
11. Ali, A. M., Farag, A. A. & El-Baz, A. S. Graph cuts framework for kidney segmentation with prior shape constraints. In *International conference on medical image computing and computer-assisted intervention*, 384–392 (Springer, 2007).
12. Shehata, M. *et al.* 3d kidney segmentation from abdominal diffusion MRI using an appearance-guided deformable boundary. *PLoS One* **13**, 2 (2018).
13. Heller, N. *et al.* The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. arXiv preprint arXiv:1912.01054 (2019).
14. Sharma, K. *et al.* Automatic segmentation of kidneys using deep learning for total kidney volume quantification in autosomal dominant polycystic kidney disease. *Sci. Rep.* **7**, 1–10 (2017).
15. Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition* 3431–3440 (2015).
16. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241 (Springer, 2015).
17. Han, X. Automatic liver lesion segmentation using a deep convolutional neural network method. arXiv preprint arXiv:1704.07239 (2017).
18. Isensee, F. & Maier-Hein, K. H. An attempt at beating the 3d u-net. arXiv preprint arXiv:1908.02182 (2019).
19. Zheng, Q., Delingette, H., Duchateau, N. & Ayache, N. 3-d consistent and robust segmentation of cardiac images by deep learning with spatial propagation. *IEEE Trans. Med. Imaging* **37**, 2137–2148 (2018).
20. Bai, W. *et al.* Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *J. Cardiovasc. Magn. Reson.* **20**, 65 (2018).
21. Irving, B. *et al.* Deep quantitative liver segmentation and vessel exclusion to assist in liver assessment. In *Annual Conference on Medical Image Understanding and Analysis*, 663–673 (Springer, 2017).
22. Basty, N. *et al.* Automated measurement of pancreatic fat and iron concentration using multi-echo and t1-weighted MRI data. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 345–348 (IEEE, 2020).
23. Attar, R. *et al.* Quantitative cmr population imaging on 20,000 subjects of the uk biobank imaging study: Lv/rv quantification pipeline and its evaluation. *Med. Image Anal.* **56**, 26–42 (2019).
24. Langner, T., Ahlström, H. & Kullberg, J. Large-scale biometry with interpretable neural network regression on uk biobank body MRI. arXiv preprint arXiv:2002.06862 (2020).
25. Östling, A. Automated kidney segmentation in magnetic resonance imaging using u-net (2019).
26. Malmberg, F., Nordenskjöld, R., Strand, R. & Kullberg, J. Smartpaint: A tool for interactive segmentation of medical volume images. *Comput. Methods Biomech. Biomed. Eng.* **5**, 36–44 (2017).
27. Iglovikov, V. & Shvets, A. Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. ArXiv e-prints (2018). arXiv:1801.05746.
28. Bilic, P. *et al.* The liver tumor segmentation benchmark (lits). arXiv preprint arXiv:1901.04056 (2019).
29. Roseman, D. A. *et al.* Clinical associations of total kidney volume: The framingham heart study. *Nephrol. Dial. Transplant.* **32**, 1344–1350 (2017).
30. Wang, X. *et al.* Age, kidney function, and risk factors associate differently with cortical and medullary volumes of the kidney. *Kidney Int.* **85**, 677–685 (2014).
31. Gandy, S., Armoogum, K., Nicholas, R., McLeay, T. & Houston, J. A clinical MRI investigation of the relationship between kidney volume measurements and renal function in patients with renovascular disease. *Br. J. Radiol.* **80**, 12–20 (2007).
32. Yang, G. *et al.* Automatic kidney segmentation in CT images based on multi-atlas image registration. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 5538–5541 (IEEE, 2014).
33. Ceranka, J. *et al.* Registration strategies for multi-modal whole-body MRI mosaicing. *Magn. Reson. Med.* **79**, 1684–1695 (2018).
34. Robinson, R. *et al.* Automated quality control in image segmentation: Application to the uk biobank cardiovascular magnetic resonance imaging study. *J. Cardiovasc. Magn. Reson.* **21**, 18 (2019).

Author contributions

T.L. wrote the main manuscript text and conducted the experiments, A.Ö. conducted an early precursor of the single-operator cross-validation[25], L.M. provided the reference segmentations of dataset *A*, A.K. and D.O. provided the reference segmentations of dataset *B*, D.L., A.W., and L.L. implemented the 2.5D network architecture in a related project, R.S. and J.K. supervised the project and data access from the UK Biobank, H.A. provided medical expertise. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-77981-4.

Correspondence and requests for materials should be addressed to T.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.