Journal of
Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Benchmarking Wilms' tumor in multisequence MRI data: why does current clinical practice fail? Which popular segmentation algorithms perform well?

Sabine Müller
Iva Farag
Joachim Weickert
Yvonne Braun
André Lollert
Jonas Dobberstein
Andreas Hötker
Norbert Graf

Sabine Müller, Iva Farag, Joachim Weickert, Yvonne Braun, André Lollert, Jonas Dobberstein, Andreas Hötker, Norbert Graf, "Benchmarking Wilms' tumor in multisequence MRI data: why does current clinical practice fail? Which popular segmentation algorithms perform well?," *J. Med. Imag.* **6**(3), 034001 (2019), doi: 10.1117/1.JMI.6.3.034001.

Benchmarking Wilms' tumor in multisequence MRI data: why does current clinical practice fail? Which popular segmentation algorithms perform well?

Sabine Müller,[a,b,*] Iva Farag,[a] Joachim Weickert,[b] Yvonne Braun,[a] André Lollert,[c] Jonas Dobberstein,[a] Andreas Hötker,[d] and Norbert Graf[a]
[a]Saarland University, Medical Center, Department of Pediatric Oncology and Hematology, Homburg, Germany
[b]Saarland University, Faculty of Mathematics and Computer Science, Mathematical Image Analysis Group, Saarbrücken, Germany
[c]Johannes Gutenberg University, Medical Center, Department of Diagnostic and Interventional Radiology, Mainz, Germany
[d]University Hospital Zürich, Department of Diagnostic Radiology, Zürich, Switzerland

Abstract. Wilms' tumor is one of the most frequent malignant solid tumors in childhood. Accurate segmentation of tumor tissue is a key step during therapy and treatment planning. Since it is difficult to obtain a comprehensive set of tumor data of children, there is no benchmark so far allowing evaluation of the quality of human or computer-based segmentations. The contributions in our paper are threefold: (i) we present the first heterogeneous Wilms' tumor benchmark data set. It contains multisequence MRI data sets before and after chemotherapy, along with ground truth annotation, approximated based on the consensus of five human experts. (ii) We analyze human expert annotations and interrater variability, finding that the current clinical practice of determining tumor volume is inaccurate and that manual annotations after chemotherapy may differ substantially. (iii) We evaluate six computer-based segmentation methods, ranging from classical approaches to recent deep-learning techniques. We show that the best ones offer a quality comparable to human expert annotations. [DOI: 10.1117/1.JMI.6.3.034001]

Keywords: kidney; magnetic resonance imaging; segmentation.

Paper 19028RR received Jan. 29, 2019; accepted for publication Jun. 24, 2019; published online Jul. 19, 2019.

1 Introduction

Wilms' tumor, or nephroblastoma, accounts for 5% of all cancers in children and adolescents. It constitutes the most frequent malignant kidney tumor in childhood.[1] About 75% of all patients are younger than 5 years—with a peak between 2 and 3 years.[2,3] In Europe, diagnosis and therapy follow the guidelines of the International Society of Pediatric Oncology (SIOP).[4,5] One of the most important characteristics of this therapy protocol is a preoperative chemotherapy. Clinicians categorize patients as high-, intermediate-, or low-risk candidates according to histology, local stage, and tumor volume after chemotherapy. Postoperative treatment ranges from no chemotherapy (low risk stage I) up to chemotherapy with irradiation of the tumor bed (high risk, stages II and III).

The most common histological subtypes of regressive and mixed type actually belong to the intermediate-risk tumors. However, if, after chemotherapy, these tumors have a volume of more than 500 ml, they are highly malignant and the patients are treated according to the high-risk group protocol.[6] In order to avoid exposing children to unnecessary medical burden on the one hand and to maximize their chances of survival on the other hand, an exact determination of the tumor volume is indispensable.

Current practices of segmentation by human experts: Radiologists traditionally model the tumor through a time-intensive manual segmentation procedure involving the outlining of the gross tumor volume on numerous two-dimensional imaging "slices." Alternatively, they estimate the tumor volume by measuring three axes of tumor extension and assuming the nephroblastoma to have an ellipsoid shape.[6] Usually both variants are conducted using either computed tomography or magnetic resonance imaging (MRI) data. The reliability and consistent reproducibility of expert delineations of Wilms' tumors has not been investigated so far.

Computer-based segmentation algorithms: One obvious step to avoid the reproducibility problem is to replace human segmentations by automatic ones. Fully automatic segmentation of Wilms' tumors is a challenging task as these tumors do not show a discriminative texture, might have intensities overlapping with the surrounding tissue, and can be directly attached to the remaining kidney. To the best of our knowledge, there is no method available so far that does not need massive user interaction. Moreover, the scientific literature on computer-based segmentation algorithms for Wilms' tumors is fairly limited and shall be discussed next.

An initial idea for segmentation is to extend user-marked seed points in the tumor by region growing based on intensity thresholding.[7] A refined approach is to initialize an active contour inside the tumor and to expand the segmentation according to image intensities and gradients.[7] More recently, a more advanced energy-based method for segmentation of nephroblastoma has been proposed.[8] User-set scribbles are employed to approximate the gray value distributions of tumor and surrounding tissues. The energy is then regularized by an image metric induced by a state-of-the-art edge detection. However, this method still needs user interaction.

*Address all correspondence to Sabine Müller, E-mail: smueller@mia.uni-saarland.de

2329-4302/2019/$28.00 © 2019 SPIE

In spite of the fact that segmentation is an active research field in image analysis for quite some decades, it is remarkable that many well-established classes of algorithms have not been evaluated in the context of Wilms' tumor segmentation. Moreover, a comparative evaluation of these algorithms is prevented by the fact that there is no public benchmark available. So far the few computer-based algorithms for Wilms' tumor segmentation have been tested on different data sets.

1.1 *Our Contributions*

The goal of our paper is to offer solutions to the before mentioned problems in a threefold way.

i. We establish the first publicly available heterogeneous benchmark data set for Wilms' tumors. (Currently, the archives are password protected and can be accessed at the website in Ref. 9, password: spie2018. We have no information about who is accessing those files.)

   It allows clinicians to train their segmentation abilities and computer scientists to evaluate their algorithms. Our benchmark consists of multisequence MRI data before and after chemotherapy. Ground truth segmentations are approximated by consensus truth of five human experts.

ii. Based on this benchmark, we scrutinize the widely used ellipsoid approximation to the tumor volume as well as the interrater variability of manual delineations. Both results will reveal substantial shortcomings of the current standards.

iii. As a second benchmark application, we evaluate six algorithms w.r.t. their usefulness for Wilms' tumor segmentation. Although most of these segmentation algorithms are popular and time-proven methods in the computer vision community, none of them has been used for Wilms' tumor segmentation yet. Our algorithms include a fully automatic method based on Chan–Vese active contours,[10] a random forest classifier,[11] a support vector machine,[12] a k-means clustering algorithm,[13] and a clustering of superpixels.[14] Since the Wilms' tumor data are necessarily limited, most segmentation methods based on deep learning cannot be applied due to an insufficient amount of training data. One of the few methods that can be used is the U-Net,[15] which we are also evaluating.

In computer vision, benchmarking and performance evaluation have established themselves as important triggers for scientific progress in key areas ranging from motion analysis[16–18] over optimization algorithms[19] to segmentation methods.[20] Pure benchmarking and performance evaluation have become equally influential in medical image analysis,[21,22] e.g., in registration[23,24] and various segmentation problems.[25–28] The authors of these publications typically follow the clear scientific practice not to mix benchmark data with own unpublished algorithms, since this enables a fair comparison and avoids conflicts of interests. We adhere to these standards and refrain from proposing algorithms. We focus on evaluating the performance of popular fully automatic segmentation methods when being applied to Wilms' tumor data.

1.2 *Paper Organization*

Section 2 introduces our new multisequence benchmark for Wilms' tumor segmentation. We analyze interoperator variability and compare the determined volumes with volume approximations used in clinical practice. The third section evaluates human segmentations, and Sec. 4 is devoted to the evaluation of computer-based segmentation algorithms. Our conclusions are summarized in Sec. 5.

2 Benchmark Data

To describe our benchmark data set, we first present details on the acquired MRI data and the chosen method for ground truth approximation. Afterward, we introduce our error metrics and evaluate the interoperator variability on the proposed data set. In the end, we compare volume variability among human expert raters, ground truth, and ellipsoid shapes.

2.1 *Data Sets*

Our image data set consists of 28 multisequence MR scans from 17 Wilms' tumor patients (5 male and 12 female), out of which 15 have been acquired from intermediate risk tumor [histological diagnosis: stromal predominant (2), mixed histology (6), or regressive type (7)] and 2 from high risk tumor types (histological diagnosis: blastemal predominant). For 11 patients, we have both data before and after chemotherapy. The remaining ones are missing either data before or after chemotherapy. Figure 1 shows the age distribution of the children. Only patients with histologically confirmed Wilms' tumors were eligible for inclusion. The MRI sequences before and after chemotherapy for one of these patients are shown in Fig. 2.

Since it is difficult to obtain a comprehensive and representative set of Wilms' tumor data, the images have been acquired at different centers over the course of several years, using MR scanners from different manufacturers, varying field strength (1.5T and 3T) and implementations of the imaging sequences. The data sets used in our benchmark share the following three MRI settings.

- T2: T_2-weighted images, axial two-dimensional (2-D) acquisition with 3.6 to 9.1 mm slice thickness and inslice-sampling ranging from 0.3 to 1.4 mm.
- T1: T_1-weighted images, native image, axial 2-D acquisition with 2.5 to 9.1 mm slice thickness and inslice-sampling ranging from 0.5 to 1.6 mm.

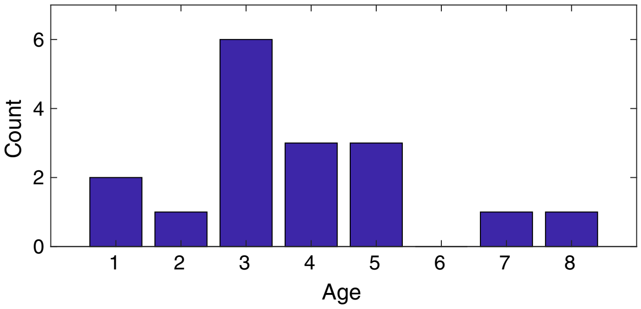

Fig. 1 Age distribution of patients whose images are made available anonymously.

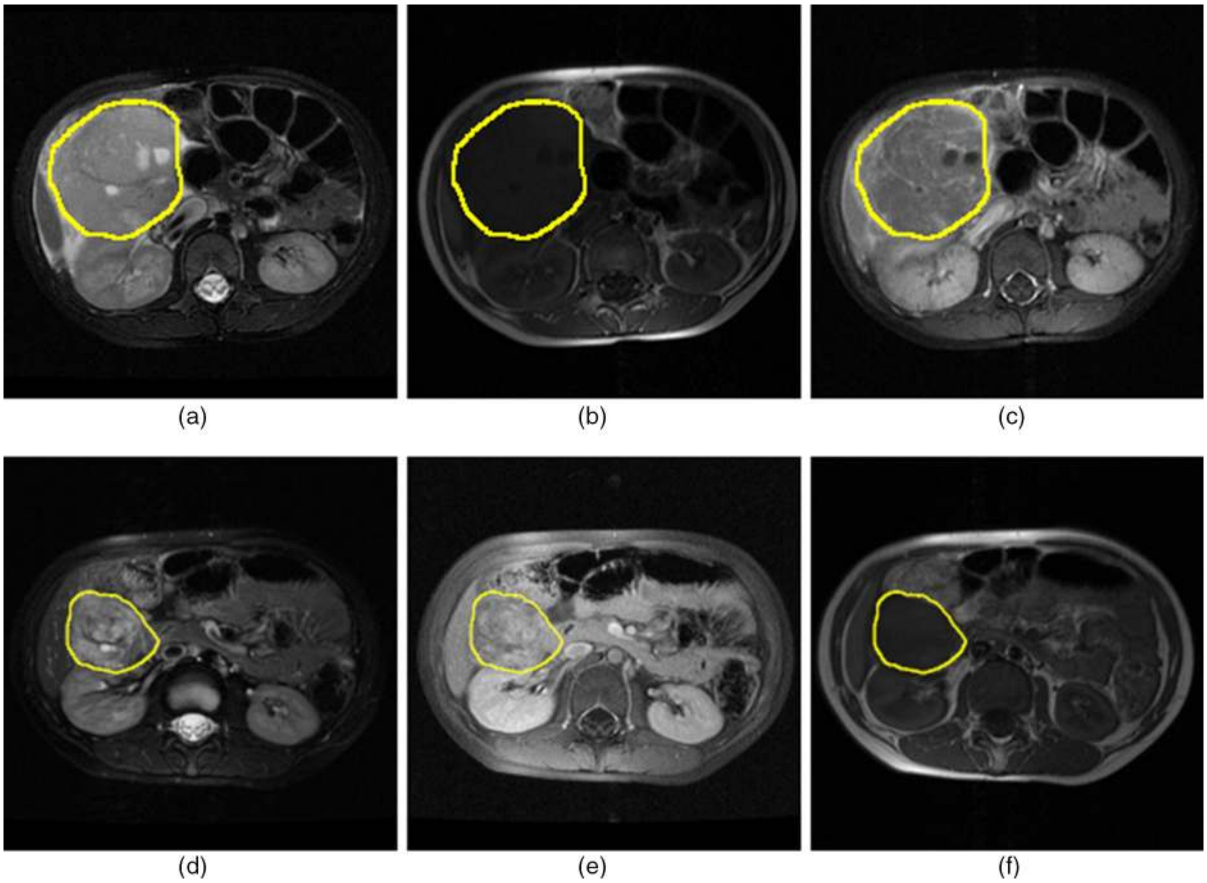

Fig. 2 Example of Wilms' tumor (training data) (a)–(c) before and (d)–(f) after chemotherapy with experts' consensus truth. From left to right: T2, T1, and T1c.

- T1c: T_1-weighted and contrast enhanced (gadolinium) images, axial 2-D acquisition with 1.8 to 7.7 mm slice thickness and inslice-sampling ranging from 0.5 to 1.6 mm.

The different MRI sequences were spatially coregistered on the T2 sequence using a rigid transformation. We balanced the number of slices with tumor areas before and after chemotherapy; see Table 1. Subtypes are not balanced among the data sets.

We deploy all images in NRRD-file format.[29] NRRD stands for "nearly raw raster data" and is a standard file format for storing medical image data, fully anonymized and without sensitive patient information.

2.2 *Annotations by Human Experts*

The images were manually annotated by five human expert raters coauthoring this publication. Rater-1 and rater-4 are experienced radiologists with several years of experience in Wilms' tumor analysis. Rater-2 is a physician familiar with Wilms' tumors. Rater-3 is an M.D. student previously trained in MRI imaging with advanced experience in the field. Rater-5 is an experienced oncologist with decades of practice in Wilms' tumor exploration. Segmentations were performed using the MITK software from Ref. 30, and experts outlined tumor structures in T2-sequences in every axial slice.

2.3 *Ground Truth Generation*

Since the generation of error-free ground truth information for medical images is usually not possible, we rely on expert votes to approximate the tumor area. Majority voting for each voxel has been shown to be useful in several contexts.[31,32] Unfortunately, this simple approach neither regards variability in quality or performance among the human raters nor does it provide guidance as to how many experts should agree before

Table 1 Image properties before and after chemotherapy. The values in brackets indicate the average occurrence.

	Training set		Test set	
	Slices	Slices with tumor	Slices	Slices with tumor
Prechemotherapy	19 to 55 (31)	9 to 25 (15)	26 to 50 (35)	11 to 28 (18)
Postchemotherapy	19 to 44 (30)	6 to 26 (12)	29 to 70 (54)	6 to 23 (13)

Table 2 Estimated quality parameters of each expert before and after chemotherapy. Rater-1, radiologist; rater-2, physician; rater-3, M.D. student; rater-4, radiologist; rater-5, oncologist.

	Rater-1	Rater-2	Rater-3	Rater-4	Rater-5
			Prechemotherapy		
Sensitivity	0.76	0.71	0.80	0.65	0.58
Specificity	0.65	0.75	0.73	0.82	0.74
			Postchemotherapy		
Sensitivity	0.78	0.60	0.77	0.70	0.67
Specificity	0.72	0.57	0.75	0.85	0.82

a voxel is labeled as tumor. Hence, we decide to use the STAPLE framework[33] to produce consensus segmentations.

The STAPLE algorithm uses expectation maximization. Let $D_{\mathbf{x},j}$, $j = 1, \ldots, n$ be the expert decisions and $\hat{\mathbf{G}}$ be the true consensus segmentation. The performance of each expert is rated on the basis of the sensitivity $p_j = \mathtt{Pr}(D_{\mathbf{x},j} = 1|\hat{\mathbf{G}} = 1)$ and the specificity $q_j = \mathtt{Pr}(D_{\mathbf{x},j} = 0|\hat{\mathbf{G}} = 0)$. It iterates between estimating the conditional probability of $\hat{\mathbf{G}}$ in relation to the expert decisions and previous estimates of the performance parameters and estimation of updated reliability parameters. Before chemotherapy, convergence is on average reached with less than 33 iterations. After chemotherapy, the algorithm converged on average after 52 iterations. The estimated quality parameters of each expert are shown in Table 2 and indicate high interrater variability.

Figure 3 shows annotations from all five human experts and the final ground truth approximation.

2.4 *Error Metrics*

We show results in terms of the metrics suggested in Ref. 28 and compute precision and recall as

$$P_{\hat{\mathbf{G}},\mathbf{G}} := \frac{|\hat{\mathbf{G}} \cap \mathbf{G}|}{|\mathbf{G}|}, \quad R_{\hat{\mathbf{G}},\mathbf{G}} := \frac{|\hat{\mathbf{G}} \cap \mathbf{G}|}{|\hat{\mathbf{G}}|}, \tag{1}$$

where $\hat{\mathbf{G}}$ is the experts' consensus truth, and $\mathbf{G}$ is the algorithmic prediction. The harmonic mean of precision and recall is called Dice score. It relates the area of a cluster to its voxelwise overlap with the approximated ground truth. The average Dice score determines the overall segmentation accuracy.

Another class of error measures evaluates the distance between the segmentation boundaries, i.e., the surface distance. The best known example of this is the Hausdorff distance.[34] It calculates for a given volume the shortest distance to all points on the surface of another volume and vice versa and finally extracts the maximal distance. However, the return of the maximum over all surface distances makes the Hausdorff measurement very susceptible to small remote subregions in either ground truth or segmentation result. In the evaluation of the fully automated methods, predictions with few false-positive areas, which only marginally influence the overall quality of the segmentation, can also dramatically influence Hausdorff's overall result. Therefore, we refrain from evaluating this error measure. It is not conclusive in our scenario.

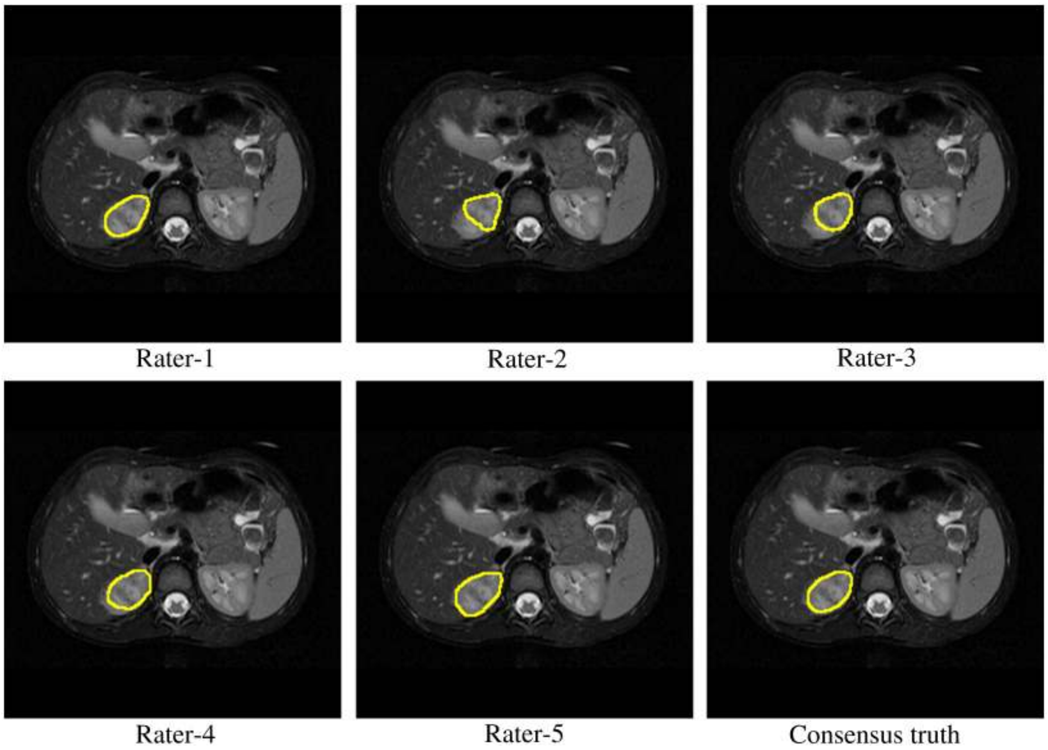

Fig. 3 Example annotations by human expert raters. Rater-1, radiologist; rater-2, physician; rater-3, M.D. student; rater-4, radiologist; rater-5, oncologist.

Table 3 Interoperator variability before and after chemotherapy in terms of Dice score. Rater-1, radiologist; rater-2, physician; rater-3, M.D. student; rater-4, radiologist; rater-5, oncologist.

	Rater-1	Rater-2	Rater-3	Rater-4
		Prechemotherapy		
Rater-2	0.85 ± 0.13			
Rater-3	0.89 ± 0.11	0.89 ± 0.08		
Rater-4	0.85 ± 0.13	0.90 ± 0.05	0.88 ± 0.08	
Rater-5	0.83 ± 0.13	0.89 ± 0.05	0.87 ± 0.07	0.89 ± 0.05
		Postchemotherapy		
Rater-2	0.63 ± 0.37			
Rater-3	0.83 ± 0.24	0.65 ± 0.37		
Rater-4	0.84 ± 0.10	0.65 ± 0.36	0.80 ± 0.24	
Rater-5	0.84 ± 0.10	0.64 ± 0.35	0.80 ± 0.24	0.89 ± 0.05

3 Evaluation of Human Expert Segmentations

3.1 *Accuracy*

3.1.1 *Interoperator variability*

We calculate the interoperator variability using all 28 data sets of all 17 patients. In order to do so, we compute the disagreement of the outlined volume marked by each physician with each volume outline prepared by each of the other four clinicians for the same data set. This process was repeated for each patient to provide a data set comprising the average disagreement between the five contours for each data set. We also divide the data sets based on their acquisition time relative to chemotherapy, i.e., before and after chemotherapy. Table 3 shows the interoperator variability in terms of Dice score before chemotherapy and after chemotherapy, respectively.

Before chemotherapy, the average Dice score between human experts shows their agreement on average with 0.87 ± 0.09 on tumor areas. After chemotherapy, when tumor tissues are barely visible, the average Dice score between human expert raters drops to 0.78 ± 0.24 indicating a high interrater variability. Especially after chemotherapy, rater-2 seems to be the bottleneck in agreement of the human experts. Therefore, we also computed the average Dice scores excluding this annotator. It turns out that average Dice score and standard deviation between human expert raters before chemotherapy slightly decreases to 0.87 ± 0.1. After chemotherapy, it improves to 0.83 ± 0.17, but still shows a high variability.

We also evaluated our expert annotations with McNemar's statistical test[35]

$$\chi^2 = \frac{|b - c|^2}{b + c}. \tag{2}$$

This χ^2-test for paired nominal data, based on the contingency matrix of these samples, provides information on whether there is a statistically significant difference. We calculated the corresponding matrix according to Table 4. Here, the first entry per field refers to the first expert to be compared and the second to the other. For example, field b means that the first of the two has labeled a pixel as tumor and the other as nontumor region. Furthermore, a significance level of $\alpha = 0.05$ corresponds approximately to a fiducial level of $\chi^2 = 3.8415$.

Table 4 Confusion matrix for McNemar's statistical test.

a: no / no	b: yes / no
c: no / yes	d: yes / yes

The results in Table 5 highlight the differences in expert annotations analogous to our previous analyses: all results of McNemar's test reject the null hypothesis that the annotations are similar with high values for all rater combinations. Unfortunately, it is not possible to compare these test results before and after chemotherapy: on the one hand, the tumor shrinks during therapy, and on the other hand, the resolution of the images is usually not the same. Both result in a different number of pixels in the contingency matrix.

3.1.2 *Deviation from ground truth*

The average Dice score before chemotherapy of human experts in comparison to ground truth is 0.93 ± 0.05. After chemotherapy, the contrast of tumor regions is usually lower and the tumor outlines are more ambiguous. Consequently, human experts agree less on tumor areas. The average Dice score decreases to 0.85, and variability increases dramatically to 0.16.

3.2 *Volume Variability*

Tumor expansion after preoperative chemotherapy is an important metric used to categorize patients as high-, intermediate-, or low-risk candidates. High-risk patients receive an additional postoperative chemotherapy aligned with an irradiation.

Table 5 Interoperator variability before and after chemotherapy in terms of McNemar's test averaged on all data sets. Rater-1, radiologist; Rater-2, physician; Rater-3, M.D. student; Rater-4, radiologist; rater-5, oncologist.

	Rater-1	Rater-2	Rater-3	Rater-4
		Prechemotherapy		
Rater-2	11,169			
Rater-3	7594	7158		
Rater-4	15,840	4683	8636	
Rater-5	19,106	7195	10,538	5916
		Postchemotherapy		
Rater-2	10,898			
Rater-3	2423	10,152		
Rater-4	11,668	11,765	9187	
Rater-5	13,733	14,598	11,293	1502

Therefore, an accurate determination of tumor volume is critical. The clinical volume equals the volume information used in therapy and treatment planning. It approximates the tumor by an ellipsoid shape. It is computed as $\texttt{width} \times \texttt{height} \times \texttt{depth} \times 0.524$,[6] where `width`, `height`, and `depth` of tumor denote the maximal expansion of tumor tissue on MR images. Note that the volume of the largest ellipsoid that fits in a cuboid is $\pi/6 \approx 0.524$ times the cuboid volume. Starting with the assumption that the true tumor volume is found through the consensus of our five human experts, we compare human expert annotations and clinical volumes in terms of percental volume differences in relation to the ground truth volume before and after chemotherapy, respectively. It turns out that clinical volumes differ before chemotherapy on average by $22.62 \pm 16.12\%$, and after chemotherapy by $35.07 \pm 41.01\%$ from the ground truth volumes. Before and after chemotherapy, clinical volumes are on average smaller than the ground truth volume, i.e., 85.71% before and 92.86% after chemotherapy. In contrast, human experts differ before chemotherapy on average by $10.58 \pm 5.90\%$ and after chemotherapy by $25.98 \pm 34.57\%$ from the ground truth volume.

This shows that assuming an ellipsoid shape for Wilms' tumors is an erroneous oversimplification, and human expert annotations are helpful to determine tumor volumes more precisely.

4 Evaluation of Segmentation Algorithms

In the following, we conduct example evaluations on our new benchmark data with six fully automatic methods:

- Chan–Vese active contours[10] with two level sets.
- K-means clustering[13] with intensities.
- Entropy rate superpixel segmentation.[14]
- Classification with a support vector machine[12] with intensities and HOG-features.[36]
- Random-forest classification,[11] either with intensities or HOG-features.[36]
- Segmentation with a U-Net.[15]

To guarantee a fair evaluation, we equally split the data sets in training and test data, each containing seven data sets before and after chemotherapy. For each segmentation approach, we include information from all modalities. Since the sampling rate in depth direction is substantially lower than in the other directions, we prefer to restrict ourselves to 2-D segmentations in the present manuscript. Exploring three-dimensional segmentations is reserved for future research. Let us now sketch each of the evaluated segmentation approaches.

4.1 *Chan–Vese Active Contours*

We consider a cubic data domain $\Omega \subset \mathbb{R}^3$ and a volumetric data set $\boldsymbol{f}: \Omega \to \mathbb{R}^3$. In our setting, the codomain describes the different MRI modalities T2, T1, and T1c. Then a segmentation of $\mathbf{f}$ by means of the Chan–Vese active contour model[10] minimizes the cost function

$$E(\mathbf{u}, C) = \lambda_{\mathtt{in}} \int_{C_{\mathtt{in}}} \|\mathbf{u}_{\mathtt{in}} - \mathbf{f}\|^2 \mathrm{d}\mathbf{x} + \lambda_{\mathtt{out}} \int_{C_{\mathtt{out}}} \|\mathbf{u}_{\mathtt{out}} - \mathbf{f}\|^2 \mathrm{d}\mathbf{x} + \nu \ell(C), \tag{3}$$

where the data domain Ω is split into two regions C_{in} and C_{out}. The function $\mathbf{f}$ is approximated by a piecewise constant function where $\mathbf{u}_{\mathtt{in}}$ and $\mathbf{u}_{\mathtt{out}}$ are the arithmetic means of $\mathbf{f}$ inside and outside the segment boundaries C, respectively. The positive weights $\lambda_{\mathtt{in}}$ and $\lambda_{\mathtt{out}}$ control the influence of each region to the final partitioning, $\|.\|$ denotes the Euclidean norm in $\mathbb{R}^3$, and C are the segment boundaries with a (Hausdorff) length of $\ell(C)$. This length is weighted with a parameter $\nu > 0$.

4.2 *K-means Clustering*

K-means clustering[13] is a vector quantization method that partitions n observations into k clusters. Data points are assigned to cluster centers, prototypes of corresponding classes, with minimal Euclidean distance. In our application, we want to split the observations into two classes, tumor, and nontumor points.

Given a set of data points $\mathbf{f}: \Omega \to D$ with $D \subset \mathbb{R}^3$ and $\Omega \subset \mathbb{R}^3$, k-means minimizes

$$E(D_1, D_2) = \int_{D_1} \|\xi - \mathbf{u_1}\|^2 \mathrm{d}\xi + \int_{D_2} \|\xi - \mathbf{u_2}\|^2 \mathrm{d}\xi, \quad D = D_1 \cup D_2, \quad D_1 \cap D_2 = \emptyset, \tag{4}$$

where $\mathbf{u}_1$ and $\mathbf{u}_2$ are the arithmetic means of both classes. In this case, k-means clustering is equivalent to Otsu's method.[37]

4.3 *Support Vector Machine*

Support vector machines[12] are based on the concept of hyperplanes in a multidimensional space, separating between sets of objects having different classes, e.g., tumor and nontumor points. In our application, we use a fivefold cross validation to find optimized hyperparameters. Training was performed using MATLAB[38] and the problem was solved via sequential minimal optimization.[39] Furthermore, we used Gaussian-like kernels and the classification error, i.e., the weighted fraction of misclassified observations, as loss function.

4.4 *Random-Forest Classification*

Ensemble methods employ a finite set of different learning algorithms to get better predictive performance than using a single learning algorithm. Random forests[11] are ensemble approaches for classification combining a group of decision trees. A single tree is highly sensitive to noise, while the average of many decorrelated trees is not. Training all decision trees of a random forest on the same training data would result in strongly correlated trees. Bagging (bootstrap aggregation) generates new training sets $\mathbf{K}$ by sampling from the original training set $\mathbf{Y}$ uniformly and with replacement. In this way, decision trees are decorrelated by using different training data. In addition, random forests use feature bagging, i.e., features are randomly sampled for each decision tree.[40] To estimate how well the results can be generalized, we use two-fold-cross validation, i.e., we train two sets of models.

4.5 *Entropy Rate Superpixel Segmentation*

The method of Liu et al.[14] formulates the superpixel segmentation problem as maximization of the entropy rate of cuts in the graph. Optimizing this entropy rate encourages the clustering of compact and homogeneous regions, which also favors the superpixels to overlap with only one single object on the perceptual boundaries.

This technique starts with each pixel being considered as a separate cluster. Clusters are then gradually merged into larger superpixels. In this way, during segmentation, a hierarchy of superpixels is created until finally only one superpixel, the image itself, is left. In our case, we want to segment a tumor, i.e., we use the hierarchy of superpixels to divide the image into three groups: tumor, body, and background. Unfortunately, we do not know in advance which superpixel contains which class. This objective function is optimized with a greedy algorithm.

4.6 *U-Net*

In many areas of medical image processing, deep learning and especially convolutional neural networks (CNN) have proven to be very powerful tools. Within these, the U-net architecture[15] is one of the standard CNNs in the field of medical image segmentation. It learns segmentation in an end-to-end setting and only needs a few training examples. Since our benchmark consists of real clinical data, they are available in different resolutions. Some of them also contain other parts of the body, e.g., the arms. Therefore, the amount of nontumor areas outweighs the tumor areas substantially, such that it becomes necessary to balance the classes. This is done in three steps: first we determine the connected components, i.e., connected parts of the body and remove everything except the largest one. Then we determine the maximum extent of the existing object and extract this part to a new, smaller image; see Fig. 4. This is then rescaled to a size of 512×512 pixels. We use the implementation presented in Ref. 41 to solve our segmentation problem and set up the network with batch size 5 and 50 epochs.

4.7 *Results*

In Table 6, we present the mean precision, recall, and Dice score over the 14 test data sets of the different segmentation algorithms. Since the Chan–Vese method is region-based, it suffers from the fact that the visual appearance of Wilms' tumors can be highly heterogeneous. Our experiments show that intensities are an important feature to identify tumor areas, resulting in high precision values for the pixel-based classifiers *k*-means clustering and random forests. However, spatial information is essential as intensities of a tumor can overlap with those of the surrounding tissue. Accordingly, the pixel-based methods suffer from low recall. Using HOG-features in addition to intensities improves *k*-means clustering after chemotherapy, SVM classification as well as random forests both before and after chemotherapy.

The results of the superpixel-based method are unexpectedly poor both before and after chemotherapy. The optimum number of superpixels depends strongly on the image, and it is also difficult to identify the respective segments. We could not find

Table 6 Results on the proposed benchmark data set (test data). *k*-means, *k*-means clustering; CV, Chan–Vese active contours; RF, random forest classification; SVM, support vector machine; INT, intensity values; HOG, HOG-features; PP, postprocessing. Best results are depicted in boldface.

Method	Dice score	Precision	Recall
	Prechemotherapy		
CV[10]	0.57	0.48	0.69
k-means[13] (INT)	0.53	0.76	0.41
Superpixel[14]	0.41	0.33	0.56
SVM[12] (INT + HOG[36])	0.71	0.71	0.72
RF[11] (INT + HOG[36])	**0.92**	**0.92**	0.91
U-net[15]	0.64	0.49	**0.94**
	Postchemotherapy		
CV[10]	0.41	0.32	0.58
k-means[13] (INT)	0.35	0.50	0.27
Superpixel[14]	0.41	0.29	0.68
SVM[12] (INT + HOG[36])	0.68	0.69	0.67
RF[11] (INT + HOG[36])	**0.81**	**0.73**	**0.92**
U-net[15]	0.30	0.25	0.61

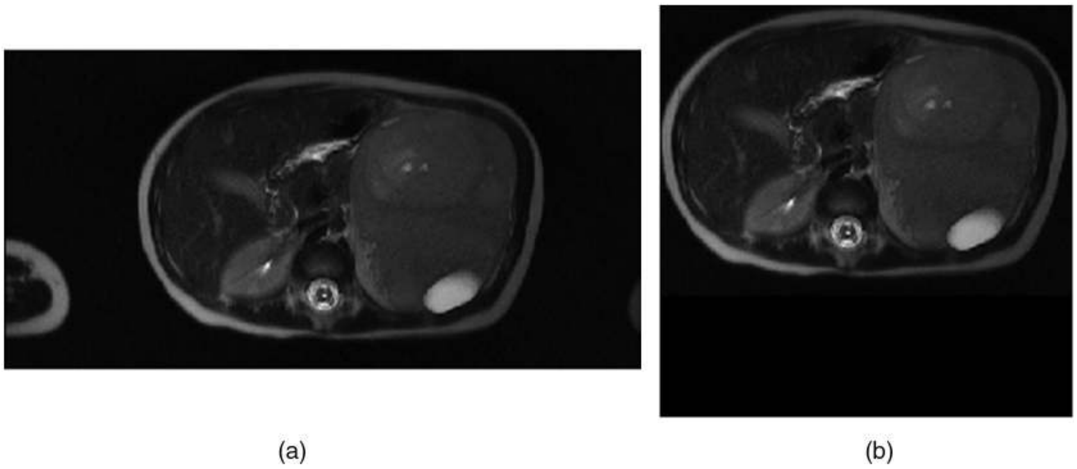

Fig. 4 Exemplary preprocessing step for the U-Net. (a) Original image containing abdomen and extremities and (b) image after preprocessing.

a parameter set that worked on all data sets. Deep learning methods usually require a large amount of training data. The U-net used here deviates from this paradigm and can also be trained with smaller amounts of data. Table 6 shows that it gives a high mean recall but a low mean precision. This indicates that although the network can recognize the basic structure of the nephroblastoma, it is not able to distinguish it from similar tissue.

Overall, segmentation with random forests provides the best results before chemotherapy but is also the leading approach after chemotherapy, yielding the highest quality measures. Therefore, we suggest random forests trained on HOG-features as well as intensities as the baseline method for this benchmark data set. Since the tumor volume after chemotherapy is decisive for postoperative treatment planning, it is currently the optimal method for this purpose. The segmentation quality lies within the variability of human experts.

In order to ensure spatial consistency, we also apply Chan–Vese active contours on the predicted probabilities of the random forest. It turns out that predictions of this method lack too much global information and the resulting segmentation loses quality. These observations highlight the challenges in the data set.

5 Conclusions

We have proposed the first multisequence benchmark for segmentation of Wilms' tumors. In spite of the fact that such a data set involving tumors in children is necessarily limited in size, its amount of information is rich: there are multisequence MRT images for all patients, and for 11 patients both pre- and postchemotherapy images. That is supplemented by manual annotations by five independent human experts, as well as histological diagnoses.

Our benchmark allows several important conclusions.

We have demonstrated that human expert annotations suffer from a large interoperator variability especially after preoperative chemotherapy. Furthermore, we have shown that the popular tumor volume determination based on ellipsoid shapes tends to be highly erroneous.

Our data set also allowed evaluation of six computer-based algorithms. At this time, all fully automatic segmentations apart from random forests undersegment the tumor volume compared to human expert raters. Thus, their precision is insufficient, especially after chemotherapy. Our experiments indicate that segmentation with random forests[11] is the most appropriate tool for Wilms' tumors. Its results lie within the variability of the contouring performed by human expert raters on the same data. Moreover, it offers the advantage that it is much faster than a full segmentation by human experts.

In our ongoing research, we plan to include more anatomical knowledge into our segmentation strategies and to constantly enlarge the number of available data sets. It is our hope that our benchmark data set for segmentation of nephroblastoma will stimulate a growing interest in this research field which is challenging both from a medical and a computer vision viewpoint. Most importantly, we are confident that the resulting progress will help to maximize the survival chances of the affected children.

Disclosures

We do not have conflicts of interest.

Acknowledgments

This work was partially funded by the European Union's seventh framework program under the project Computational Horizons in Cancer (Grant Agreement No. 600841). J.W. has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No. 741215, ERC Advanced Grant No. INCOVID).

References

1. G. Pastore et al., "Malignant renal tumours incidence and survival in European children (1978–1997): report from the automated childhood cancer information system project," *Eur. J. Cancer* **42**(13), 2103–2114 (2006).
2. A. M. Davidoff, "Wilms' tumor," *Curr. Opin. Pediatr.* **21**(3), 357–364 (2009).
3. S. Kim and D. H. Chung, "Pediatric solid malignancies: neuroblastoma and Wilms' tumor," *Surg. Clin. N. Am.* **86**(2), 469–487 (2006).
4. N. Graf, M.-F. Tournade, and J. de Kraker, "The role of preoperative chemotherapy in the management of Wilms' tumor: the SIOP studies," *Urol. Clin. N. Am.* **27**(3), 443–454 (2000).
5. S. C. Kaste et al., "Wilms tumour: prognostic factors, staging, therapy and late effects," *Pediatr. Radiol.* **38**(1), 2–17 (2008).
6. N. Graf, H. Reinhard, and J. O. Semler, "SIOP 2001/GPOH Therapieoptimierungsstudie zur Behandlung von Kindern und Jugendlichen mit einem Nephroblastom," http://www.kinderkrebsinfo.de (2003).
7. R. David et al., "Clinical evaluation of DoctorEye platform in nephroblastoma," in *Proc. 5th Int. Adv. Res. Workshop In Silico Oncol. and Cancer Invest.*, IEEE, pp. 1–4 (2012).
8. S. Müller et al., "Robust interactive multi-label segmentation with an advanced edge detector," *Lect. Notes Comput. Sci.* **9796**, 117–128 (2016).
9. www.mia.uni-saarland.de/wilms-benchmark.
10. T. F. Chan, B. Y. Sandberg, and L. A. Vese, "Active contours without edges for vector-valued images," *J. Visual Commun. Image Represent.* **11**(2), 130–141 (2000).
11. L. Breiman, "Random forests," *Mach. Learn.* **45**(1), 5–32 (2001).
12. B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. Fifth Annu. Workshop Comput. Learn. Theory*, ACM, New York, pp. 144–152 (1992).
13. S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory* **28**, 129–137 (1982).
14. M.-Y. Liu et al., "Entropy rate superpixel segmentation," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, IEEE, Providence, Rhode Island, pp. 2097–2104 (2011).
15. O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," *Lect. Notes Comput. Sci.* **9351**, 234–241 (2015).
16. J. L. Barron, D. J. Fleet, and S. S. Beauchemin, "Performance of optical flow techniques," *Int. J. Comput. Vision* **12**(1), 43–77 (1994).
17. S. Baker et al., "A database and evaluation methodology for optical flow," *Int. J. Comput. Vision* **92**(1), 1–31 (2011).
18. A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, IEEE, pp. 3354–3361 (2012).
19. J. Kappes et al., "A comparative study of modern inference techniques for discrete energy minimization problems," *Int. J. Comput. Vision* **115**(2), 155–184 (2015).
20. D. Martin et al., "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. Eighth IEEE Int. Conf. Comput. Vision*, IEEE, Vol. 2, pp. 416–423 (2001).
21. A. Hanbury, H. Müller, and G. Langs, *Cloud-Based Benchmarking of Medical Image Analysis*, Springer, Cham (2017).
22. C.-W. Wang et al., "A benchmark for comparison of dental radiography analysis algorithms," *Med. Image Anal.* **31**, 63–76 (2016).
23. K. Murphy et al., "Evaluation of registration methods on thoracic CT: the EMPIRE10 challenge," *IEEE Trans. Med. Imaging* **30**, 1901–1920 (2011).

24. Z. Xu et al., "Evaluation of six registration methods for the human abdomen on clinically acquired CT," *IEEE Trans. Biomed. Eng.* **63**, 1563–1572 (2016).
25. O. Bernard et al., "Standardized evaluation system for left ventricular segmentation algorithms in 3D echocardiography," *IEEE Trans. Med. Imaging* **35**(4), 967–977 (2016).
26. R. Karim et al., "Evaluation of state-of-the-art segmentation algorithms for left ventricle infarct from late Gadolinium enhancement MR images," *Med. Image Anal.* **30**, 95–107 (2016).
27. O. Maier et al., "ISLES 2015: a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI," *Med. Image Anal.* **35**, 250–269 (2017).
28. B. H. Menze et al., "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imaging* **34**(10), 1993–2024 (2015).
29. "NRRD: nearly raw raster data," http://teem.sourceforge.net/nrrd/index.html (25 April 2019).
30. I. Wolf et al., "The medical imaging interaction toolkit (MITK): a toolkit facilitating the creation of interactive software by extending VTK and ITK," *Proc. SPIE* **5367** (2004).
31. R. A. Heckemann et al., "Automatic anatomical brain MRI segmentation combining label propagation and decision fusion," *NeuroImage* **33**, 115–126 (2006).
32. V. C. Raykar et al., "Supervised learning from multiple experts: whom to trust when everyone lies a bit," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, ACM, New York, pp. 889–896 (2009).
33. S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation," *IEEE Trans. Med. Imaging* **23**(7), 903–921 (2004).
34. R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Vol. **317**, Springer Science & Business Media, Berlin (2009).
35. Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika* **12**(2), 153–157 (1947).
36. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, San Diego, California, pp. 886–893 (2005).
37. D. Liu and J. Yu, "OTSU method and K-means," in *Proc. IEEE Ninth Int. Conf. Hybrid Intell. Syst.*, IEEE, Vol. 1, pp. 344–349 (2009).
38. MATLAB, version 9.4 (R2018a), The MathWorks Inc., Natick, Massachusetts, www.mathworks.com/products/matlab (2018).
39. R.-E. Fan, P.-H. Chen, and C.-J. Lin, "Working set selection using second order information for training support vector machines," *J. Mach. Learn. Res.* **6**, 1889–1918 (2005).
40. T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(8), 832–844 (1998).
41. J. Akeret et al., "Radio frequency interference mitigation using deep convolutional neural networks," *Astron. Comput.* **18**, 35–39 (2017).

Sabine Müller received her BSc degree in computer science from Saarland University, Saarbrücken, Germany, in 2011, and her MSc degree in visual computing in 2014. Currently, she is a PhD candidate in the Mathematical Image Analysis Group at Saarland University and research assistant at the Department of Pediatric Oncology and Hematology, Saarland University Medical Center. Her research interests are in the areas of medical image segmentation, computer vision, and pattern recognition.

Iva Farag received her BSc degree in computer science from Saarland University in 2015 and her master's degree in computer science with specialization in the field of exploratory data analysis in 2018. Her research interests include pattern recognition, data analysis and summarization, and machine learning.

Joachim Weickert is a professor of mathematics and computer science at Saarland University, Saarbrücken, Germany, where he heads the Mathematical Image Analysis Group. He graduated and received his PhD from the University of Kaiserslautern, Germany, in 1991 and 1996. He worked as a postdoctoral researcher at the University Hospital of Utrecht and the University of Copenhagen, and as assistant professor at the University of Mannheim. He is the editor-in-chief of the *Journal of Mathematical Imaging and Vision*.

Yvonne Braun received her PhD from the University of Münster, Germany, in 2005, where she carried out research on the topic of telomerase as therapeutic target in pediatric tumors. Afterwards, she worked as a clinical research associate in the field of oncology and diabetes. Since 2011, she has worked as a research associate in the Department of Pediatric Oncology and Hematology, Saarland University Medical Center, Germany.

André Lollert graduated from medical school in Mainz, Germany, on May 6, 2011. On July 5, 2011 he was certified as a medical doctor. After internship and residency at the Department of Diagnostic and Interventional Radiology (08/2011-09/2013 and 10/2014-01/2017), and Section of Pediatric Radiology (10/2013-09/2014), Medical Center of the Johannes Gutenberg University Mainz, he was certified as a radiologist on January 25, 2017. Since then he has been working as a radiologist at the Section of Pediatric Radiology.

Jonas Dobberstein received his general qualification for university entrance (A-levels) in 2011 at Tilemannschule in Limburg, Germany. He started his medical studies in 2011 at Saarland University, Germany. Currently, he is a medical student at the Department of Pediatric Oncology and Hematology at Saarland University Medical Center, Germany.

Andreas Hötker graduated from medical school in 2009 and continued as a resident in the Department of Radiology of the University Medical Center Mainz, where he was responsible for urogenital imaging as an attending assistant, after becoming a board-certified radiologist in 2016. He completed postdoctoral research fellowships at Memorial Sloan Kettering Cancer Center, New York, from 2013 to 2014 and in 2016, and has published numerous original research articles, particularly investigating functional MR techniques in urogenital oncology.

Norbert Graf is a professor of pediatrics and director of pediatric oncology and hematology at Saarland University. He is the chairman of the Renal Tumour Study Group of the International Society of Paediatric Oncology, an associate member of the Children's Oncology Group of North America, an external reviewer for the Japan Science and Technology Agency, a member of the board of the VPH-Institute and has more than 25 years of experience in running clinical trials.