MELBA Journal Machine Learning for Biomedical Imaging

Robust Renal Mass Segmentation on CT: A Validation Study of an AI-Based Framework

Sarah **de Boer** [1], Hartmut **Häntze** [1,2], Kiran Vaidhya **Venkadesh** [1], Myrthe A. D. **Buser** [1], Gabriel E. **Humpire Mamani** [1], Lina **Xu** [2], Lisa C. **Adams** [4], Jawed **Nawabi** [3], Keno K. **Bressem** [4,5], Bram **van Ginneken** [1,6], Mathias **Prokop** [1], Alessa **Hering** [1]

1 Department of Medical Imaging, Radboudumc, Nijmegen, The Netherlands
2 Department of Radiology, Charité - Universitätsmedizin Berlin, Berlin, Germany
3 Department of Neuroradiology, Charité - Universitätsmedizin Berlin, Berlin, Germany
4 Department of Diagnostic and Interventional Radiology, Klinikum rechts der Isar, TUM University Hospital, Technical University of Munich, Munich, Germany
5 Department of Cardiovascular Radiology and Nuclear Medicine, German Heart Center, TUM University Hospital, Technical University of Munich, Munich, Germany
6 Fraunhofer MEVIS, Bremen, Germany

Abstract

Renal mass segmentation has important potential to enhance the clinical workflow, especially in settings requiring quantitative assessments. Kidney volume could serve as an important biomarker for renal diseases, with changes in volume correlating directly with kidney function. Currently, clinical practice often relies on subjective visual assessment for evaluating kidney size and kidney lesions, including tumors and cysts, which are typically staged based on diameter, volume, and anatomical location. To support a more objective and reproducible approach, this research aims to develop a robust, thoroughly validated renal mass segmentation algorithm, named Renal-Net. We employ publicly available training datasets and leverage the state-of-the-art medical image segmentation framework nnU-Net. Validation is conducted using both proprietary and public test datasets, with segmentation performance quantified by Dice coefficient and the 95th percentile Hausdorff distance. Furthermore, we analyze robustness across subgroups based on patient sex, age, CT contrast phases, and tumor histologic subtypes. Our findings demonstrate that our segmentation algorithm, trained exclusively on publicly available data, generalizes effectively to external test sets and outperforms existing state-of-the-art models across all tested datasets. Subgroup analyses reveal consistent high performance, indicating strong robustness and reliability. The developed algorithm and associated code are publicly accessible at `https://github.com/DIAGNijmegen/oncology-kidney-abnormality-segmentation`.

Keywords

Deep Learning, Medical Imaging, Segmentation, Kidney cancer, Renal Cell Carcinoma, Renal mass

Article informations

`https://doi.org/10.59275/j.melba.2026-67g5`

Volume 2026, Received: 2026-01-30, Published 2026-05-22

Corresponding author: sarah.deboer@radboudumc.nl

1. Introduction

Kidney cancer has a global incidence rate of approximately 400,000 new cases annually, leading to 175,000 deaths (Cirillo et al., 2024). It is often detected incidentally during imaging performed for unrelated medical reasons, most often in computed tomography (CT). Treatment options for suspected malignant renal masses include radical and partial nephrectomy (Motzer et al., 2022). Ablation techniques are also offered in various medical facilities as a less invasive treatment option for small tumors (Filippiadis et al., 2019; Mershon et al., 2020). For less aggressive tumors and for patients with reduced life expectancy due to other reasons like age or comorbidities, active surveillance is a treatment alternative (Capitanio and Montorsi, 2016; Kutikov et al., 2010). Given the variety

of available treatment options, patient stratification plays a crucial role in optimizing clinical decision-making. In clinical practice, the evaluation of the kidneys and lesions relies on the subjective visual assessment of kidney volume, tumor longest diameter and tumor location which are used to accurately stratify patients (Kutikov and Uzzo, 2009). Furthermore, in other kidney diseases, like autosomal dominant polycystic kidney disease (ADPKD), the assessment of total kidney volume (TKV) is an indicator for disease severity and disease progression (Gaur et al., 2019).

Manual segmentation of organs and tumors on CT is time consuming and the accuracy and reproducibility of the segmentations are subject to inter-rater variability (Joskowicz et al., 2019; Meyer et al., 2006). This might be caused by differences in experience levels, but also imaging characteristics. In the medical imaging field, many research has been done on AI-based segmentation of organs of interest and tumor regions (Bilic et al., 2023; Staal et al., 2004; Litjens et al., 2014). Within the last few years researchers have taken immense steps towards solving this task, starting with the U-Net (Ronneberger et al., 2015) and resulting in the nnU-Net (Isensee et al., 2021). The nnU-Net is an automated, self-configuring segmentation model that optimizes its hyperparameters based on a given annotated dataset and is trained in a supervised learning setting. Depending on the specific segmentation task, the size and characteristics of available training data, and hardware resources, users can select among multiple configurations provided by the nnU-Net. Multi-organ segmentation models built upon nnU-Net have emerged, with TotalSegmentator (Wasserthal et al., 2023) being the most widely adopted framework.

While general medical image segmentation models have made considerable progress, they may not always perform optimally for specialized tasks such as renal mass segmentation. In such contexts, organ-specific models can achieve superior performance through the use of carefully curated training data and the integration of domain-specific knowledge. Kidney tumors often considerably alter the organ's shape, size, and internal structures, posing challenges for general segmentation models primarily trained on healthy or standard anatomy. Such structural variations necessitate the development of dedicated AI methods tailored to specific clinical scenarios. For instance, studies have specifically addressed segmentation of large kidney tumors (Yang et al., 2022), while other works have targeted segmentation of kidney lesions in patients with ADPKD (Rombolotti et al., 2022). Beyond tumor and lesion segmentation, AI research has explored additional kidney-related segmentation tasks, such as delineating the renal cortex and medulla for assessing kidney donor candidacy (Korfiatis et al., 2022), and segmenting renal veins and arteries to facilitate surgical planning (Khan et al., 2025). Moreover, general tumor segmentation models, which are not specific to the kidneys, have also been investigated (Pang et al., 2020; de Grauw et al., 2025).

Kidney cancer diagnosis not only relies on CT scans, but Magnetic Resonance (MR) imaging is used for, among other reasons, determining tumor subtypes and characterizing kidney cysts (Ljungberg et al., 2019). U-Net based algorithms for kidney tumor and cyst segmentation on MRI have been proposed (Gregory et al., 2021; Haghighi et al., 2018; Zöllner et al., 2021), as well as segmentation algorithms for total kidney volume segmentation (Raj et al., 2022) and liver cyst segmentation for patients with ADPKD (Chookhachizadeh Moghadam et al., 2024). Additionally, kidney tumor models for MRI can be developed more quickly if equivalent CT models already exist (Häntze et al., 2025). Finally, the inclusion of Positron Emission Tomography (PET) imaging can improve kidney segmentation performance on CT (Leube et al., 2024).

The Kidney Tumor Segmentation (KiTS) challenge has considerably advanced the development of segmentation models for kidney cancer (Heller et al., 2021, 2023). Held three times to date, the challenge has produced numerous studies detailing methods for achieving high performance on the challenge dataset. The KiTS dataset was utilized to train an AI model to generate annotations for various cancer collections from the National Cancer Institute (NCI) Imaging Data Commons (IDC) (Murugesan et al., 2024). Furthermore, an nnU-Net model trained on the KiTS dataset and evaluated on in-house data, and vice versa, demonstrated a decline in external performance (Raman et al., 2025). This performance drop underscores the need for more heterogeneous training datasets and diverse, clinically relevant, external validation cohorts. The winners of the 2023 version of the KiTS challenge introduced Auto3DSeg (Myronenko et al., 2024), which is now integrated in the MONAI (MONAI, 2024) framework along with an implementation for training it on the KiTS dataset. However, recently, nn-UNet is shown to be the framework that leads to state-of-the-art performance on well known benchmarks, while the out-of-the-box performance of Auto3DSeg underperformed compared to nnU-Net (Isensee et al., 2024). On the KiTS23 dataset, nnU-Net ResEnc L achieved a Dice score of 88%, while Auto3DSeg SegResNet achieved 81%. The runner up on the KiTS23 challenge investigated different 3D-UNet setups and introduced a multi-scale post-processing strategy (Uhm et al., 2024). The third place of the challenge introduced a cascaded approach where an low resolution network first segments the kidneys and an ROI is used as input to the second network making high resolution predictions (George, 2022).

This research is designed as a large-scale validation study rather than a methodological contribution, addressing a critical gap between algorithmic development and clinical adoption (Chouvarda et al., 2025; Mercaldo et al.,

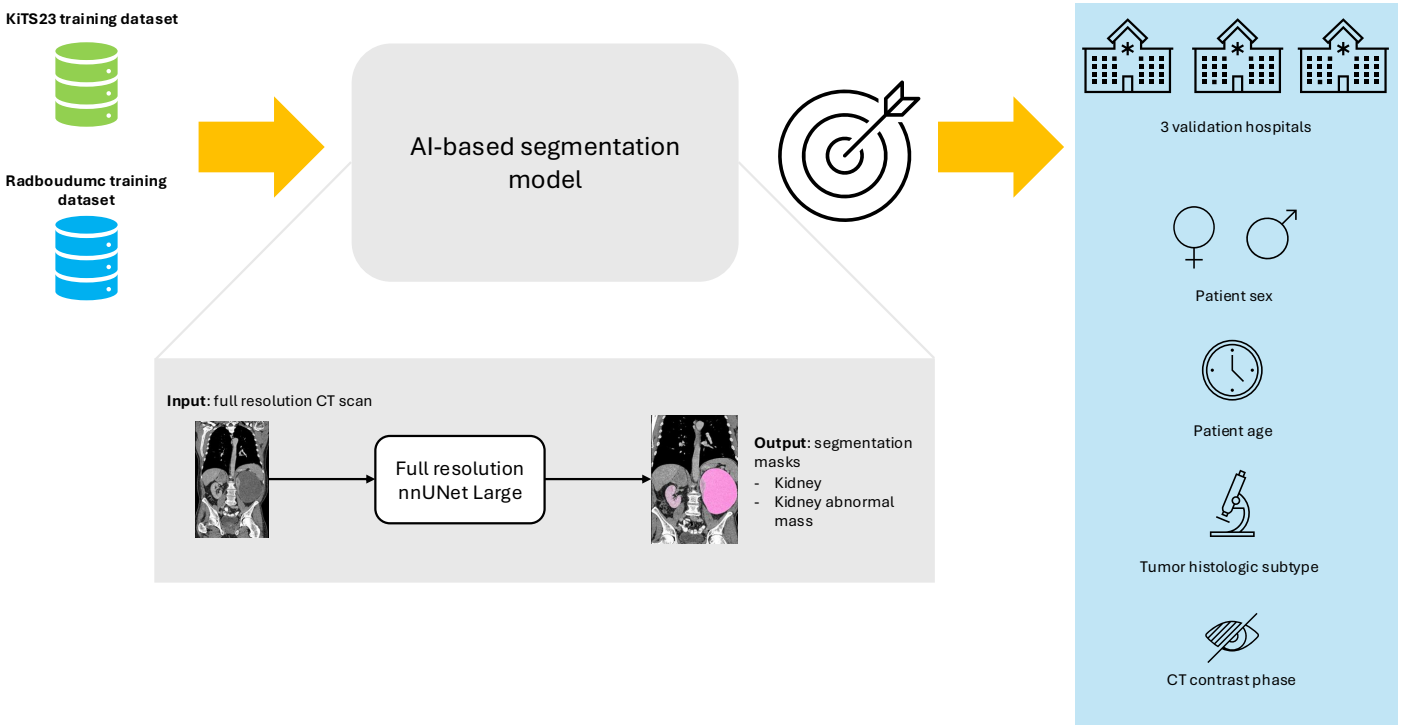

Figure 1: Overview of the research. The proposed AI-based segmentation model is trained on two publicly available datasets. The proposed method consists of a trained nnU-Net that segments the kidney and lesions in the kidney if present. We thoroughly validate the proposed method on three test datasets from three different medical centers, and test performance across patient sex, patient age, tumor histologic subtype and CT contrast phase.

2025). Our work makes several distinctive contributions: (1) We develop a robust kidney and renal mass segmentation model (named Renal-Net) that maintains performance across diverse patient populations and imaging protocols; (2) We provide comprehensive validation across three independent datasets totaling over 1,500 CT scans, representing, to the best of our knowledge, the largest external validation in this domain; (3) We conduct detailed subgroup analyses to identify potential biases in model performance across patient demographics, imaging parameters, and tumor characteristics; and (4) We make our validated model publicly available to facilitate clinical testing and further research. The algorithm is available on grand-challenge.org for easy use (`https://grand-challenge.org/algorithms/kidney-abnormality-segmentation/`) and our code is available on Github: `https://github.com/DIAGNijmegen/oncology-kidney-abnormality-segmentation`.

2. Methods

In this section, the training (Section 2.1) and testing (Section 2.2) datasets are introduced. We outline how we combined the different training datasets and show the characteristics of all datasets used in this study. Furthermore, we introduce the deep learning framework (Section 2.3) and layout our post-processing steps. The methodology is outlined in Figure 1.

2.1 Training Datasets

2.1.1 KiTS23 public training dataset

The 2023 Kidney Tumor Segmentation challenge training dataset consists of 489 CT scans coming from patients who underwent cryoablations, partial nephrectomy or radical nephrectomy for suspected renal cancer between 2010 and 2022 at a medical center in the United States of America (Heller et al., 2023). The most recent contrast-enhanced preoperative scan in either the corticomedullary or nephrogenic phase was selected and annotated by a pool of experts, trainees and lay-people using the following labels: kidney, tumor and cyst. The kidney region includes all parenchyma and the non-fat tissue within the hilum. The tumor regions comprises of masses found on the kidney that were suspected of being malignant and the cyst regions comprise of masses that were radiologically or pathologically determined to be cysts. The average age of the patients in the dataset is 59 ($\pm$ 14) years, with 37% identifying as female. 91% of the tumors is histologically confirmed to be malignant.

Table 1: Data set characteristics, for the training (KiTS and Radboudumc training set) and testing (Radboudumc test set, TCGA-KIRC, and Charité) datasets. Besides country of origin and number of scans, presented are the mean values averaged over the cases followed by the standard deviations. Lesion and kidney volumes are averaged over the number of lesions and kidneys in that case. Kidney volumes include lesion volumes. Although the Radboudumc dataset contains cases without lesions, these lesion-free cases were excluded when calculating the average lesion count and lesion volumes. However, these cases were included when calculating average kidney volumes, resulting in lower average kidney volumes compared to the other datasets.

	Training		Testing		
	KiTS	Radboudumc training set	Radboudumc test set	TCGA-KIRC	Charité Universitäts-medizin Berlin
Country	USA	Netherlands	Netherlands	USA	Germany
Number of scans	489	215	50	28	1510
Number of lesions	2.7 ± 2.8	3.0 ± 5.0	3.9 ± 2.8	3.3 ±2.6	2.0 ± 3.0
Lesion volume ($\times 10^3$ mm^3)	101 ± 236	17 ± 84	15 ± 47	136 ± 185	118 ± 218
Kidney volume ($\times 10^3$ mm^3)	278 ± 153	182 ± 104	132 ± 64	268 ± 158	277 ± 232
In-plane resolution (mm)	0.80 ± 0.11	0.75 ± 0.07	0.76 ± 0.06	0.77 ± 0.08	1.0 ± 0.0
Slice thickness (mm)	3.35 ± 1.70	1.22 ± 0.37	1.17 ± 0.38	2.81 ± 1.48	1.0 ± 0.0

2.1.2 Radboudumc public dataset

The KiTS23 dataset is derived solely from U.S.-based sources, limiting its generalizability. To enhance model robustness, we incorporated additional external data from the publicly available Kidney Abnormality dataset (Humpire-Mamani et al., 2023; Mamani et al., 2023), collected at Radboudumc Nijmegen in the Netherlands. This dataset contains in total 215 contrast-enhanced thorax-abdomen CT scans from oncology patients. Out of the 215 scans, 113 cases are without lesions in the kidney, and 102 cases are with one or more kidney lesions, including cysts, lesions, masses, metastases, and tumors. Seventeen patients underwent left nephrectomy, and eighteen underwent right nephrectomy. The kidney region was annotated by radiologists and medical students, encompassing the renal cortex, medulla, and pyramid, while the renal mass region included only masses connected to the kidney parenchyma. Lesions in the collecting system were excluded. The patient cohort contains 44% of patients identifying as female, with an average age of 60 years, ranging from 22 to 84 years (Mamani et al., 2023).

2.1.3 Merging of training datasets

To combine the KiTS23 dataset with the Kidney Abnormality dataset from Radboudumc (hereafter referred to as the Radboudumc training dataset), a few adjustments were made. The primary difference between the datasets lies in their annotation protocols, for full annotation protocols see (Heller et al., 2021, 2023; Mamani et al., 2023). The KiTS23 dataset distinguishes between cysts and tumors, while in the Radboudumc dataset these categories were combined. To ensure consistency across datasets, the cyst and tumor masks in the KiTS23 dataset were merged into one category, aligning with the Radboudumc dataset. Besides practical challenges, the clinical boundary between renal tumors and cysts is inherently ambiguous. Kidney lesions include malignant solid tumors, benign solid masses (e.g., oncocytoma) (Kang et al., 2014), and cystic masses ranging from simple cysts (Bosniak categories I and II) to complex lesions with malignant potential (Bosniak categories III and IV) (Silverman et al., 2019). In this study, we therefore define the semantic segmentation task of background, kidney, and renal masses.

Secondly, the annotation protocols for the kidney region differ between the two datasets: KiTS includes the hilum as part of the kidney region, whereas Radboudumc excludes it. An example illustrating these differences is provided in Figure 2. We decided not to re-annotate either of the kidney regions.

2.2 Test datasets

The robustness of the proposed deep learning framework was evaluated using one publicly available dataset and two proprietary datasets. Table 1 presents the characteristics of the datasets and compares them to the training set.

2.2.1 Radboudumc private test set

Alongside the publicly available Radboudumc kidney abnormality dataset, a separate private test set was collected. This set consists of 20 cases without renal masses and 30 cases with renal masses. The set is annotated following the same protocol used for the public dataset. Additionally, to enable assessment of inter-reader variability, an independent observer (an experienced medical student) also segmented

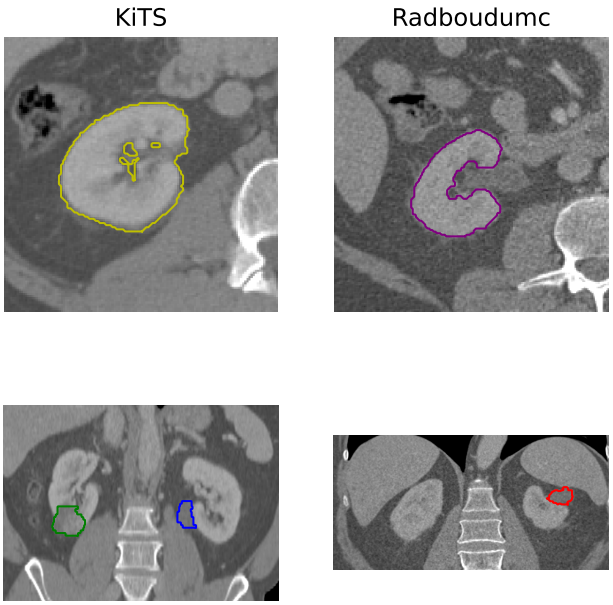

Figure 2: Visualization of the annotation protocols of the two training datasets. On the left, a segmentation mask from the KiTS dataset is shown in ■, which shows that the hilum is included in the annotation. On the right, a segmentation mask from the Radboudumc dataset is shown in ■, which shows that the hilum was left out of the kidney region while annotating. In the second row, we visualize that KiTS annotations distinguish between cysts ■ and tumors ■ and Radboudumc annotations only include an abnormal mass region ■.

all cases in this private set. This inter-observer annotation facilitates evaluating the model's performance relative to human observer variability. The patient characteristics distribution is the same as reported for the Radboudumc training set.

2.2.2 TCGA-KIRC

Under the AIMI Annotations initiative, a subset of The Cancer Genome Atlas Kidney Renal Clear Cell Carcinoma Collection (TCGA-KIRC) (Akin et al., 2016) is annotated for kidney, tumor and cyst regions, first by a trained AI model and corrected by an experienced radiologist (Murugesan et al., 2024). This set contains 28 valid imaging-mask pairs all from patients with clear cell renal cell carcinoma, incomplete imaging-mask pairs were excluded from analysis. The cohort is 41% female, with a mean age of 57 (± 10) years.

2.2.3 Charité Universitätsmedizin Berlin private test set

The second proprietary dataset originates from Charité Universitätsmedizin Berlin, Germany. It comprises 1510 preoperative CT scans from 917 patients who underwent nephrectomy due to suspected renal tumors. The dataset includes scans from the early venous (n=764), delayed venous (n=220), and arterial (n=526) contrast phases. Each scan is accompanied by annotations for one kidney, including tumors and cysts, performed by a medical student and two radiologists with five and four years of experience, respectively. The contra-lateral kidney and metastases are not annotated. The dataset includes the following tumor subtypes: clear cell renal cell carcinoma (ccRCC, n=1069), papillary renal cell carcinoma (pRCC, n=216), chromophobe renal cell carcinoma (chrRCC, n=74), renal oncocytoma (RO, n=100), and other rare tumor subtypes (n=51). The patient cohort is 30% female, with a mean age of 63 (± 11) years.

2.3 Deep learning framework

The deep learning method used in this paper is the nnU-Netv2 framework (Isensee et al., 2021, 2024). The nnU-Net framework eliminates the need for model architecture tweaking and hyperparameter tuning. The framework expects a supervised training dataset, and based on fixed parameters, rule-based parameters and empirical parameters the optimal network architecture and training parameters are determined. In this research, we used the newly introduced Residual Encoder (ResEnc) presets (Isensee et al., 2024), in particular the L(arge) version. Both full-resolution and low-resolution configurations were trained, along with a cascade model. Cross-validation results indicated that the full-resolution model performed best for this task. The resulting final model setup is presented in Table 4.

The model (named Renal-Net) is made available in two ways, by making the model weights and the code to run the model accessible on Zenodo(`https://doi.org/10.5281/zenodo.15315330`) and Github (`https://github.com/DIAGNijmegen/oncology-kidney-abnormality-segmentation`) and by providing a ready to use algorithm on `grand-challenge.org` (`https://grand-challenge.org/algorithms/kidney-abnormality-segmentation/`).

2.3.1 Postprocessing

After the prediction is made by the nnU-Net algorithm, a postprocessing pipeline is executed. Predicted renal mass regions that are not attached to the kidney region were removed. However, if the volume of the predicted lesion is bigger than 100 cm^3, it was kept. We noticed that large tumors can suppress the kidneys to such an extend that the model does not recognize it as kidney anymore. This threshold is based on the average kidney volume, measured on ultrasound imaging, 146 cm^3 for the left kidney and 134 cm^3 for the right kidney (Emamian et al., 1993). For predicted masses attached to the kidney, only those with an axial diameter greater than three mm were retained.

According to the RECIST criteria (Eisenhauer et al., 2009), lesions smaller than ten mm are not considered target lesions, and those smaller than five mm are regarded as non-measurable. A threshold of three mm was chosen to allow the inclusion of some very small lesions while maintaining a low rate of false positives.

3. Experiments

Model training was done on a single NVIDIA A100 40 GB GPU. In our experiments, we used the ensemble model of the 5-fold cross validation training.

3.1 Evaluation

The proposed model (Renal-Net) is evaluated on multiple grounds. First, we evaluated the segmentation and detection performance on the three independent test datasets comparing the model with and without post-processing. Secondly, for the TCGA-KIRC and Charité Universitätsmedizin Berlin test sets we performed subgroup analysis to assess potential performance variations.

Performance is reported for three regions: (1) kidney only, (2) kidney + masses, and (3) renal masses only. This structured evaluation enables a clear comparison, isolating kidney segmentation performance - particularly relevant for comparison with TotalSegmentator, which was trained on healthy kidneys - and assessing our model's robustness in segmenting kidneys without lesions. Following the KiTS challenge evaluation (Heller et al., 2023), performance metrics are also reported on the combined segmentation of kidneys and lesions.

3.1.1 Reference models

To benchmark Renal-Net, we compared it to two publicly available models, TotalSegmentator (Wasserthal et al., 2023) and the BAMF model (Murugesan et al., 2024), which is the only publicly available model specifically trained for renal mass segmentation. The BAMF model, similar to the proposed model, is trained primarily on CT.

TotalSegmentator (Wasserthal et al., 2023) provides separate labels for right and left kidney as well as for cysts in each kidney. To align with our evaluation approach, we combined right and left kidney regions into one kidney region and right and left kidney cysts into renal mass region.

Similar to the proposed model, the BAMF model is nnU-Net-based and specifically trained for kidney tumor segmentation. The training dataset includes the KiTS dataset and the TCGA-KIRC dataset. For the latter, annotations were obtained via active learning, where model predictions on unlabeled cases were expert-corrected and iteratively added for retraining.

3.1.2 Radboudumc test set

We compared the performance of our proposed model with TotalSegmentator (Wasserthal et al., 2023) and the BAMF model (Murugesan et al., 2024) on the Radboudumc test sets. The Radboudumc private test dataset contains two subsets, B20 with 20 patients who have healthy kidneys, and B30 with 30 patients who have at least one lesion in the kidney. Performance is reported on the two subsets separately.

3.1.3 TCGA-KIRC test set

To evaluate the proposed model on the TCGA-KIRC test set, we compared its performance against TotalSegmentator. We excluded the BAMF model from this comparison because TCGA-KIRC is part of its training set, which would bias the evaluation.

3.1.4 Charité Universitätsmedizin Berlin test set

Due to the relatively large size of the Charité Universitätsmedizin Berlin test set and the previously observed poor performance on the renal mass region on the other two test sets, TotalSegmentator was excluded from this analysis.

In our proposed model, segmentation was performed for both the left and right kidney. However, for the Charité dataset, annotations were available for only a single kidney per scan, specifically the kidney with a lesion. To address this, we limited our analysis to the annotated kidney. Specifically, overlap between each predicted region and the available annotation was assessed (overlap was considered if at least one voxel overlapped) and only the region that overlapped with the reference was selected for evaluation. If no overlap was found, an evaluation score of zero was assigned.

3.2 Statistical tests

To investigate whether the proposed model significantly outperforms baseline methods in terms of Dice score, the Mann-Whitney U test was performed. The Mann-Whitney U test was selected over rank-based alternatives, as it operates on the full score distributions rather than average ranks, providing greater statistical power for pairwise superiority testing. We tested the one-sided alternative hypothesis that the proposed model achieves superior Dice scores relative to each comparison model. We only performed the significance test for the Dice metric, as this is our main performance metric. A hierarchical statistical analysis plan was used, where we first tested for superiority over TotalSegmentator and if this condition is met we tested for superiority over the BAMF-model. Since TotalSegmentator is not trained specifically for kidney lesions, significance testing was not

performed for this comparison. We corrected for multiple testing by using the Holm-Bonferroni adjustment for the multiple datasets and segmentation regions we tested for each model comparison. A significance threshold (α) of 0.05 was used.

3.3 Segmentation performance

The Dice metric was used as overlap-based metric to compare with previous research but a boundary-based metric was added as is recommended in the Metrics Reloaded paper (Maier-Hein et al., 2024). For boundary-based metrics we report Hausdorff distance at 95 percentile.

3.4 Detection performance

Detection performance of renal masses was assessed by reporting false-positive and false-negative findings of the model. A renal mass was considered successfully detected if the overlap between the predicted region and the annotated region exceeded a specified threshold. In this study, an overlap threshold of 0.5 was used, with additional results at a threshold of 0 provided in Appendix C. The overlap is defined as the intersection over union. In this study, the main focus was segmentation performance and detection performance is derived from segmentation predictions. We therefore focus on precision, recall and F1-scores rather than doing FROC analysis.

3.5 Subgroup analysis

We performed subgroup analyses for the TCGA-KIRC and Charité Universitätsmedizin Berlin test datasets to identify potential biases in model performance. The Radboudumc test set was excluded from this analysis due to the absence of detailed patient-level characteristics We investigated the robustness across different subgroups of patient sex, patient age, contrast phase of the scan and tumor histologic subtype.

4. Results

Segmentation and detection outcomes are reported separately, with segmentation performance provided for each anatomical region. Significance test are performed solely for the Dice metric. In Appendix B, we present all segmentation metrics for each test set in tables.

4.1 Segmentation results

4.1.1 Healthy kidneys

The Radboudumc private test set contains a subset (B20) with patients with healthy kidneys. Figure 3 presents the results of the proposed model (Renal-Net) and the reference models on this test dataset. The inter-observer variability, as reported by an independent human observer with a mean Dice of 0.94 ($\pm$0.01) on the B20 subset (Mamani et al., 2023), is comparable to the performance achieved by the proposed model (0.93 $\pm$ 0.06) and TotalSegmentator (0.95 $\pm$ 0.01), while the BAMF model achieves 0.90 $\pm$ 0.02. The proposed model does not significantly outperform TotalSegmentator (corrected p-value $>$ 0.05) based on Dice on the healthy kidney subset, therefore superiority over the BAMF-model is not tested. Based on the performance measured in HD95, TotalSegmentator and the proposed model achieve mean scores of 7.39 $\pm$ 23.71 and 55.80 $\pm$ 101.21 respectively. From the boxplots in Figure 3 it can be observed that the median HD95 of the proposed model is close to the median HD95 of TotalSegmentator, however the quartiles differ substantially indicating a few cases where false positives for structures physically far from the kidney region are segmented.

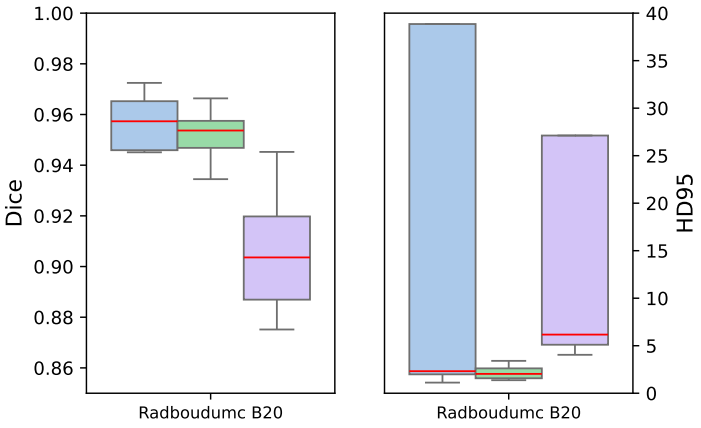

Figure 3: Segmentation results on the healthy cohort of the Radboudumc test set expressed in Dice (left) and Hausdorff distance (mm) at 95th percentile (right). We present Renal-Net (ours) ■, TotalSegmentator ■ and the BAMF model ■. The boxplots show the median (horizontal red line in the box), quartiles (presented by the box) and the full distribution (presented by the whiskers). For the purpose of readability we omitted the outliers from these plots.

4.1.2 Kidney region

The first row in Figure 4 presents the segmentation results for the kidney region. Median Dice scores for all models across the three datasets are around 0.90. The proposed model achieved the highest mean Dice scores on both the TCGA-KIRC (0.95 $\pm$ 0.03) and Radboudumc (0.94 $\pm$ 0.05) datasets, outperforming TotalSegmentator (0.80 $\pm$ 0.14 and 0.92 $\pm$ 0.10). On the Radboudumc dataset, the proposed model also outperformed the BAMF model (0.90 $\pm$ 0.04). Inter-observer variability yielded mean Dice scores of 0.925

(±0.051) on this test set (Mamani et al., 2023).

For HD95, the proposed model demonstrated substantially lower values on the TCGA-KIRC dataset (3.79 ± 5.89) compared to TotalSegmentator (25.90 ± 32.41). A similar trend as for the healthy kidneys was observed on the Radboudumc test set, where the proposed model segments for a few cases structures physically far from the kidneys resulting in a higher HD95 score (44.32 ± 99.98).

On the Charité Universitätsmedizin Berlin private test set, the proposed model and BAMF model reached comparable performance, both in terms of mean Dice (0.85 ± 0.10 and 0.83 ± 0.13, respectively) and HD95 (9.84 ± 16.11 and 12.19 ± 18.90, respectively).

Across all datasets for the kidney region, the proposed model yielded significantly higher Dice scores than the reference models (corrected p-values < 0.001).

4.1.3 Kidney+Mass region

The second row in Figure 4 shows the results on the kidney+mass region. We see very similar trends as for the kidney region alone. The proposed model outperforms TotalSegmentator on the TCGA-KIRC dataset. On the Radboudumc test set, the BAMF model underperforms compared to both the proposed model and TotalSegmentator. Inter-observer variability yielded mean Dice scores of 0.941 (±0.017) on the Radboudumc test set (Mamani et al., 2023), where the proposed model achieved 0.95 (±0.02). The proposed model and the BAMF model performed similarly in terms of Dice on the Charité Universitätsmedizin Berlin private test set. Across all datasets for the kidney+mass region, the proposed model achieved significantly higher Dice scores than the reference models with corrected p-values < 0.001.

The mean HD95 scores as presented in Appendix B highlight that the proposed model (43.66 ± 100.00) and the BAMF model (70.58 ± 88.95) segment for a few cases regions physically far from the kidney region, while TotalSegmentator (15.40 ± 37.83) does not show the same pattern.

4.1.4 Renal mass region

Lastly, the renal mass results are presented in the bottom row of Figure 4. Both the Dice and HD95 scores are more widely spread, but the proposed model consistently outperforms TotalSegmentator and the BAMF model. TotalSegmentator performs much worse on the TCGA-KIRC dataset compared to the proposed model both on the mean Dice (0.05 ± 0.13 and 0.86 ± 0.25, respectively) and mean HD95 metric (120.18 ± 61.30 and 14.94 ± 25.81, respectively). On the Radboudumc test set, the proposed model and the BAMF-model both have varying performance on the renal mass region when looking at the mean Dice score

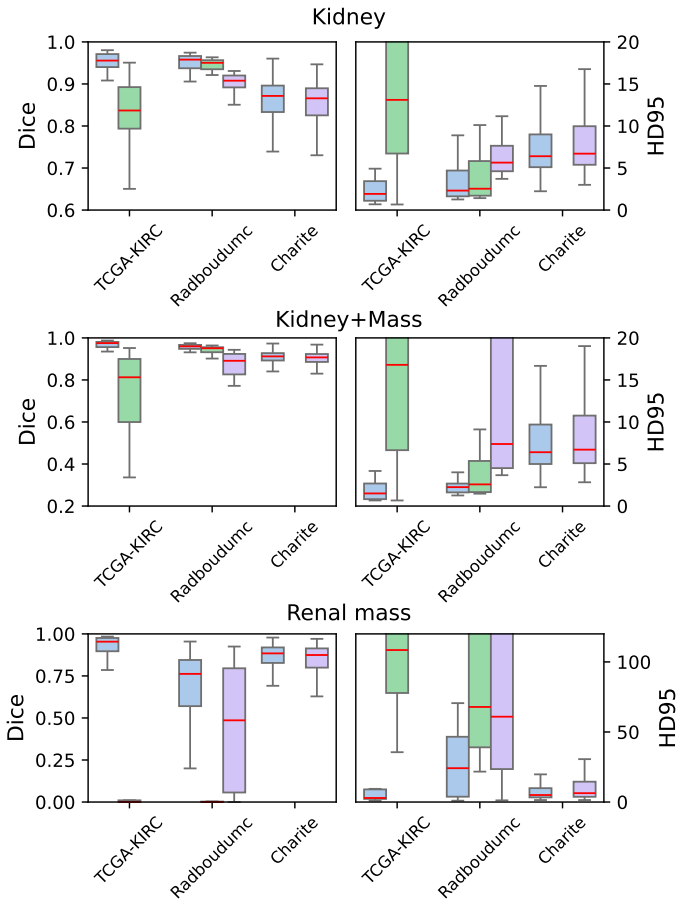

Figure 4: Segmentation results expressed in Dice (left column) and Hausdorff distance (mm) at 95th percentile (right column). The results are shown for three segmentation regions, Kidney, Kidney+Mass and Renal mass. We present results per dataset (excluding the patients with healthy kidneys in the Radboudumc test set) and show Renal-Net (ours) ■, TotalSegmentator ■ and the BAMF-model ■. The boxplots show the median (horizontal red line in the box), quartiles (presented by the box) and the full distribution (presented by the whiskers). For the purpose of readability we omitted the outliers from these plots. The BAMF-model is excluded in the TCGA-KIRC analysis due to training data overlap and TotalSegmentator is excluded from the Charité analysis due to observed poor performance on the other two test sets.

(0.66 ± 0.29 and 0.45 ± 0.35 respectively), although the proposed model on average scores better than the BAMF-model. Next to that, the HD95 metric shows superior performance of the proposed model. Inter-observer variability yielded a mean Dice score of 0.664 (±0.274) for the renal mass region (Mamani et al., 2023).

Similarly to the other tested segmentation regions, the performance of the proposed model and the BAMF model on the Charité Universitätsmedizin Berlin private test set is close together. Measured in mean Dice the proposed model

and the BAMF model reach 0.83 ± 0.17 and 0.81 ± 0.19 respectively, in HD95 the models reach 10.75 ± 16.86 and 13.43 ± 18.47 respectively. To investigate the differences in performance between the models one step further we compare the 5th percentile, which focuses on the difficult cases. These results are presented in Table 2. We can see that the BAMF-model is under performing compared to the proposed model on all regions but especially the renal mass region (5th percentile Dice of 0.53 and 0.38 for the proposed and BAMF-model). By looking at the 95th percentile we can see that this advantage is not followed by worse performance on the easy cases (both models achieve 0.95 Dice on the 95th percentile).

The proposed model achieved significantly higher Dice scores than the reference models across all datasets for the renal mass region with corrected p-values < 0.001.

Table 2: Quantile investigation of the model performance on the Charité Universitätsmedizin Berlin test set. Dice scores are reported at the 5th and 95th percentiles to characterize the distribution of segmentation performance across patients, capturing both typical failure cases and upper performance bounds. PP = post-processing, best values per metric per region are presented in bold.

Kidney	Dice 5th percentile	Dice 95th percentile
BAMF-model	0.65	0.91
Renal-Net (ours)	0.71	0.92
Renal-Net + PP (ours)	**0.71**	**0.92**
Kidney+mass		
BAMF-model	0.78	0.94
Renal-Net (ours)	0.84	0.95
Renal-Net + PP (ours)	**0.84**	**0.95**
Renal Mass		
BAMF-model	0.39	0.95
Renal-Net (ours)	0.53	0.95
Renal-Net + PP (ours)	**0.53**	**0.95**

4.2 Detection results

Table 3 presents the detection results of the proposed model and the reference models on all test sets. We present the detection results using an overlap threshold of 0.5, based on intersection over union, the results for a lower threshold can be found in Appendix C. The proposed model outperforms the reference models on all three metrics on all three test datasets. The difference in scores between the datasets highlights the different nature of the three test datasets.

4.3 Subgroup analysis

Subgroup analysis was performed on the TCGA-KIRC and Charité Universitätsmedizin Berlin private test sets using the proposed model (results of the BAMF-model can be found in Appendix D). We investigated the robustness across different subgroups of patient sex, patient age, contrast phase of the scan and tumor histologic subtype. Figure 5 presents the results for the TCGA-KIRC test set, Figure 6 presents the results for the Charité test set. An additional subgroup analysis of the BAMF model on the Charité test set can be found in Appendix D.

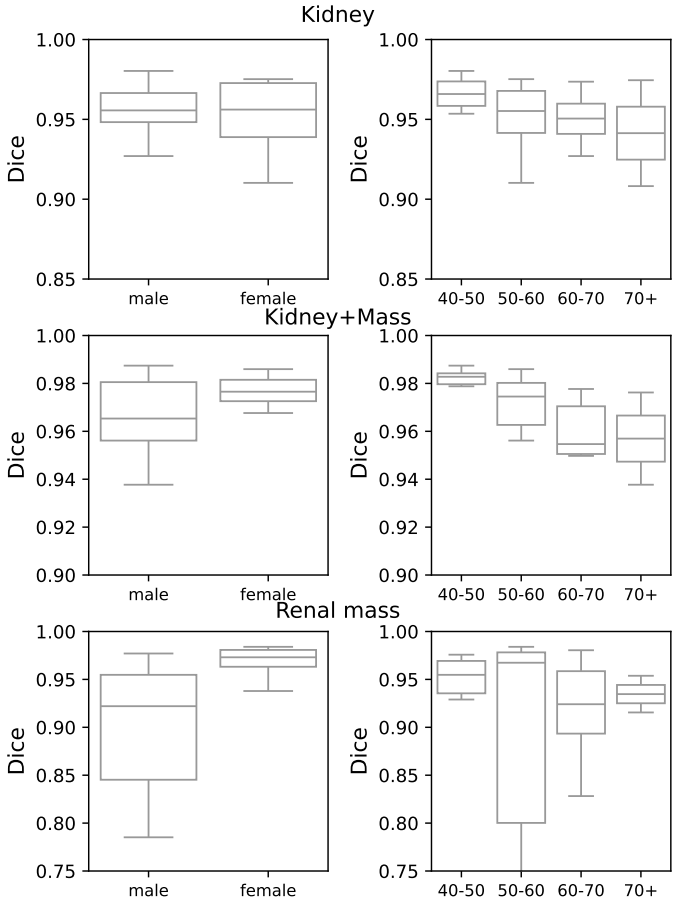

Figure 5: Subgroup analysis performed on the TCGA-KIRC test set. Subgroups that are tested for are patient sex and patient age. We show performance measured in Dice. The boxplots show the median (horizontal line in the box), quartiles (presented by the box) and the full distribution (presented by the whiskers). For the purpose of readability we omitted the outliers from these plots.

4.4 Qualitative results

In this section, we present qualitative results of the segmentations of the proposed and reference models. Figure 7 shows two cases from the Radboudumc test set where both the proposed and BAMF-model (baseline) are able to de-

Table 3: Detection performance measured in precision, recall and F1-score, where an overlap threshold of 0.5 is used to determine if a lesion is detected. We present the averages and standard deviations for all models and all three test datasets. PP = post-processing, best values per metric per dataset are presented in bold.

Radboudumc	Precision	Recall	F1-score
TotalSegmentator	0.00 (± 0.00)	0.00 (± 0.00)	0.00 (± 0.00)
BAMF-model	0.34 (± 0.28)	0.44 (± 0.35)	0.34 (± 0.24)
Renal-Net (ours)	0.56 (± 0.35)	0.52 (± 0.34)	0.51 (± 0.30)
Renal-Net + PP (ours)	**0.65 (± 0.30)**	**0.52 (± 0.34)**	**0.59 (± 0.27)**
TCGA-KIRC	Precision	Recall	F1-score
TotalSegmentator	0.00 (± 0.00)	0.00 (± 0.00)	0.00 (± 0.00)
Renal-Net (ours)	0.75 (± 0.28)	0.56 (± 0.33)	0.64 (± 0.27)
Renal-Net + PP (ours)	**0.75 (± 0.28)**	**0.56 (± 0.33)**	**0.64 (± 0.27)**
Charité Universitäts-medizin Berlin	Precision	Recall	F1-score
BAMF-model	0.68 (± 0.34)	0.79 (± 0.34)	0.70 (± 0.32)
Renal-Net (ours)	0.76 (± 0.32)	0.83 (± 0.30)	0.77 (± 0.29)
Renal-Net + PP (ours)	**0.77 (± 0.31)**	**0.83 (± 0.31)**	**0.78 (± 0.29)**

tect the renal mass (reference mask in yellow, predictions in orange), while TotalSegmentator misses the lesions in the first case. The biggest lesion in the top row of Figure 7 presents as a homogeneously fluid-filled, non-enhancing mass consistent with a simple or low-complexity cyst. The segmentation of the renal mass as predicted by the proposed model more closely mimics the annotation mask compared to the segmentation predicted by the BAMF-model (baseline). The kidneys (blue) are similar between all models.

Figure 8 presents two cases from the Charité Universitätsmedizin Berlin test set. The papillary renal cell carcinoma in the first row (yellow ground truth) has grown around the kidney and the characteristic kidney shape is not detectable anymore. The second row shows a tumor case on a remnant kidney. The proposed model did segment the tumors (orange) in both cases and did correctly capture the remnant kidney (blue). The BAMF-model did segment parts of the remnant kidney but could not correctly delineate the uncommon tumor shapes.

In Appendix E, we present additional interesting cases on the Charité Universitätsmedizin Berlin test set. In Figure 10, both models were able to correctly delineate the kidneys (blue) but failed to segment the clear cell renal cell carcinomas (yellow ground truth). In Figure 11, both models correctly segment the kidneys (blue) and are able to localize the tumors (orange). However, the BAMF-model (baseline) was better in capturing the whole tumor. The scan in the upper row has visible horizontal lines through the scan, possibly caused by metal. In Figure 12, we present a case where both models were able to correctly segment a transplanted kidney (blue) despite its uncommon orientation and location in the pelvic area.

5. Discussion

In this research, we validated and investigated the robustness of a segmentation model (named Renal-Net) based on the state-of-the-art nnU-Net framework and trained on two publicly available training datasets. We tested the segmentation and detection performance on three test datasets. Two test datasets are external for the proposed model. Except for the healthy kidney subset, our model significantly outperforms the reference models. Additionally, on the Radboudumc private test set, the model's performance was within the range of human inter-observer variability, indicating that it performs comparable to an independent human annotator. Furthermore, these results suggest the model's predictions are realistic and clinically relevant. Our subgroup analysis showed that differences in performance between subgroups in sex, age, and tumor histologic subtypes are small and that the model is robust to use of different contrast phases.

For the healthy kidney subset, TotalSegmentator achieves a lower median HD95 than the proposed model. We noticed that the proposed model predicted in a few cases structures physically far from the kidneys, which resulted in higher HD95 scores. In deployment, these structures could be flagged after connected component analysis or a region of interest crop could be made before or after the segmentation model using atlas based registration techniques (Buddenkotte et al., 2024). While such shifts have less impact on volumetric Dice scores, they can noticeably affect HD95.

The detection performance of our proposed model is superior to the reference models, although it still generates some false positives and false negatives, as is reflected in the

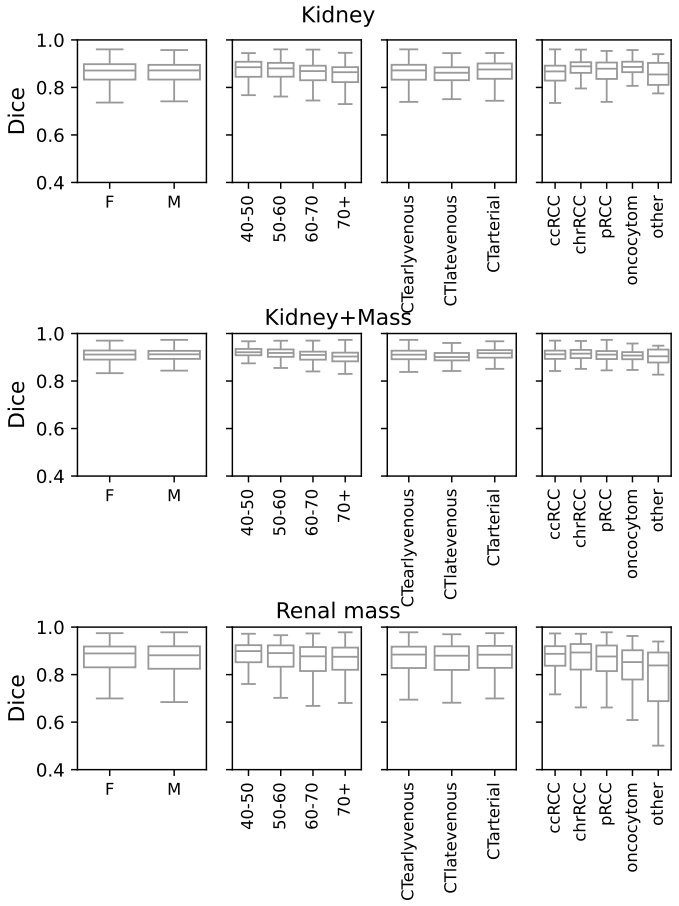

Figure 6: Subgroup analysis performed on the Charité Universitätsmedizin Berlin test set. Subgroups that are tested for are patient sex, patient age, CT contrast phase and tumor histologic subtype. We show performance measured in Dice. The boxplots show the median (horizontal line in the box), quartiles (presented by the box) and the full distribution (presented by the whiskers). For the purpose of readability we omitted the outliers from these plots.

precision, recall and F1-score. Future work should focus on reducing these detection errors to better match the clinical requirements. Clinically, this segmentation model could be used in multiple scenarios. One possible scenario is using the model as a standalone tool to assist radiologists directly. In this setting, the radiologist reviews the AI-generated segmentation and subsequently decides whether to accept or reject its predictions. False-positive detections could then be easily identified and dismissed by the radiologist. However, false-negative detections present a higher risk in this scenario, as they might incorrectly reassure the radiologist that no lesions are present in the kidneys, potentially leading to missed diagnoses. Another scenario of using this algorithm in clinical practice is to integrate it in a broader automated kidney diagnostic pipeline. The radiologist might not see the in-between results of the pipeline but only the outcome. In this case, both false negatives and false positives can negatively impact the outcomes of such workflows.

The qualitative results showed that the kidney segmentation performance is good for both the proposed and reference models. However, renal mass segmentation was for some cases better captured by the proposed model. The proposed model also showed more robust performance across a wide range of kidney anatomies.

In this study, we deliberately chose not to re-annotate either of the training datasets to standardize the inclusion or exclusion of the kidney hilum. We show preliminary evidence to support the hypothesis that exposure to annotation variability during training improves model robustness in real-world settings. By combining these heterogeneous datasets, our model is exposed to a broad range of anatomical variations, scanning protocols, and slice thicknesses, promoting robustness against distribution shifts. Clinically, achieving a high level of robustness is ultimately more valuable than minor incremental gains in segmentation accuracy. A potential application of our model in clinical practice includes evaluating kidney volumes and tracking relative volume changes over multiple time points. This can serve as a diagnostic marker to guide procedures such as active surveillance. Another interesting use case could be predicting the Mayo Imaging Classification score (Bais et al., 2024) for patients with ADPKD.

Furthermore, we trained the model to detect and segment renal masses as a single merged class, rather than separating cystic and solid lesions. This decision was motivated by the considerable overlap in imaging characteristics between these subtypes, e.g., Bosniak IV cystic lesions may present with both septations and solid components (Silverman et al., 2019), making a strict class boundary ambiguous. Our results demonstrate that this approach is effective for renal mass detection and segmentation. Whether two separate models, one for cystic and one for solid lesions, could further improve performance remains an open question and represents a direction for future work.

Besides CT imaging, MR imaging is a widely used modality for renal imaging. The model presented in this work is trained and validated on CT imaging only. Recent work has shown that with some extra preprocessing steps, a CT-only model can be applied to MR imaging (Häntze et al., 2025). Although the results are not yet satisfactory, this setup could be used to accelerate annotating MRI for training an MRI specific model.

The proposed model has an important limitation, the model is not able to recognize kidneys and renal masses on CT scans without contrast. This problem could potentially be mitigated by retraining the model on a training dataset that does include non-contrast scans. While contrast phase information was available for one test dataset, contrast agent specifications were not recorded across sites. We

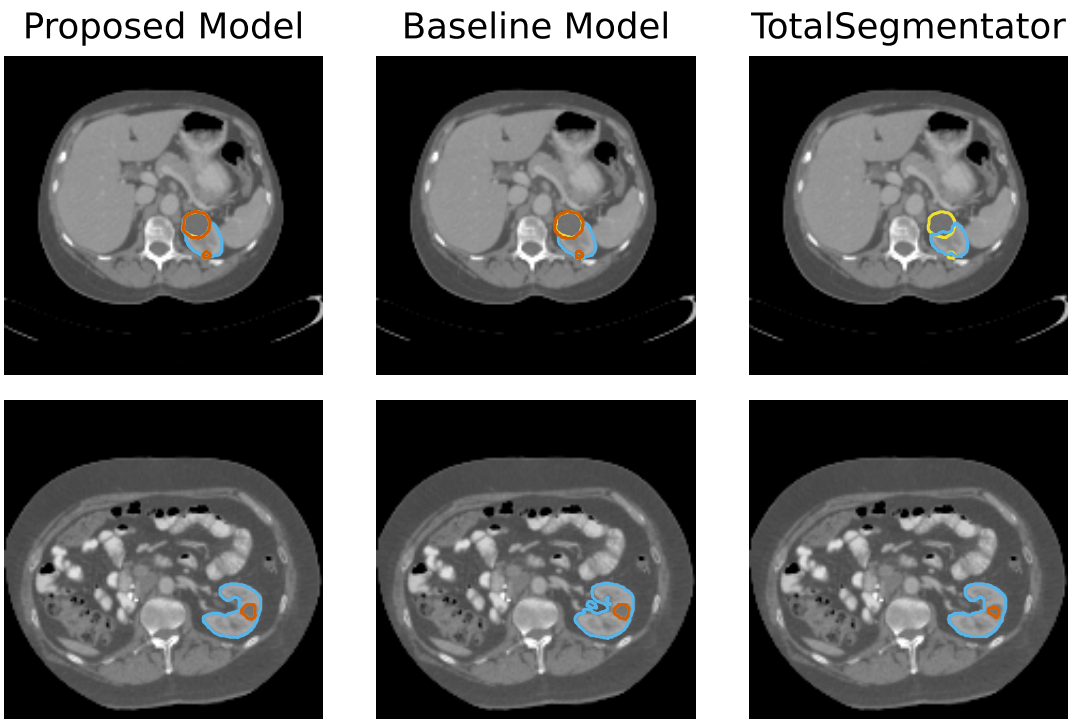

Figure 7: Presented are two cases (top and bottom) from the Radboudumc test set and the predictions of the proposed model, the BAMF-model (baseline) and TotalSegmentator. The reference renal mass mask and the renal mass predictions are shown. The predicted kidney is also highlighted. The biggest lesion in the top row presents as a homogeneously fluid-filled, non-enhancing mass consistent with a simple or low-complexity cyst.

acknowledge that a systematic analysis of contrast agent variability would have provided further insight into model generalizability. Nevertheless, the consistently high performance observed across all test datasets indirectly demonstrates that the model is robust to variability in scanner type, acquisition protocol, and other site-specific factors. A further limitation is the incomplete metadata availability across test sets, which limited a fully consistent subgroup analysis. Analyses were conducted for two of the three test sets where demographic metadata was available. Future work with complete metadata across all test sets would enable a more comprehensive assessment of model fairness across demographic subgroups. Finally, we acknowledge that our validation does not include pediatric cases nor data from Africa, South-America, Asia or Australia. Future validation studies should be performed to further test the generalizability of the model.

6. Conclusion

In this study, we validated our proposed renal mass segmentation model (Renal-Net) across three independent test datasets, demonstrating superior segmentation and detection performance compared to existing publicly available models. Through subgroup analyses, we confirmed that the model performs robustly across various patient subgroups and imaging conditions. Our results indicate that combining publicly available datasets with the state-of-the-art nnU-Net framework is sufficient for developing a robust and generalizable segmentation model for renal masses. We encourage researchers and clinicians to utilize our publicly available model and further evaluate its performance using imaging data from their own medical centers.

Acknowledgments

This research is funded by the European Union under HORIZON-HLTH-2022: COMFORT (101079894). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them.

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or

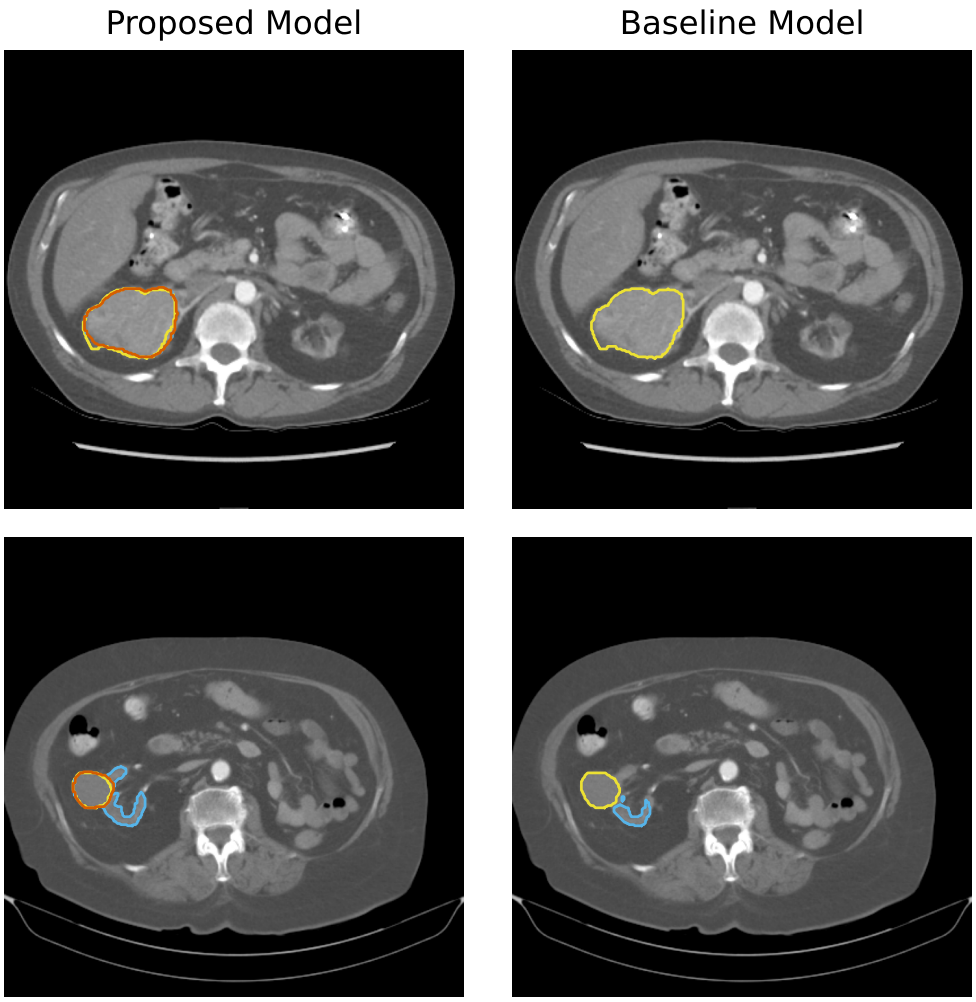

Figure 8: Presented are two cases (top and bottom) from the Charité Universitätsmedizin Berlin test set and the predictions of the proposed model and the BAMF-model (baseline). The reference renal mass mask and the renal mass predictions are shown. The predicted kidney is also highlighted.

human subjects.

Conflicts of Interest

Bram van Ginneken is a shareholder of Thirona BV and Plain Medical. Kiran Vaidhya Venkadesh is a shareholder of Plain Medical. The other authors declare that they do not have any conflicts of interest.

Data availability

The data used for training the model are publicly available via `https://kits-challenge.org/kits23/` and `https://zenodo.org/records/8014290`. The TCGA test set is publicly available via `https://www.cancerimagingarchive.net/collection/tcga-kirc/` and `https://zenodo.org/records/13244892`. The test datasets from Radboudumc and Charité Universitätsmedizin Berlin were used under approval for the current study. Restrictions apply to the availability of these datasets and so they are not publicly available.

References

Oguz Akin, Pierre Elnajjar, Matthew Heller, Rose Jarosz, Bradley J. Erickson, Shanah Kirk, Yueh Lee, Marston W. Linehan, Rabindra Gautam, Raghu Vikram, Kimberly M. Garcia, Charles Roche, Ermelinda Bonaccio, and Joe Filippini. The Cancer Genome Atlas Kidney Renal Clear Cell Carcinoma Collection (TCGA-KIRC), 2016.

Thomas Bais, Paul Geertsema, Martine G.E. Knol, Maatje D.A. van Gastel, Robbert J. de Haas, Esther Meijer, and Ron T. Gansevoort. Validation of the Mayo Imaging Classification System for Predicting Kidney Outcomes in ADPKD. Clinical Journal of the American Society of Nephrology : CJASN, 19(5):591–601, May 2024. .

Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, Fabian Lohöfer, Julian Walter Holch, Wieland Sommer, Felix Hofmann, Alexandre Hostettler, Naama Lev-Cohain, Michal Drozdzal, Michal Marianne Amitai, Refael Vivanti, Jacob Sosna, Ivan Ezhov, Anjany Sekuboyina, Fernando Navarro, Florian Kofler, Johannes C. Paetzold, Suprosanna Shit, Xiaobin Hu, Jana Lipková, Markus Rempfler, Marie Piraud, Jan Kirschke, Benedikt Wiestler, Zhiheng Zhang, Christian Hülsemeyer, Marcel Beetz, Florian Ettlinger, Michela Antonelli, Woong Bae, Míriam Bellver, Lei Bi, Hao Chen, Grzegorz Chlebus, Erik B. Dam, Qi Dou, Chi-Wing Fu, Bogdan Georgescu, Xavier Giró-i-Nieto, Felix Gruen, Xu Han, Pheng-Ann Heng, Jürgen Hesser, Jan Hendrik Moltz, Christian Igel, Fabian Isensee, Paul Jäger, Fucang Jia, Krishna Chaitanya Kaluva, Mahendra Khened, Ildoo Kim, Jae-Hun Kim, Sungwoong Kim, Simon Kohl, Tomasz Konopczynski, Avinash Kori, Ganapathy Krishnamurthi, Fan Li, Hongchao Li, Junbo Li, Xiaomeng Li, John Lowengrub, Jun Ma, Klaus Maier-Hein, Kevis-Kokitsi Maninis, Hans Meine, Dorit Merhof, Akshay Pai, Mathias Perslev, Jens Petersen, Jordi Pont-Tuset, Jin Qi, Xiaojuan Qi, Oliver Rippel, Karsten Roth, Ignacio Sarasua, Andrea Schenk, Zengming Shen, Jordi Torres, Christian Wachinger, Chunliang Wang, Leon Weninger, Jianrong Wu, Daguang Xu, Xiaoping Yang, Simon Chun-Ho Yu, Yading Yuan, Miao Yue, Liping Zhang, Jorge Cardoso, Spyridon Bakas, Rickmer Braren, Volker Heinemann, Christopher Pal, An Tang, Samuel Kadoury, Luc Soler, Bram van Ginneken, Hayit Greenspan, Leo Joskowicz, and Bjoern Menze. The Liver Tumor Segmentation Benchmark (LiTS). Medical image analysis, 84: 102680, 2023.

Thomas Buddenkotte, Roland Opfer, Julia Krüger, Alessa Hering, and Mireia Crispin-Ortuzar. CTARR: A fast and robust method for identifying anatomical regions on CT images via atlas registration. Machine Learning for Biomedical Imaging, 2:2067–2088, 2024. .

Umberto Capitanio and Francesco Montorsi. Renal cancer. The Lancet, 387(10021):894–906, 2016.

Mina Chookhachizadeh Moghadam, Mohit Aspal, Xinzi He, Dominick J Romano, Arman Sharbatdaran, Zhongxiu Hu, Kurt Teichman, Hui Yi Ng He, Usama Sattar, Chenglin Zhu, Hreedi Dev, Daniil Shimonov, James M Chevalier, Akshay Goel, George Shih, Jon D Blumenfeld, Mert R Sabuncu, and Martin R Prince. Deep learning-based liver cyst segmentation in MRI for autosomal dominant polycystic kidney disease. Radiology Advances, 1(2), 2024.

Ioanna Chouvarda, Sara Colantonio, Ana S. C. Verde, Ana Jimenez-Pastor, Leonor Cerdá-Alberich, Yannick Metz, Lithin Zacharias, Shereen Nabhani-Gebara, Maciej Bobowicz, Gianna Tsakou, Karim Lekadir, Manolis Tsiknakis, Luis Martí-Bonmati, and Nikolaos Papanikolaou. Differences in technical and clinical perspectives on AI validation in cancer imaging: Mind the gap! European Radiology Experimental, 9:7, 2025. .

Luigi Cirillo, Samantha Innocenti, and Francesca Becherucci. Global epidemiology of kidney cancer. Nephrology Dialysis Transplantation, 39(6):920–928, 2024.

M. J. J. de Grauw, E. Th. Scholten, E. J. Smit, M. J. C. M. Rutten, M. Prokop, B. van Ginneken, and A. Hering. The ULS23 challenge: A baseline model and benchmark dataset for 3D universal lesion segmentation in computed tomography. Medical Image Analysis, 102:103525, 2025.

E. A. Eisenhauer, P. Therasse, J. Bogaerts, L. H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, and J. Verweij. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). European Journal of Cancer, 45(2):228–247, 2009.

Seyed Alireza Emamian, Michael Bachmann Nielsen, Jan Fog Pedersen, and Lars Ytte. Kidney dimensions at sonography: correlation with age, sex, and habitus in 665 adult volunteers. AJR. American journal of roentgenology, 160(1):83–86, 1993.

D. Filippiadis, G. Mauri, P. Marra, G. Charalampopoulos, N. Gennaro, and F. De Cobelli. Percutaneous ablation techniques for renal cell carcinoma: Current status and future trends. International Journal of Hyperthermia, 36 (2):21–30, 2019.

Pritika Gaur, Wladyslaw Gedroyc, and Peter Hill. ADPKD—what the radiologist should know. The British Journal of Radiology, 92(1098):20190078, 2019.

Yasmeen George. A Coarse-to-Fine 3D U-Net Network for Semantic Segmentation of Kidney CT Scans. In Nicholas Heller, Fabian Isensee, Darya Trofimova, Resha Tejpaul, Nikolaos Papanikolopoulos, and Christopher Weight, editors, Kidney and Kidney Tumor Segmentation, volume 13168, pages 137–142. Springer International Publishing, Cham, 2022. .

Adriana V. Gregory, Deema A. Anaam, Andrew J. Vercnocke, Marie E. Edwards, Vicente E. Torres, Peter C. Harris, Bradley J. Erickson, and Timothy L. Kline. Semantic Instance Segmentation of Kidney Cysts in MR Images: A Fully Automated 3D Approach Developed

Through Active Learning. Journal of Digital Imaging, 34 (4):773–787, 2021.

Marzieh Haghighi, Simon K. Warfield, and Sila Kurugol. Automatic renal segmentation in DCE-MRI using convolutional neural networks. In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pages 1534–1537, 2018.

Hartmut Häntze, Lina Xu, Maximilian Nikolas Rattunde, Leonhard Donle, Felix J. Dorfner, Alessa Hering, Jawed Nawabi, Lisa C. Adams, and Keno K. Bressem. MRI annotation using an inversion-based preprocessing for CT model adaptation. European Radiology Experimental, 9 (1):93, 2025.

Nicholas Heller, Fabian Isensee, Klaus H. Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, Guang Yao, Yaozong Gao, Yao Zhang, Yixin Wang, Feng Hou, Jiawei Yang, Guangwei Xiong, Jiang Tian, Cheng Zhong, Jun Ma, Jack Rickman, Joshua Dean, Bethany Stai, Resha Tejpaul, Makinna Oestreich, Paul Blake, Heather Kaluzniak, Shaneabbas Raza, Joel Rosenberg, Keenan Moore, Edward Walczak, Zachary Rengel, Zach Edgerton, Ranveer Vasdev, Matthew Peterson, Sean McSweeney, Sarah Peterson, Arveen Kalapara, Niranjan Sathianathen, Nikolaos Papanikolopoulos, and Christopher Weight. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge. Medical Image Analysis, 67:101821, January 2021.

Nicholas Heller, Fabian Isensee, Dasha Trofimova, Resha Tejpaul, Zhongchen Zhao, Huai Chen, Lisheng Wang, Alex Golts, Daniel Khapun, Daniel Shats, Yoel Shoshan, Flora Gilboa-Solomon, Yasmeen George, Xi Yang, Jianpeng Zhang, Jing Zhang, Yong Xia, Mengran Wu, Zhiyang Liu, Ed Walczak, Sean McSweeney, Ranveer Vasdev, Chris Hornung, Rafat Solaiman, Jamee Schoephoerster, Bailey Abernathy, David Wu, Safa Abdulkadir, Ben Byun, Justice Spriggs, Griffin Struyk, Alexandra Austin, Ben Simpson, Michael Hagstrom, Sierra Virnig, John French, Nitin Venkatesh, Sarah Chan, Keenan Moore, Anna Jacobsen, Susan Austin, Mark Austin, Subodh Regmi, Nikolaos Papanikolopoulos, and Christopher Weight. The KiTS21 Challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase CT. arXiv, (arXiv:2307.01984), 2023.

Gabriel E. Humpire-Mamani, Luc Builtjes, Colin Jacobs, Bram Van Ginneken, Mathias Prokop, and Ernst Th. Scholten. Dataset for: Kidney abnormality segmentation in thorax-abdomen CT scans. Zenodo, June 2023.

Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. Nature Methods, 18(2):203–211, 2021.

Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F. Jaeger. nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation. arXiv, (arXiv:2404.09556), April 2024.

Leo Joskowicz, D. Cohen, N. Caplan, and J. Sosna. Inter-observer variability of manual contour delineation of structures in CT. European Radiology, 29(3):1391–1399, 2019.

Stella K. Kang, William C. Huang, Pari V. Pandharipande, and Hersh Chandarana. Solid Renal Masses: What the Numbers Tell Us. American Journal of Roentgenology, 202(6):1196–1206, 2014. .

Rashid Khan, Chao Chen, Asim Zaman, Jiayi Wu, Haixing Mai, Liyilei Su, Yan Kang, and Bingding Huang. RenalSegNet: Automated segmentation of renal tumor, veins, and arteries in contrast-enhanced CT scans. Complex & Intelligent Systems, 11(2):131, 2025.

Panagiotis Korfiatis, Aleksandar Denic, Marie E. Edwards, Adriana V. Gregory, Darryl E. Wright, Aidan Mullan, Joshua Augustine, Andrew D. Rule, and Timothy L. Kline. Automated Segmentation of Kidney Cortex and Medulla in CT Images: A Multisite Evaluation Study. Journal of the American Society of Nephrology, 33(2):420, 2022.

Alexander Kutikov and Robert G. Uzzo. The R.E.N.A.L. Nephrometry Score: A Comprehensive Standardized System for Quantitating Renal Tumor Size, Location and Depth. The Journal of Urology, 182(3):844–853, 2009.

Alexander Kutikov, Brian L. Egleston, Yu-Ning Wong, and Robert G. Uzzo. Evaluating Overall Survival and Competing Risks of Death in Patients With Localized Renal Cell Carcinoma Using a Comprehensive Nomogram. Journal of Clinical Oncology, 28(2):311–317, 2010.

Julian Leube, Matthias Horn, Philipp E. Hartrampf, Andreas K. Buck, Michael Lassmann, and Johannes Tran-Gia. PSMA-PET improves deep learning-based automated CT kidney segmentation. Zeitschrift Fur Medizinische Physik, 34(2):231–241, May 2024. .

Geert Litjens, Robert Toth, Wendy van de Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, Robin Strand, Filip Malmberg, Yangming Ou, Christos

Davatzikos, Matthias Kirschner, Florian Jung, Jing Yuan, Wu Qiu, Qinquan Gao, Philip "Eddie" Edwards, Bianca Maan, Ferdinand van der Heijden, Soumya Ghose, Jhimli Mitra, Jason Dowling, Dean Barratt, Henkjan Huisman, and Anant Madabhushi. Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge. Medical Image Analysis, 18(2):359–373, 2014.

Börje Ljungberg, Laurance Albiges, Yasmin Abu-Ghanem, Karim Bensalah, Saeed Dabestani, Sergio Fernández-Pello, Rachel H. Giles, Fabian Hofmann, Milan Hora, Markus A. Kuczyk, Teele Kuusk, Thomas B. Lam, Lorenzo Marconi, Axel S. Merseburger, Thomas Powles, Michael Staehler, Rana Tahbaz, Alessandro Volpe, and Axel Bex. European Association of Urology Guidelines on Renal Cell Carcinoma: The 2019 Update. European Urology, 75(5):799–810, 2019.

Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D. Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, Mauricio Reyes, Michael A. Riegler, Manuel Wiesenfarth, A. Emre Kavur, Carole H. Sudre, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzel, Tim Rädsch, Laura Acion, Michela Antonelli, Tal Arbel, Spyridon Bakas, Arriel Benis, Matthew Blaschko, M. Jorge Cardoso, Veronika Cheplygina, Beth A. Cimini, Gary S. Collins, Keyvan Farahani, Luciana Ferrer, Adrian Galdran, Bram van Ginneken, Robert Haase, Daniel A. Hashimoto, Michael M. Hoffman, Merel Huisman, Pierre Jannin, Charles E. Kahn, Dagmar Kainmueller, Bernhard Kainz, Alexandros Karargyris, Alan Karthikesalingam, Hannes Kenngott, Florian Kofler, Annette Kopp-Schneider, Anna Kreshuk, Tahsin Kurc, Bennett A. Landman, Geert Litjens, Amin Madani, Klaus Maier-Hein, Anne L. Martel, Peter Mattson, Erik Meijering, Bjoern Menze, Karel G. M. Moons, Henning Müller, Brennan Nichyporuk, Felix Nickel, Jens Petersen, Nasir Rajpoot, Nicola Rieke, Julio Saez-Rodriguez, Clara I. Sánchez, Shravya Shetty, Maarten van Smeden, Ronald M. Summers, Abdel A. Taha, Aleksei Tiulpin, Sotirios A. Tsaftaris, Ben Van Calster, Gaël Varoquaux, and Paul F. Jäger. Metrics reloaded: Recommendations for image analysis validation. Nature Methods, 21(2):195–212, 2024.

Gabriel Efrain Humpire Mamani, Nikolas Lessmann, Ernst Th Scholten, Mathias Prokop, Colin Jacobs, and Bram van Ginneken. Kidney abnormality segmentation in thorax-abdomen CT scans. arXiv, (arXiv:2309.03383), 2023.

Sarah F. Mercaldo, James M. Hillis, and Jeffrey D. Blume. Evaluating the Performance and Clinical Utility of AI-driven Diagnostic Tools in Radiology. Radiology, 317(2): e243935, 2025. .

J. Patrick Mershon, Mei N. Tuong, and Noah S. Schenkman. Thermal ablation of the small renal mass: A critical analysis of current literature. Minerva Urologica e Nefrologica, 72(2), 2020.

Charles R. Meyer, Timothy D. Johnson, Geoffrey McLennan, Denise R. Aberle, Ella A. Kazerooni, Heber MacMahon, Brian F. Mullan, David F. Yankelevitz, Edwin J.R. Van Beek, Samuel G. Armato, Michael F. McNitt-Gray, Anthony P. Reeves, David Gur, Claudia I. Henschke, Eric A. Hoffman, Peyton H. Bland, Gary Laderach, Richie Pais, David Qing, Chris Piker, Junfeng Guo, Adam Starkey, Daniel Max, Barbara Y. Croft, and Laurence P. Clarke. Evaluation of Lung MDCT Nodule Annotation Across Radiologists and Methods. Academic Radiology, 13(10):1254–1265, 2006.

Consortium MONAI. MONAI: Medical Open Network for AI. Zenodo, 2024.

Robert J. Motzer, Eric Jonasch, Neeraj Agarwal, Ajjai Alva, Michael Baine, Kathryn Beckermann, Maria I. Carlo, Toni K. Choueiri, Brian A. Costello, Ithaar H. Derweesh, Arpita Desai, Yasser Ged, Saby George, John L. Gore, Naomi Haas, Steven L. Hancock, Payal Kapur, Christos Kyriakopoulos, Elaine T. Lam, Primo N. Lara, Clayton Lau, Bryan Lewis, David C. Madoff, Brandon Manley, M. Dror Michaelson, Amir Mortazavi, Lakshminarayanan Nandagopal, Elizabeth R. Plimack, Lee Ponsky, Sundhar Ramalingam, Brian Shuch, Zachary L. Smith, Jeffrey Sosman, Mary A. Dwyer, Lisa A. Gurski, and Angela Motter. Kidney Cancer, Version 3.2022. Journal of the National Comprehensive Cancer Network, 20(1):71–90, 2022.

Gowtham Krishnan Murugesan, Diana McCrumb, Mariam Aboian, Tej Verma, Rahul Soni, Fatima Memon, Keyvan Farahani, Linmin Pei, Ulrike Wagner, Andrey Y. Fedorov, David Clunie, Stephen Moore, and Jeff Van Oss. AI-Generated Annotations Dataset for Diverse Cancer Radiology Collections in NCI Image Data Commons. Scientific Data, 11(1):1165, October 2024. ISSN 2052-4463. .

Andriy Myronenko, Dong Yang, Yufan He, and Daguang Xu. Automated 3D Segmentation of Kidneys and Tumors in MICCAI KiTS 2023 Challenge. In Nicholas Heller, Andrew Wood, Fabian Isensee, Tim Rädsch, Resha Teipaul, Nikolaos Papanikolopoulos, and Christopher Weight, editors, Kidney and Kidney Tumor Segmentation, pages 1–7, Cham, 2024. Springer Nature Switzerland.

Shuchao Pang, Anan Du, Mehmet A. Orgun, Zhenmei Yu, Yunyun Wang, Yan Wang, and Guanfeng Liu. CTumorGAN: A unified framework for automatic computed tomography tumor segmentation. European Journal of

Nuclear Medicine and Molecular Imaging, 47(10):2248–2268, 2020.

Anish Raj, Fabian Tollens, Laura Hansen, Alena-Kathrin Golla, Lothar R. Schad, Dominik Nörenberg, and Frank G. Zöllner. Deep Learning-Based Total Kidney Volume Segmentation in Autosomal Dominant Polycystic Kidney Disease Using Attention, Cosine Loss, and Sharpness Aware Minimization. Diagnostics, 12(5), May 2022. ISSN 2075-4418. .

Alex G. Raman, David Fisher, Joseph M. Rich, Christopher Weight, Nicholas Heller, Mihir Desai, Inderbir Gill, Assad Oberai, and Vinay A. Duddalwar. Evaluation of nnU-Net for kidney tumor segmentation on a large external patient cohort. European Journal of Radiology Artificial Intelligence, 3:100035, 2025.

Maria Rombolotti, Fabio Sangalli, Domenico Cerullo, Andrea Remuzzi, and Ettore Lanzarone. Automatic cyst and kidney segmentation in autosomal dominant polycystic kidney disease: Comparison of U-Net based methods. Computers in Biology and Medicine, 146:105431, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv, (arXiv:1505.04597), 2015.

Stuart G. Silverman, Ivan Pedrosa, James H. Ellis, Nicole M. Hindman, Nicola Schieda, Andrew D. Smith, Erick M. Remer, Atul B. Shinagare, Nicole E. Curci, Steven S. Raman, Shane A. Wells, Samuel D. Kaffenberger, Zhen J. Wang, Hersh Chandarana, and Matthew S. Davenport. Bosniak Classification of Cystic Renal Masses, Version 2019: An Update Proposal and Needs Assessment. Radiology, 292 (2):475–488, 2019.

J. Staal, M.D. Abramoff, M. Niemeijer, M.A. Viergever, and B. van Ginneken. Ridge-based vessel segmentation in color images of the retina. IEEE Transactions on Medical Imaging, 23(4):501–509, 2004.

Kwang-Hyun Uhm, Hyunjun Cho, Zhixin Xu, Seohoon Lim, Seung-Won Jung, Sung-Hoo Hong, and Sung-Jea Ko. Exploring 3D U-Net Training Configurations and Post-processing Strategies for the MICCAI 2023 Kidney and Tumor Segmentation Challenge. In Nicholas Heller, Andrew Wood, Fabian Isensee, Tim Rädsch, Resha Teipaul, Nikolaos Papanikolopoulos, and Christopher Weight, editors, Kidney and Kidney Tumor Segmentation, pages 8–13, Cham, 2024. Springer Nature Switzerland. .

Jakob Wasserthal, Hanns-Christian Breit, Manfred T. Meyer, Maurice Pradella, Daniel Hinck, Alexander W. Sauter, Tobias Heye, Daniel T. Boll, Joshy Cyriac, Shan Yang, Michael Bach, and Martin Segeroth. TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. Radiology: Artificial Intelligence, 5(5): e230024, 2023.

Ehwa Yang, Chan Kyo Kim, Yi Guan, Bang-Bon Koo, and Jae-Hun Kim. 3D multi-scale residual fully convolutional neural network for segmentation of extremely large-sized kidney tumor. Computer Methods and Programs in Biomedicine, 215:106616, 2022.

Frank G. Zöllner, Marek Kociński, Laura Hansen, Alena-Kathrin Golla, Amira Šerifović Trbalić, Arvid Lundervold, Andrzej Materka, and Peter Rogelj. Kidney Segmentation in Renal Magnetic Resonance Imaging - Current Status and Prospects. IEEE Access, 9:71577–71605, 2021. ISSN 2169-3536. .

Appendix A. nnU-Net configuration

Table 4 presents the final training and architecture setup of the proposed model.

Table 4: Overview of nnU-Net 3D full-resolution configuration used in this research. This configuration is determined by the nnU-Net framework and we did not adjust any settings.

Category	Configuration
Preprocessing	
Data identifier	nnUNetPlans_3d_fullres
Preprocessor	DefaultPreprocessor
Patch size	[160, 224, 192]
Median image size	[390, 512, 512]
Spacing (mm)	[0.7695, 0.7695, 0.7695]
Normalization	CTNormalization
Network Architecture	
Class	ResidualEncoderUNet
Number of stages	6
Features per stage	[32, 64, 128, 256, 320, 320]
Kernel sizes	$3 \times 3 \times 3$ (all stages)
Strides	[1,1,1], [2,2,2] ($\times$5)
Blocks per stage	[1, 3, 4, 6, 6, 6]
Decoder convs	[1, 1, 1, 1, 1]
Normalization	InstanceNorm3d
Nonlinearity	Leaky ReLU
Bias	True
Training Configuration	
Batch size	2
Batch Dice	True

Table 5: Segmentation results on the TCGA-KIRC test set. PP = post-processing, best values per metric per region are presented in bold. † Significantly superior to TotalSegmentator ($p < 0.05$).

Kidney region	Dice ↑	HD95 (mm) ↓
TotalSegmentator	0.80 ± 0.14	25.90 ± 32.41
Renal-Net (ours)	0.95 ± 0.03	3.79 ± 5.89
Renal-Net + PP (ours)	**0.95 ± 0.03**†	**3.79 ± 5.89**
Kidney+mass region	Dice ↑	HD95 (mm) ↓
TotalSegmentator	0.75 ± 0.18	28.50 ± 26.90
Renal-Net (ours)	0.97 ± 0.02	3.64 ± 7.85
Renal-Net + PP (ours)	**0.97 ± 0.02**†	**3.64 ± 7.85**
Renal mass region	Dice ↑	HD95 (mm) ↓
TotalSegmentator	0.05 ± 0.13	120.18 ± 61.30
Renal-Net (ours)	0.86 ± 0.25	14.94 ± 25.81
Renal-Net + PP (ours)	**0.86 ± 0.25**†	**14.94 ± 25.81**

Appendix B. Segmentation performance

In this section, we present the segmentation performance on the three test sets using the Dice similarity coefficient and Hausdorff distance at 95th percentile (Table 5, 6, 7). Significance tests are only performed for the main metric: Dice similarity coefficient. The statistical analysis plan follows a hierarchical scheme, where only if superiority of the proposed model over TotalSegmentator is established, supriority over the BAMF model is assessed.

Table 6: Segmentation results on the Charité Universitätsmedizin Berlin test set. PP = post-processing, best values per metric per region are presented in bold. ‡ Significantly superior to BAMF model ($p < 0.05$.)

Kidney region	Dice ↑	HD95 (mm) ↓
BAMF-model	0.83 ± 0.13	12.19 ± 18.90
Renal-Net (ours)	0.85 ± 0.10	9.84 ± 16.11
Renal-Net + PP (ours)	**0.85 ± 0.10**‡	**9.84 ± 16.11**
Kidney+mass region	Dice ↑	HD95 (mm) ↓
BAMF-model	0.89 ± 0.11	10.43 ± 13.68
Renal-Net (ours)	0.90 ± 0.07	9.28 ± 12.40
Renal-Net + PP (ours)	**0.90 ± 0.08**‡	**9.21 ± 11.92**
Renal mass region	Dice ↑	HD95 (mm) ↓
BAMF-model	0.81 ± 0.19	13.43 ± 18.47
Renal-Net (ours)	0.83 ± 0.17	10.75 ± 16.86
Renal-Net + PP (ours)	**0.83 ± 0.17**‡	**10.46 ± 15.83**

Appendix C. Detection results

Table 8 presents the detection performance for an overlap threshold of 0 based on the intersection of union.

Table 7: Segmentation results on the Radboudumc test set. B20 is a subset comprising of 20 patients with healthy kidneys. B30 is a subset comprising of 30 patients with kidneys containing lesions. PP = post-processing, best values per metric per region are presented in bold. † Significantly superior to TotalSegmentator ($p < 0.05$). ‡ Significantly superior to BAMF model ($p < 0.05$, tested conditional on superiority over TotalSegmentator).

Kidney region	B20		B30	
	Dice ↑	HD95 (mm) ↓	Dice ↑	HD95 (mm) ↓
TotalSegmentator	**0.95 ± 0.01**	**7.39 ± 23.71**	0.92 ± 0.10	**15.58 ± 37.87**
BAMF model	0.90 ± 0.02	57.83 ± 108.77	0.90 ± 0.04	18.44 ± 43.61
Renal-Net (ours)	0.93 ± 0.06	55.80 ± 101.21	0.94 ± 0.05	44.32 ± 99.98
Renal-Net + PP (ours)	0.93 ± 0.06	55.80 ± 101.21	**0.94 ± 0.05**†‡	44.32 ± 99.98
Kidney+mass region	B20		B30	
	Dice ↑	HD95 (mm) ↓	Dice ↑	HD95 (mm) ↓
TotalSegmentator	–	–	0.94 ± 0.02	**15.40 ± 37.83**
BAMF model	–	–	0.84 ± 0.15	70.58 ± 88.95
Renal-Net (ours)	–	–	0.94 ± 0.05	49.44 ± 102.72
Renal-Net + PP (ours)	–	–	**0.95 ± 0.02**†‡	43.66 ± 100.00
Renal Mass region	B20		B30	
	Dice ↑	HD95 (mm) ↓	Dice ↑	HD95 (mm) ↓
TotalSegmentator	–	–	0.06 ± 0.14	123.27 ± 113.94
BAMF model	–	–	0.45 ± 0.35	119.28 ± 113.74
Renal-Net (ours)	–	–	0.64 ± 0.31	50.41 ± 76.18
Renal-Net + PP (ours)	–	–	**0.66 ± 0.29**†‡	**31.84 ± 39.10**

Table 8: Detection performance with an overlap threshold of 0 based on the intersection over union. PP = post-processing, best values per metric per dataset are presented in bold.

Radboudumc test set	Precision	Recall	F1-score
TotalSegmentator	0.42 (± 0.40)	0.11 (± 0.26)	0.23 (± 0.21)
BAMF-model	0.58 (± 0.31)	**0.66 (± 0.28)**	0.57 (± 0.23)
Renal-Net (ours)	0.76 (± 0.32)	0.66 (± 0.29)	0.67 (± 0.26)
Renal-Net + PP (ours)	**0.83 (± 0.23)**	0.64 (± 0.32)	**0.74 (± 0.20)**
TCGA-KIRC	Precision	Recall	F1-score
TotalSegmentator	0.76 (± 0.34)	0.28 (± 0.35)	0.51 (± 0.24)
Renal-Net (ours)	0.91 (± 0.18)	0.69 (± 0.33)	0.79 (± 0.22)
Renal-Net + PP (ours)	**0.91 (± 0.18)**	**0.69 (± 0.33)**	**0.79 (± 0.22)**
Charité Universitäts-medizin Berlin	Precision	Recall	F1-score
BAMF-model	0.78 (± 0.27)	0.91 (± 0.23)	0.81 (± 0.23)
Renal-Net (ours)	0.83 (± 0.26)	**0.92 (± 0.22)**	0.84 (± 0.22)
Renal-Net + PP (ours)	**0.84 (± 0.25)**	0.91 (± 0.22)	**0.85 (± 0.22)**

Appendix D. Subgroup analysis

In this section, we present an additional subgroup analysis of the BAMF model on the Charité Universitätsmedizin Berlin test set. Figure 9 presents these results and it can be noted that similarly as the proposed model, the BAMF model does not seem to have a systematic bias towards a subgroup.

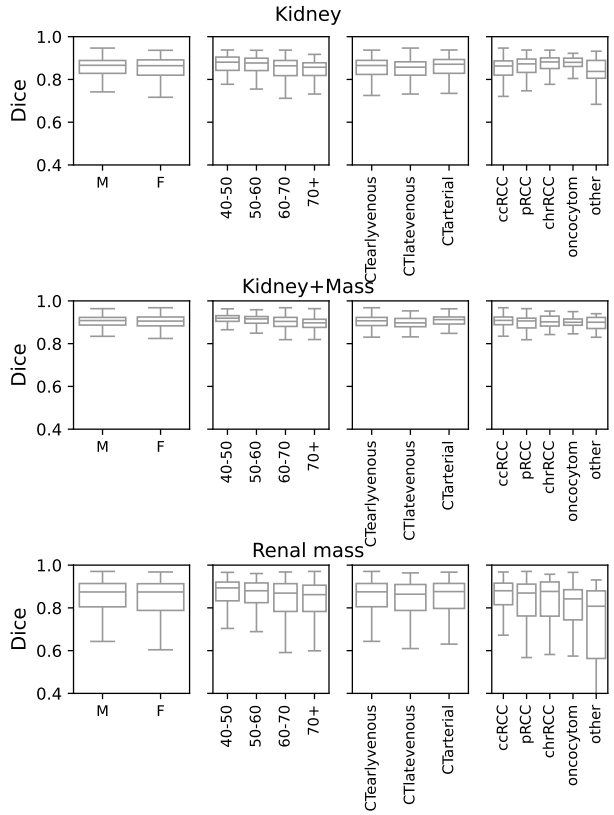

Figure 9: Subgroup analysis of the BAMF-model performed on the Charité Universitätsmedizin Berlin test set. Subgroups that are tested for are patient sex, patient age, CT contrast phase and tumor histologic subtype. We show performance measured in Dice. The boxplots show the median (horizontal line in the box), quartiles (presented by the box) and the full distribution (presented by the whiskers). For the purpose of readability we omitted the outliers from these plots.

Appendix E. Qualitative results

In this section, we present additional qualitative results for the Charité Universitätsmedizin Berlin test set. In Figure 10, both models were able to correctly delineate the kidneys (blue) but failed to segment the clear cell renal cell carcinomas (yellow ground truth). The tumor in the first row is clearly visible; it is likely the small form of the remnant kidney that confuses both models. The tumor in the second row has minimal contrast differences to the kidney, it is very hard to detect even if the reference annotation mask is known.

In Figure 11, both models correctly segment the kidneys (blue) and are able to localize the tumors (orange). However, the BAMF-model (baseline) was better in capturing the whole tumor. The scan in the upper row has visible horizontal lines through the scan, possibly caused by metal.

In Figure 12, we present a case where both models were able to correctly segment a transplanted kidney (blue) despite its uncommon orientation and location in the pelvic area.

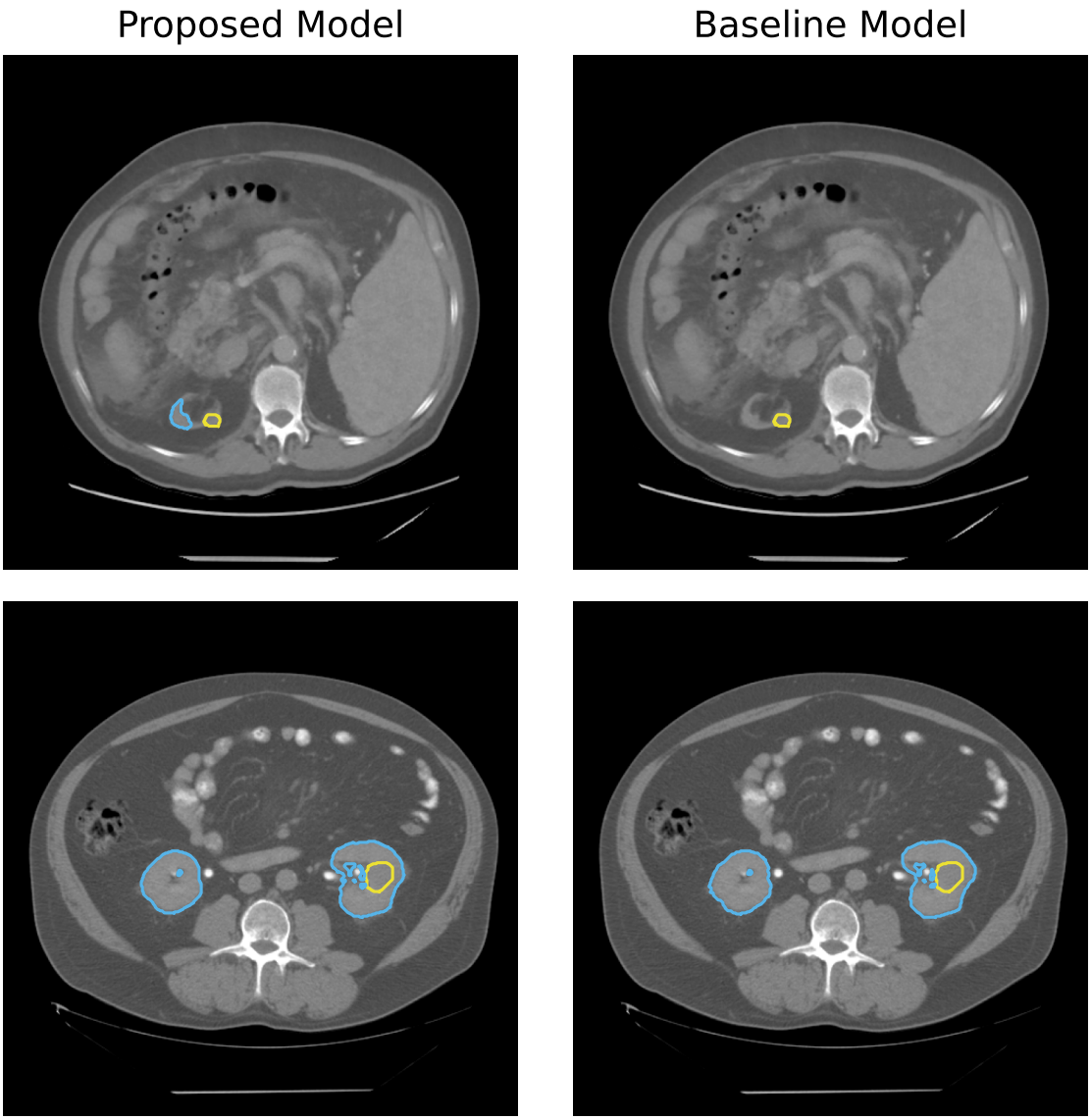

Figure 10: Presented are two cases (top and bottom) from the Charité Universitätsmedizin Berlin test set and the predictions of the proposed model and the BAMF-model (baseline). The reference renal mass mask ■ is shown, both models were not able to detect and segment these renal masses. The predicted kidneys ■ are also highlighted.

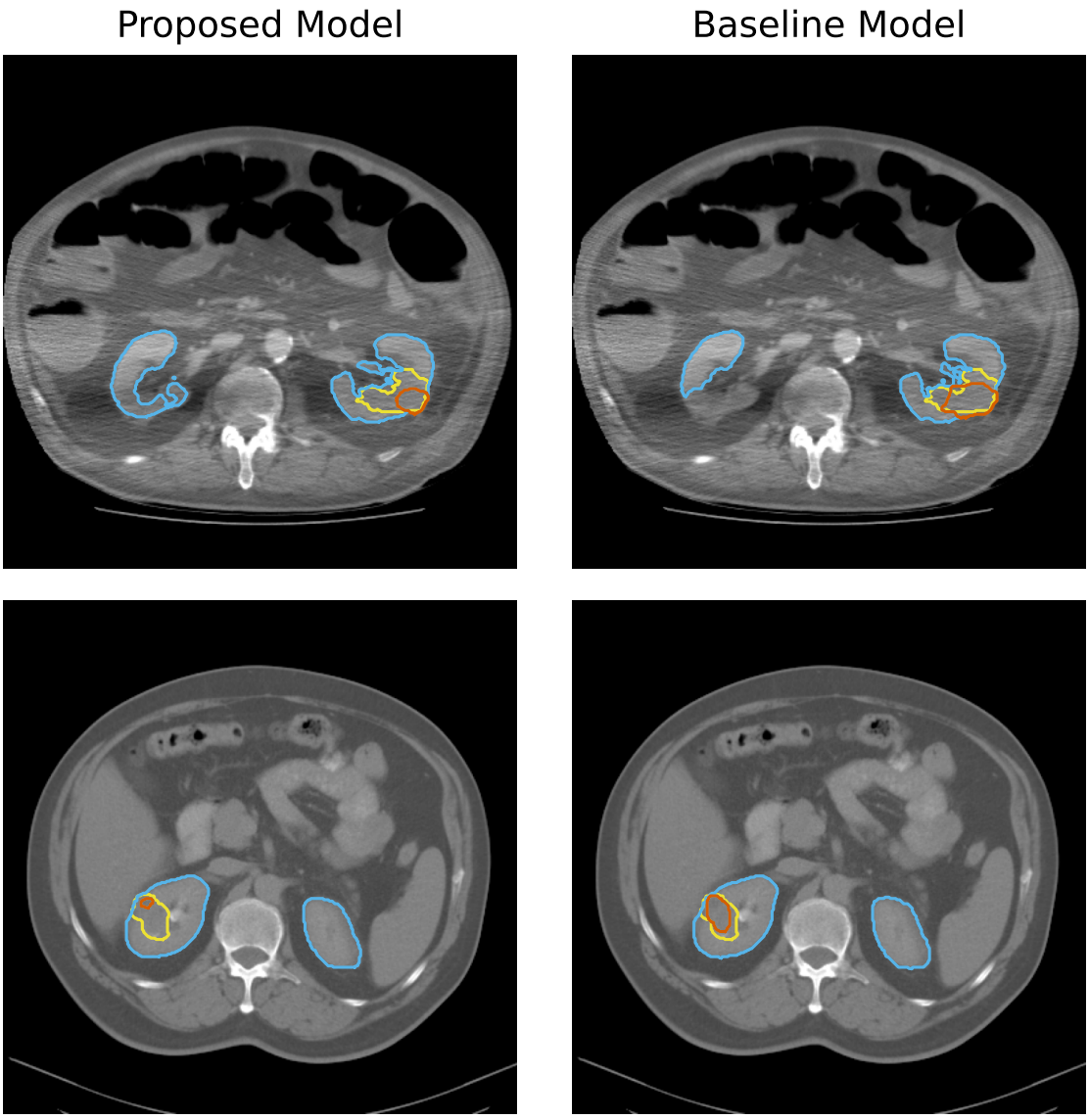

Figure 11: Presented are two cases (top and bottom) from the Charité Universitätsmedizin Berlin test set and the predictions of the proposed model and the BAMF-model (baseline). The reference renal mass mask ■ and the renal mass predictions ■ are shown. The predicted kidney ■ is also highlighted.

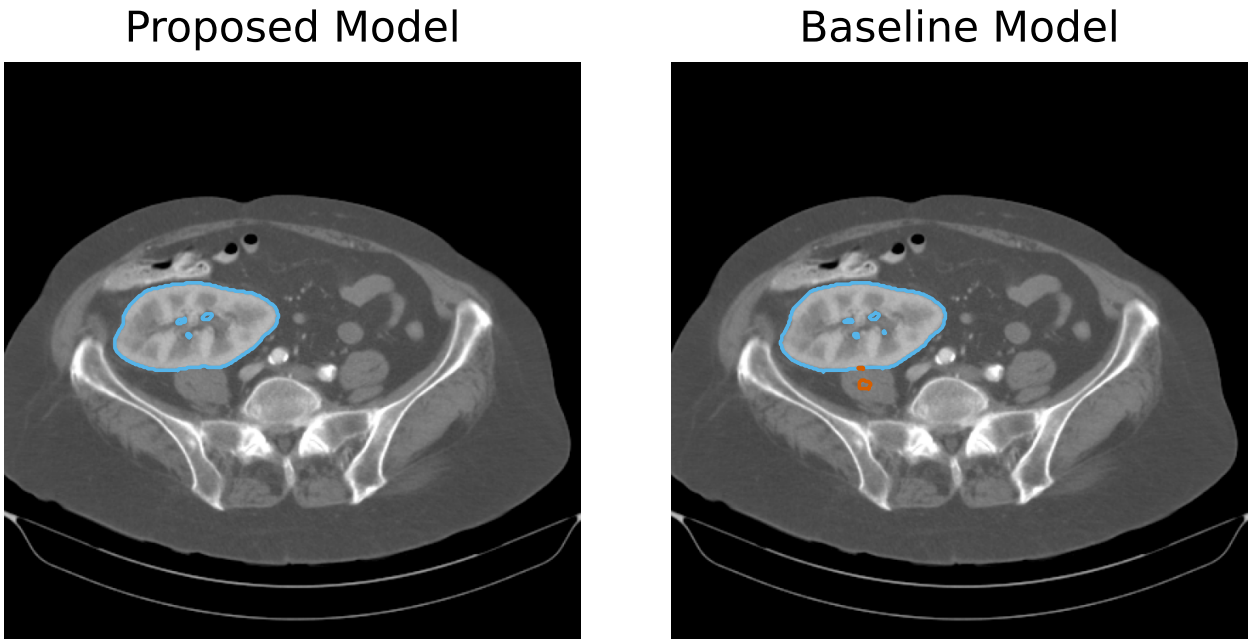

Figure 12: Presented is an interesting transplantation case from the Charité Universitätsmedizin Berlin test set and the predictions of the proposed model and the BAMF-model (baseline). The predicted kidney ■ is highlighted and shows the robustness of the model, despite the location being in the pelvic area.