Supplementary material, Table of Contents

Supplementary Methods

MRI data

All participants underwent standardized abdominal MRI-scans as part of the DIPAK-1 study [13]. Multiple sequences were scanned, but only the coronal fat saturated T2-single shot fast spin-echo sequence (HASTE) was used for this study. A more detailed description of the MRI-protocol has been published previously [14]. The manually traced kidney and liver volumes were assessed as secondary end point for the DIPAK-1 study [Meijer JAMA 2018, in press]. DICOM image data from the DIPAK-1 study was transferred to Mayo Clinic after anonymization and converted to the NIFTI file format by the dcm2nii software. The images have a reconstructed matrix size of 256 × 256 × X (with X being the number of slices that contained kidney or liver tissue, large enough to cover the full extent of the kidneys within the imaged volume). Image voxel sizes are most commonly on the order of 1.5 mm in-plane with 4 mm slice thickness and spacing between slices.

Deep learning model

The original convolutional neural network architecture[12] was extended to incorporate inception blocks with dimension reductions[14] and residual connections[15] to improve generalizability. The network architecture is shown in Supplementary Figure 1. The network input is implemented as a three-channel architecture consisting of the slice to be segmented as well as adjacent slices (posterior and anterior). The architecture consists of a series of inception block layers with dimensionality reduction, followed by strided convolutions (stride=2) and dropout (0.1). The inception block layer with dimensionality reduction consists of performing convolutions on the input in four different paths. The first path performs a 1x1 convolution, the second a 1x1 followed by a 3x3, the third a 1x1 followed by a 5x5, and the fourth max-pooling (3x3) followed by a 1x1 convolution. The outputs of each of these are then concatenated to form the inception block output. The activation function 'elu' was used throughout as well as BatchNorm. Residual connections are made between skip connections of the encoder-decoder layers. Number of kernels are doubled in each layer as 64, 128, 256, 512, and 1028 at the lowest resolution. The output is a 1x1 feature pooling convolution layer to predict whether the voxel pertains to the right kidney, left kidney, liver, or background. The model was trained with a customized Jaccard loss function, Adam optimizer with initial learning rate of 1.E-3, and batch size of 16. The model was implemented in Python using the Keras library and was run on an Nvidia Tesla P40 GPU. The model

was trained for 100 epochs. Each epoch took ~20min, and the total training time was 37 hours. Once the model was trained, inference took ~3 seconds per case.

Evaluation of automated approach

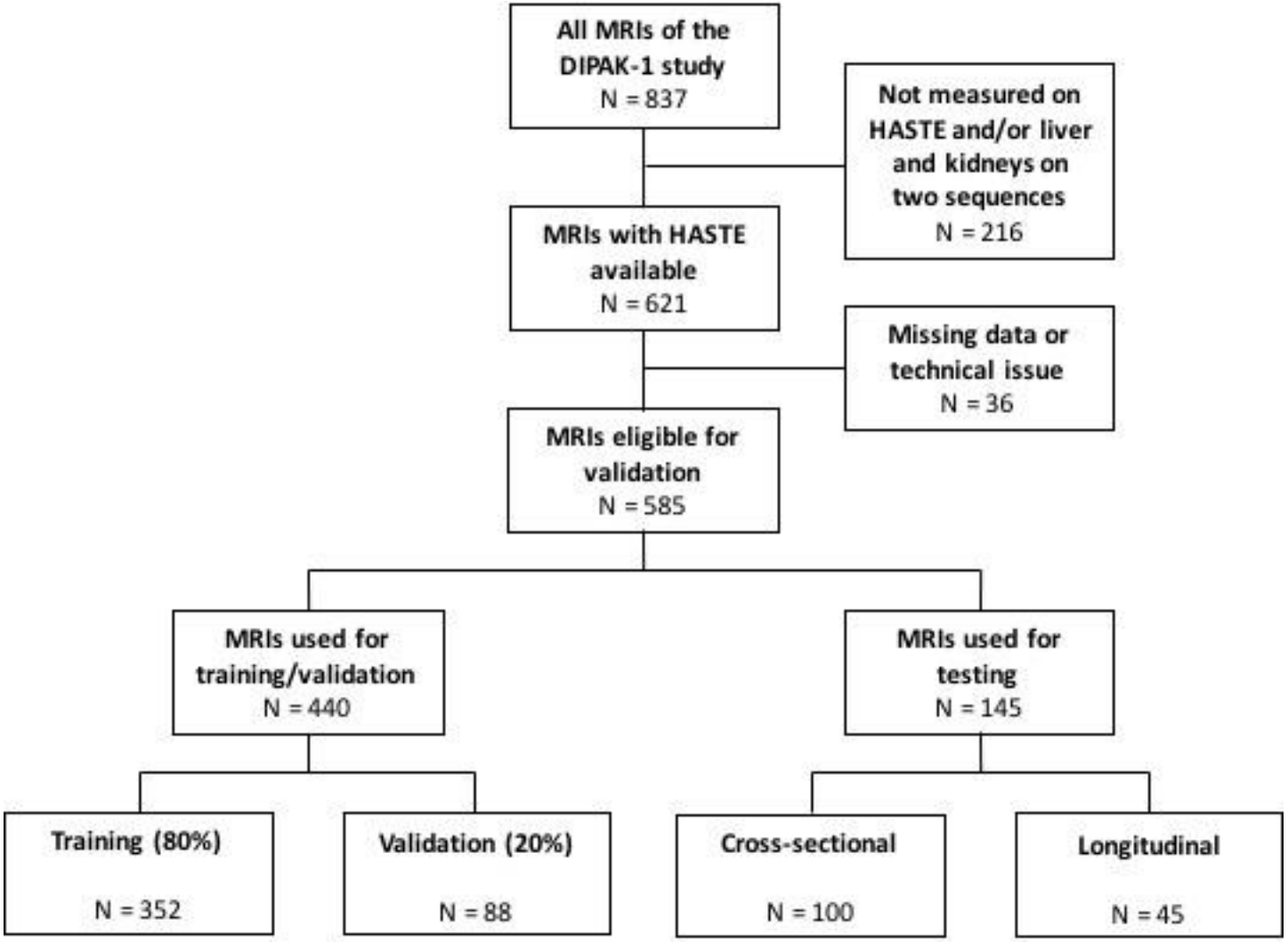
Comparison statistics were generated from the reference standard segmentations and those made by the automated approach. These comparison statistics included similarity metrics and comparison of total volume differences. For the similarity metrics, a number of commonly used metrics used to assess segmentation accuracy were calculated. These include the Dice coefficient (or similarity index) that is defined as: Dice = (2 * TP) / (2 * TP + FP + FN), where TP is true positives (i.e., both reference standard and automated approach classified a voxel as being the kidney), FP is false positives (i.e., automated approach falsely classified voxel as being the kidney), and FN are false negatives (i.e., automated approach falsely classified voxel as not being a part of the kidney), and the Jaccard coefficient (or overlap ratio), which is defined as: Jaccard = TP / (TP + FP + FN). These indices vary within the range 0 to 1, where a value closer to 1 indicates a closer similarity of the proposed method to the reference standard. Sensitivity, specificity, and precision were also calculated as well as surface distance measures. These included the mean surface distance (a measure of the average distance between the surfaces of the automated approach compared with the reference standard), as well as the Hausdorff distance (the largest difference between the surface distances).

Supplementary Table 1. Automated versus manual assessed TKV on stratification according to Mayo htTKV risk classes (A lowest risk, E highest risk). Information regarding the two misclassified cases is also given.

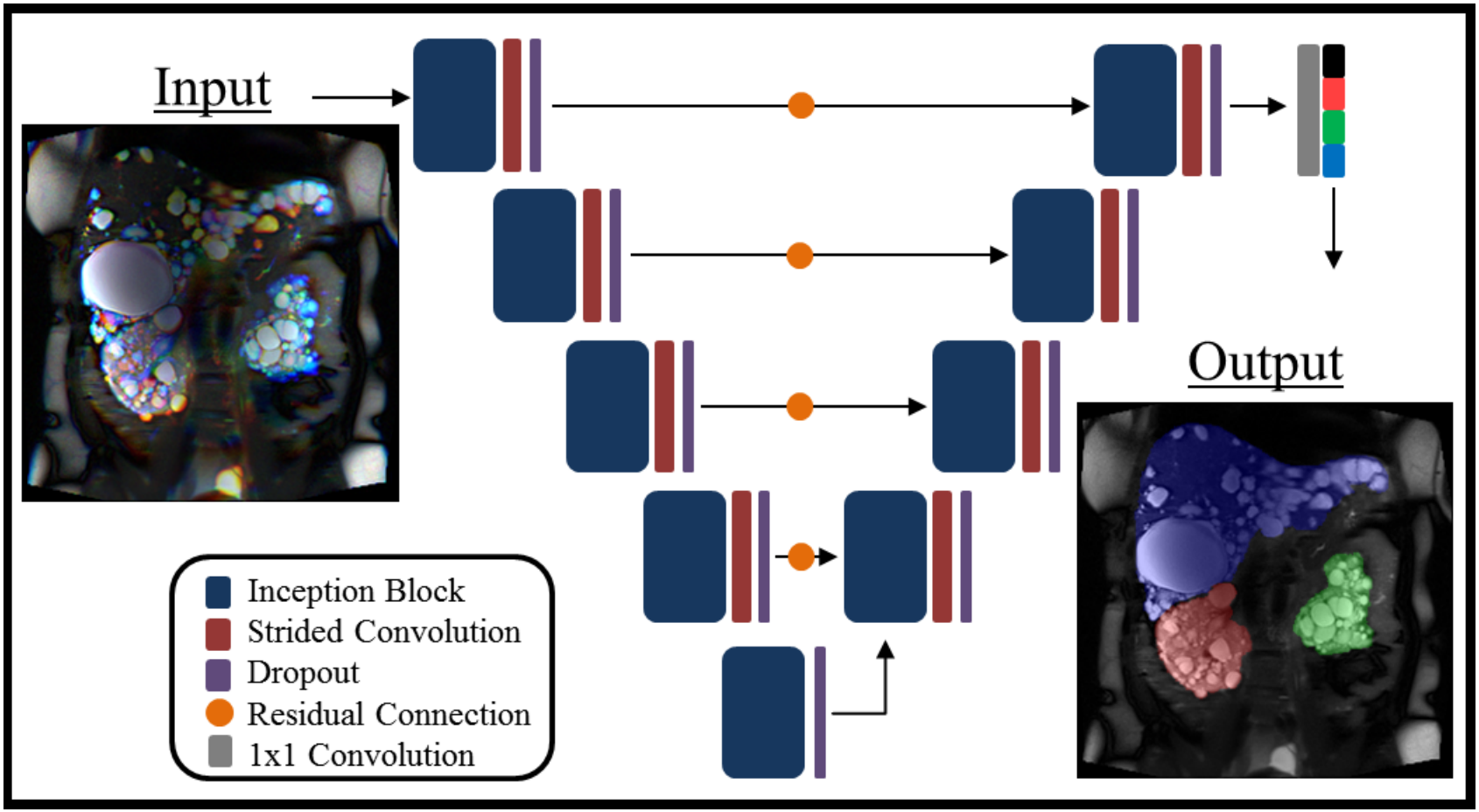
Patient	**1**	**2**
Height (m)	1.63	1.88
Age (y)	43	48
Automated hTKV	283.33	1268.23
Mayo risk class	1A	1D
Manual hTKV	308.45	1215.77
Mayo risk class	1B	1C
Cut-off point between risk classes *	284.53	1240.72

** for the given age*

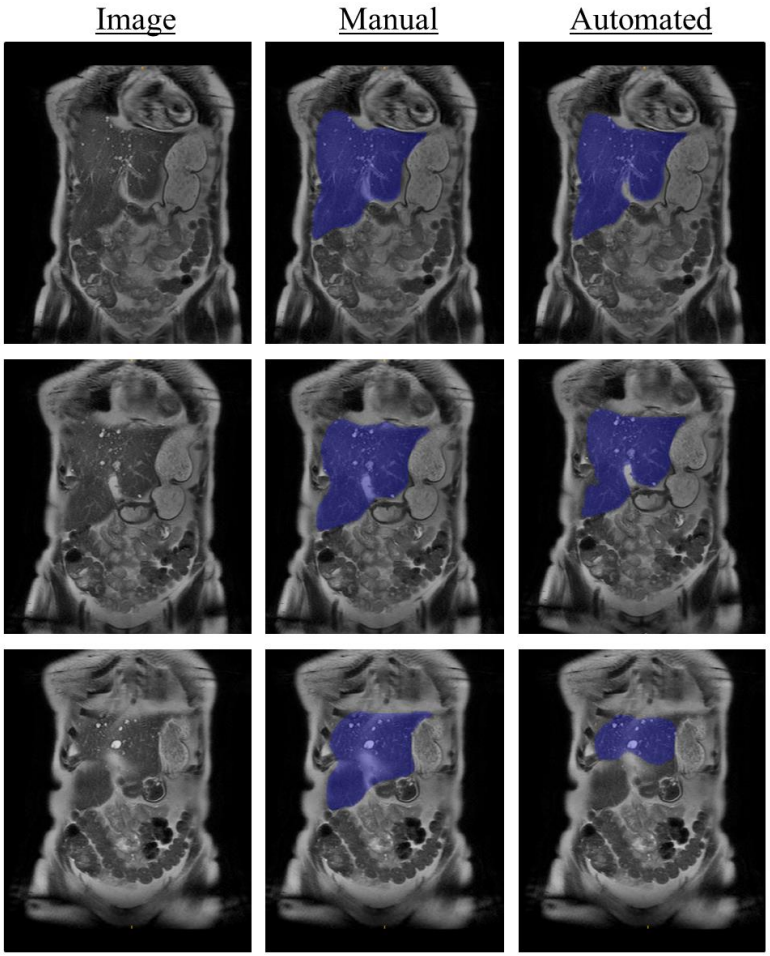
Supplementary Figure 1. Flowchart of the MRIs used.



Supplementary Figure 2. Schematic of the deep neural network architecture developed in this study. The network input is implemented as a three-channel architecture consisting of the slice to be segmented as well as adjacent slices (posterior and anterior). The architecture consists of a series of inception block layers, followed by strided convolutions (stride=2) and dropout (0.1). Residual connections are made between skip connections of the encoder-decoder layers. Number of kernels are doubled in each layer as 64, 128, 256, 512, and 1028 at the lowest resolution. The output is a 1x1 feature pooling convolution layer to predict whether the voxel pertains to the right kidney (red), left kidney (green), liver (blue), or background (black).



Supplementary Figure 3a. Examples of automated segmentations highlighting the case with the largest volume difference for liver within one patient (between manual and automated). One major source of variability is caused by the inclusion/exclusion of the portal vein (here the automated method did not include the portal vein).



Supplementary Figure 3b. Examples of automated segmentations highlighting the case with the lowest liver Dice score, which is a measure of accuracy and overlap between both methods, calculated using the amount of voxels that were positive for both methods (true positives) and the amount of voxels that were false negatives (automated method did not identify a voxel, whereas the gold standard method did) and false positives (vice versa). Another source of variability that was seen was the inclusion or exclusion of gall-bladder. Here the human reader included the gall bladder and the automated approach did not.