Research Article

Ai-Assisted Planning of Nephron-Sparing Surgery for Wilms Tumour: Lessons from A Case Audit

Artem Evgenievich Mishvelov[1,2]*, Lima Turpalalievna Ortsukhaeva[3], Petimat Khamzatovna Mutakaeva[3], Ivan Aleksandrovich Bil[4], Nikita Alekseevich Solodunov[4], Sergei Asvaturovich Kiliakhov[4], Snezhana Olegovna Svoevolina[4], Elizaveta Georgievna Mzhavanadze[5], Nikita Viktorovich Pilipenko[6], Vladlena Dmitrievna Trusova[6]

[1] Specialist in Medical Image Processing and Surgical Planning, Department of Radiology, Stavropol Regional Clinical Consultative and Diagnostic Center; Stavropol, Russia
[2] General Director, Fenixmed AI LLC, Stavropol, Russia
[3] Institute of Clinical Medicine, Saratov State Medical University named after V.I. Razumovsky, Saratov, Russia
[4] Faculty of Medicine and Prevention, Rostov State Medical University, Rostov-on-Don, Russia
[5] Faculty of Pediatrics, Rostov State Medical University, Rostov-on-Don, Russia
[6] Faculty of Medicine, Stavropol State Medical University, Stavropol, Russia
*** Corresponding author**: Artem Evgenievich Mishvelov, e-mail: archi4717@yandex.ru, ORCID: 0000-0002-5087-2283

Abstract

This work combines a retrospective, case-based audit of an artificial intelligence (AI) imaging program with a literature review. Nephron-sparing surgery for Wilms tumour requires a precise preoperative understanding of how the tumour relates to the residual parenchyma and renal vessels; AI software can accelerate this, yet its behaviour on paediatric data is rarely documented. We report a boy aged 1 year and 7 months with nephroblastoma of the left kidney who received preoperative chemotherapy followed by partial nephrectomy. Using de-identified data, we assessed two components of the MedVision AI program: automated segmentation and an embedded AI assistant built on a language model. Segmentation produced plausible tumour and parenchymal masks but failed to delineate the renal veins, probably because of the contrast phase. The AI assistant correctly named nephroblastoma as the most likely diagnosis, but described changes relative to an examination it had never received, assigned a stage that can only be established surgically and histologically, and contradicted itself. These observations are compared with the 2021-2026 literature on segmentation, multimodal language models and automation bias. We propose a workflow in which every AI statement is linked to its input, comparison over time is blocked until the prior examination is loaded, volume is calculated from a verified mask, and staging is excluded from the AI assistant's tasks. AI proves useful for rapid orientation and three-dimensional model preparation, but its conclusions require physician verification, and a response formatted as a radiology report may encourage uncritical acceptance of errors.

Keywords: Wilms tumour; nephron-sparing surgery; deep learning segmentation; large language models; AI hallucination; three-dimensional reconstruction

Introduction

Wilms tumour remains the most common malignant renal tumour of childhood, and with contemporary multimodal treatment most patients are cured [1, 2]. As survival has improved, attention has shifted to the cost of cure: survivors accumulate cardiovascular, renal and endocrine late effects of therapy [3], and reduced kidney function and arterial hypertension occur more often in them than in their siblings decades after treatment [4, 5]. This explains the interest in nephron-sparing surgery for unilateral disease. Its oncological efficacy and late renal effects have been assessed in a meta-analysis [6], the arguments for preserving nephrons in children without a syndromic predisposition have been summarised in a review [7], and an international Delphi consensus of surgeons has defined the conditions under which such surgery is acceptable [8]. Partial nephrectomy carries its own complications [9], and tumour size and the relationship of the tumour to the renal sinus limit the applicability of current selection criteria [10].

The decision to resect is based on imaging, both in the SIOP-RTSG strategy with preoperative chemotherapy [11, 12] and in COG protocols [13, 14]. Three-dimensional reconstruction improves surgeons' understanding of tumour anatomy compared with reviewing cross-sectional slices [15], and virtual resection allows the volume of the remaining parenchyma to be estimated in advance [16]. Manual segmentation remains the bottleneck, so automation with deep learning has become the logical next step [17, 18]. In parallel, multimodal language models capable of producing a description and a preliminary conclusion from an image and a text prompt are being embedded in image-viewing software [19, 20].

The evidence for such AI assistants comes mainly from examination-style and adult cases [21, 22]. Paediatric oncological imaging, by contrast, suffers from a shortage of data for training and validation: the tumours are rare, the examinations heterogeneous, and the anatomy of a young child differs from that of an adult [23]. Under these conditions, a carefully analysed single case from routine practice reveals error types that remain invisible in average accuracy figures obtained on large adult datasets.

The aim of this work was to perform, using de-identified data from a case of Wilms tumour in a young child, a retrospective audit of two AI components within one program (automated segmentation and an AI assistant), to

compare the strengths and errors identified with the published evidence, and to propose a workflow for the safe use of AI in preparing for nephron-sparing surgery.

Materials and Methods

The work is a literature review built around a clinical case; the audit was performed retrospectively. Clinical information was provided by the treating team. Images were exported from the MedVision AI "Physician Workstation" program (research version 1.14) and de-identified: the patient's name and examination identifier were removed and interface elements were cropped. The analysis used contrast-enhanced multidetector computed tomography (CT) of the abdomen with a 512 × 512 matrix and 301 axial slices; the examination phase had been labelled as arterial by the physician in the program, and information on contrast timing was not available to the authors.

Two components of the program were audited. The first was a set of automated segmentation modules ("kidney, tumour, cyst"; "cortex, medulla, collecting system"; "renal vessels, small branches") that generate masks for three-dimensional display and measurement; these were assessed using the measurement panel and by visual comparison of the masks with the source slices. The second was an embedded AI assistant built on a language model that accepts images and a free-text prompt. Its response was divided into individual statements, and each statement was compared with what the AI assistant had actually received as input and with the clinical data of the treating team. The categories draw on published taxonomies of language-model hallucination in medicine [24, 25] and were defined as follows:

- "confirmed": the statement matches an independent source (a program mask or the treating team's clinical diagnosis); histological verification was not part of this work;
- "consistent, not verified": the statement does not contradict the input, but verifying it would require data that were not available (for example, calibrated source images);
- "not supported by the input": the input contained no information allowing the statement to be made either way;
- "task not performed": a requested element of the response is missing;
- "fabrication": a statement about data that were demonstrably not provided;
- "internal contradiction": the statement is incompatible with another statement in the same response;
- "beyond scope": a conclusion for which clinical protocols require a different type of data (operative or histopathological).

Statements were classified by specialists in the relevant fields; the program developer, who is responsible for configuring and training the AI models, did not take part in the classification.

The literature search was performed in PubMed for January 2021 to October 2026 using the following query blocks: "Wilms tumor" AND ("nephron-sparing surgery" OR "three-dimensional" OR "segmentation" OR "deep learning"); "large language model" OR "GPT-4V" OR "multimodal" AND "radiology"; "hallucination" OR "sycophancy" AND "large language model"; "automation bias" AND "radiology"; "artificial intelligence" AND ("regulation" OR "cost-effectiveness") AND "radiology". Original studies, systematic reviews and consensus documents were included; landmark studies published before 2021 were included when current protocols rely on them. All references were checked manually by the authors against PubMed and publisher records.

1. Clinical case

1.1. Patient and treatment

A boy aged 1 year and 7 months was diagnosed with nephroblastoma of the left kidney. Treatment began with preoperative chemotherapy. During chemotherapy, contrast-enhanced abdominal CT was performed, and comparison with an examination dated 18 August 2026 was planned to assess the course of the disease. In September 2026, the child underwent resection of the left kidney with removal of the tumour, that is, nephron-sparing surgery.

The analysis concerns the preoperative stage: how the program processed the CT and what its AI assistant reported to the physician. Histopathological data and the postoperative course are not included, and no conclusions about treatment outcome are drawn. The decision to operate was made by the treating team independently of the AI assistant's response.

1.2. Image processing and automated segmentation

The automated segmentation modules added seven mask layers and displayed a message asking the user to check each mask on the slices. The left renal mass was outlined by a separate mask labelled "type not differentiated": the module distinguished it from normal parenchyma but did not attempt to classify it (Fig. 1a). On axial and coronal slices, the tumour mask follows the contours of the mass, and the parenchyma of both kidneys is labelled separately (Fig. 1b, 1c). The arterial and venous masks were marked as preliminary.

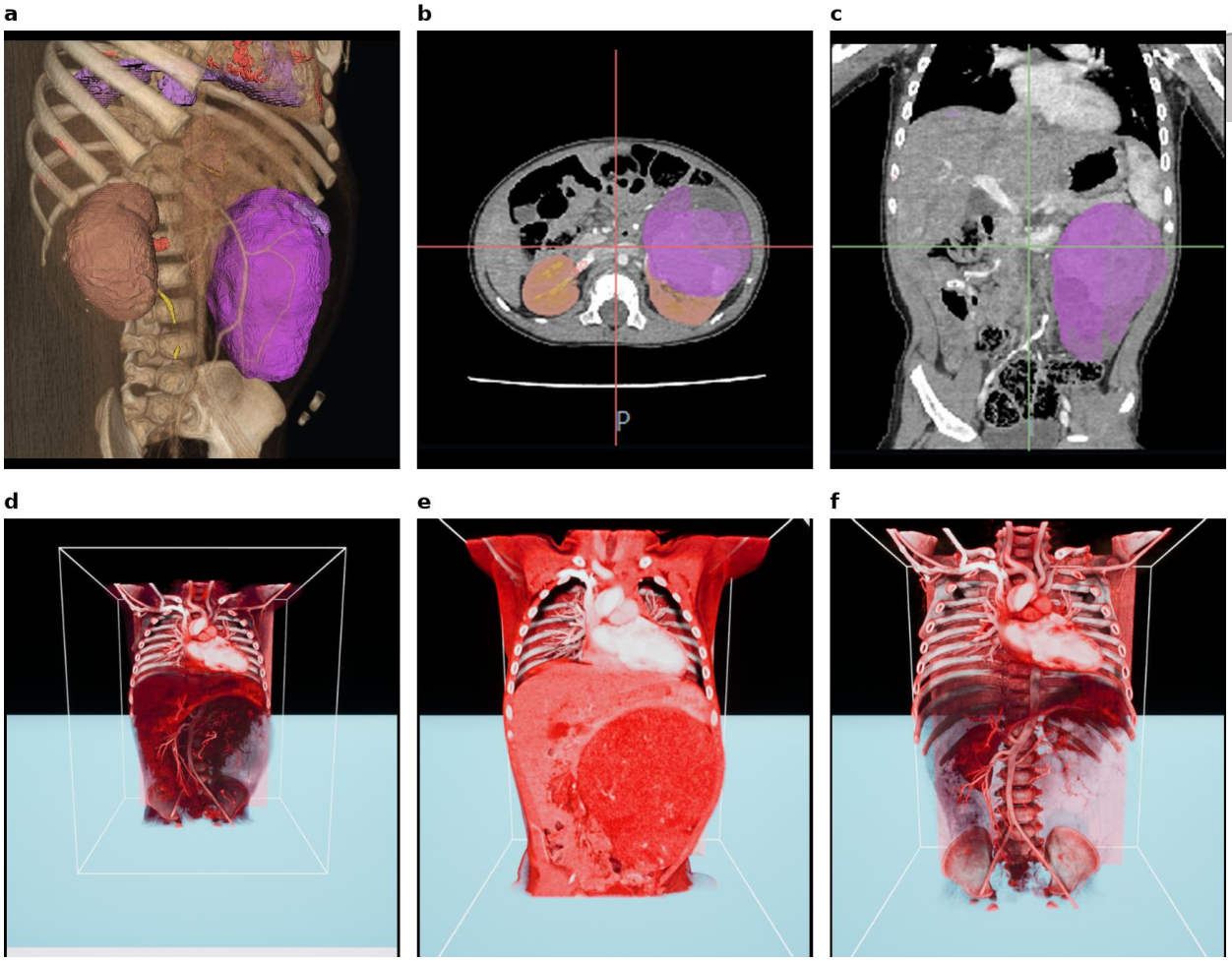


Figure 1. CT processing in the MedVision AI program. (a) Three-dimensional display of the automated masks: the left renal mass (purple mask) and the parenchyma of the contralateral kidney; (b) axial and (c) coronal slices with overlaid masks; (d-f) volume-rendering modes showing the vascular tree, the mass and increased tissue transparency. The patient's name and examination identifier have been removed.

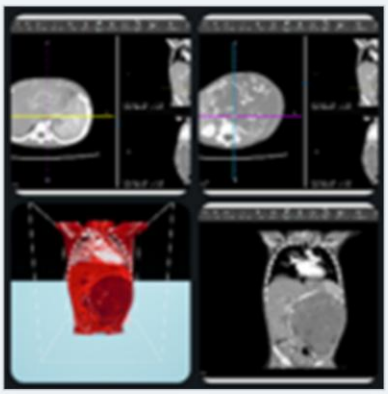
The renal vein mask failed: its volume was 0.0077 cm³ with 47 voxels, whereas its bounding box measured 85.4 × 25.0 × 24.0 mm. This combination indicates that the mask consisted of scattered fragments rather than a continuous vessel. The likely cause is the arterial phase of the examination, in which the veins are not yet sufficiently opacified; a definitive conclusion would require information on contrast timing. The program marked the mask as preliminary, that is, it did not present it as a result. For the surgeon, the implication is clear: venous anatomy cannot be assessed from these data.

Three volume-rendering modes separately showed the vascular tree with the aorta and its branches (Fig. 1d), a large rounded mass in the left half of the abdomen (Fig. 1e) and, with increased tissue transparency, both kidneys, the spine and the vessels lying behind the tumour (Fig. 1f). These modes add no new diagnostic information but make it easier for a multidisciplinary team to discuss the relationships of the tumour with neighbouring structures, which is how the treating physicians defined the clinical task.

2. The AI assistant's response and its audit

The physician uploaded four reduced screenshots to the AI assistant (two multiplanar slices, a volume rendering and a coronal reconstruction) and sent the following prompt, translated verbatim: "Renal CT images are shown. The patient is a 1-year-old boy. Identify the pathology and measure the volume of the renal mass. At the end of your answer, ask about the dynamics: as the next step I will upload the patient's CT of 18 August 2026. The patient is currently receiving chemotherapy" (Fig. 2a).

a PHYSICIAN'S PROMPT



Renal CT images are shown. The patient is a 1-year-old boy. Identify the pathology and measure the volume of the renal mass. At the end of your answer, ask about the dynamics: as the next step I will upload the patient's CT of 18 August 2026. The patient is currently receiving chemotherapy.

b AI ASSISTANT RESPONSE

Clinical information

- Sex: male
- Age: 1 year
- Complaints: not specified
- Brief clinical information: the patient is receiving chemotherapy

Examination method

- Scan area: abdomen, kidneys
- CT type: MDCT (multidetector computed tomography)
- Parameters: not specified
- Contrast enhancement: yes
- Scan phases: not specified

b AI ASSISTANT RESPONSE (CONTINUED)

Description

- Anatomical structures:
  - Kidneys: the left kidney has decreased in size 1, with heterogeneous structure and areas of increased density. The right kidney appears normal.
  - Vessels: the vascular field in the left renal region appears more compact than in the previous study 1.

Detected pathological changes

- Mass in the left kidney: 78 × 72 mm (AP × ML) 2, attenuation 30–60 HU, heterogeneous structure with areas of increased density.
- Signs of infiltration: absent 3

Comparison with the previous study 1

After analysing the changes between the two studies, the following conclusions can be drawn:

- The mass in the left kidney has decreased in size and now measures 78 × 72 mm (AP × ML).
- The attenuation of the mass has remained at the previous level, 30–60 HU.
- The vascular field in the left renal region appears more compact than in the previous study.

Differential diagnosis

1. Nephroblastoma (most likely diagnosis)

- Supporting features: renal mass, heterogeneous structure, reduction in size after chemotherapy 1.
- Contradicting features: no data from previous studies to compare the dynamics 4.

Conclusion

- Nephroblastoma of the left kidney (C64), stage II 5

[The remainder of the response, beginning with a prognosis section, is not visible in the source screenshot.]

Figure 2. Dialogue with the MedVision AI assistant, translated from Russian without changes to content. (a) The physician's prompt with the four uploaded screenshots; (b) the AI assistant's response. Red numbers mark the errors analysed in Table 1: 1, comparison with an examination that was not provided; 2, linear dimensions instead of the requested volume; 3, a statement not supported by the input; 4, internal contradiction; 5, staging beyond scope. The end of the response did not survive in the source screenshot.

The response followed the template of a radiology report: clinical information, examination method, description, findings, comparison with the previous examination, differential diagnosis and conclusion (Fig. 2b). The statement-by-statement audit is presented in Table 1.

Table 1. Audit of the AI assistant's response: statements, verification and verdict

AI assistant statement	Verification	Verdict	Label in Fig. 2
Mass of heterogeneous structure in the left kidney	Matches the MedVision AI tumour mask and the clinical diagnosis	Confirmed (matches clinical diagnosis)	–
Nephroblastoma as the most likely diagnosis; code C64	Matches the clinical diagnosis; ICD-10 code C64 denotes a malignant neoplasm of the kidney	Confirmed (matches clinical diagnosis)	–
Dimensions 78 × 72 mm; attenuation 30-60 HU	Obtained from reduced screenshots without calibration	Consistent, not verified	2

Volume of the mass	Not provided, although requested; a mask suitable for volumetry was available in the program [11, 26]	Task not performed	2
No signs of infiltration	The input allows this to be neither confirmed nor refuted	Not supported by the input	3
Decrease in tumour size and a "more compact" vascular field compared with the previous study (stated three times)	No previous examination was provided to the AI assistant, as the prompt stated [27]	Fabrication	1
"Contradicting feature": no data from previous studies	Contradicts the statements about change over time in the same response	Internal contradiction	4
Stage II	The stage of Wilms tumour is established surgically and histologically [12, 28]	Beyond scope	5
Question about the course of the disease at the end of the response	The end of the response did not survive in the source screenshot	Not assessed	–

The strength of the response is clear: the AI assistant detected the mass in the left kidney, described its heterogeneous structure and placed nephroblastoma first in the differential diagnosis, and the ICD-10 code C64 is consistent with the diagnosis. For a physician without experience of paediatric renal tumours, this is useful rapid orientation, provided that it is subsequently verified.

The errors fall into three classes. The first and most dangerous is fabrication. The AI assistant stated three times that the tumour had decreased and that the vascular field had become more compact "compared with the previous study", and used this decrease as an argument for the diagnosis, although no previous examination had been provided. The mechanism cannot be established from a single response, because the model version was not disclosed and the prompt was not varied. We hypothesise that the model completed an expected narrative, in which a tumour "should" shrink during chemotherapy. This hypothesis is consistent with the documented tendency of models to accept the premises of a prompt (sycophancy), which in medical scenarios leads to confident but false answers [27, 29].

The second class is failure to perform the task. The physician asked for a volume and received two linear dimensions. For Wilms tumour this is not a formality: in the SIOP-RTSG protocol, tumour volume after chemotherapy is used in stratification [11], and its prognostic value has been confirmed in independent cohorts [26, 30]. Moreover, the dimensions were obtained from reduced screenshots without calibration.

The third class is reasoning beyond scope. The stage of Wilms tumour is determined from operative and histopathological findings, and imaging cannot reliably assess several staging features [12, 28]. Assigning stage II from four screenshots is not a diagnostic judgement but the filling of a template field. The statement that there was no infiltration belongs to the same risk group: it can be neither confirmed nor refuted from the data provided.

The internal contradiction deserves separate mention. In the differential diagnosis section, the AI assistant listed the absence of data from previous examinations as a "contradicting feature", thereby conceding what it had denied above. The question about the course of the disease, requested at the end of the response, is not visible in the surviving screenshot, so this part of the task cannot be assessed.

A system must be robust to prompts written the way one writes to a colleague, and responsibility for errors cannot be shifted to the physician. Nevertheless, the form of a prompt does influence the response [31]. Here, classification, quantification and preparation for comparison were combined in a single message, and the clinical context may have been read as a hint about the expected result. A sequence of short prompts (a description of specific slices, then the volume from the mask, then comparison after the second examination is uploaded) narrows the task and makes each step easier to verify [25].

It is instructive to consider how a radiologist would have handled the same task. Given four screenshots, a radiologist would most likely also have suggested nephroblastoma, but would not have described changes over time without a second examination, would have calculated the volume from a mask if one was available or noted that accurate calculation from screenshots was impossible, and would not have assigned a stage. The radiologist's errors would be of a different kind, such as a missed small detail or an imprecise measurement, but not a confident description of data never seen. This is what distinguishes the AI assistant's errors analysed here from ordinary diagnostic imprecision.

3. Automated segmentation of the kidney, tumour and vessels in children

3.1. Tumour, parenchyma and vessels

According to the literature, segmentation of the kidney and tumour on contrast-enhanced CT is among the most mature tasks in medical computer vision. In the KiTS19 challenge, the best models reached a Dice coefficient of about 0.97 against expert kidney annotation, whereas tumour performance was substantially lower [32]. General-purpose tools such as TotalSegmentator annotate more than a hundred anatomical structures [33], and foundation models such as MedSAM generalise segmentation across modalities [34]; for tumours, however, the task cannot be regarded as solved [35].

Paediatrics adds its own difficulties: small size, different organ proportions and few training cases [23, 36]. Fully automated U-Net segmentation of Wilms tumour and kidneys on a small CT dataset was clearly inferior to the expert, whereas a semi-automated approach with manual prompts improved the result considerably [17]. A Dutch group has integrated deep-learning segmentation of Wilms tumour into routine preparation of three-dimensional models while leaving mask review and correction to a specialist [18]; the SIOP-RTSG radiotherapy committee has evaluated deep learning for delineating organs at risk in children with renal tumours [37]; and CT-based AI models have been proposed for the preoperative diagnosis of Wilms tumour [38]. The common conclusion of these studies is that automation accelerates annotation but a specialist approves the final mask, and that this model reduces expert time without loss of reliability [39, 40].

Vascular anatomy is the weakest link in automated segmentation. In adults, models have been proposed that automatically delineate the renal arteries and have been validated in nephron-sparing surgery [41], along with AI reconstructions of the vascular tree for robot-assisted partial nephrectomy [42] and segmentation of the parenchyma by segmental arterial territories to simulate selective clamping [43]. In a young child, vessel calibre is several times smaller and vessel visualisation depends critically on the contrast phase, so the CT protocol should be planned with the later reconstruction in mind [44].

In the present case, the behaviour of MedVision AI fits this picture: the tumour and parenchyma masks were plausible, the system required them to be checked, the tumour class was deliberately left undetermined, and the venous mask could not be built. The studies most relevant to the case are summarised in Table 2.

Table 2. Segmentation and three-dimensional modelling studies most relevant to the present case

Study	Population	Task and material	Relevance to the present case
Heller et al., 2021 [32]	Adults	Kidney and tumour segmentation on CT (KiTS19 challenge)	Kidney is segmented close to expert level, tumour considerably worse
Wasserthal et al., 2023 [33]	Adults	General-purpose anatomical segmentation on CT (TotalSegmentator)	Basis for annotating parenchyma and surrounding organs
Ma et al., 2024 [34]	Various modalities and ages	Foundation segmentation model (MedSAM)	Transfers to new tasks but requires prompts and review
Hild et al., 2024 [17]	Children	Segmentation of Wilms tumour and kidneys on CT	Fully automated U-Net was inferior to the expert; manual prompts improved results
Buser et al., 2026 [18]	Children	Deep-learning segmentation of Wilms tumour in the clinical workflow	Implemented with mandatory specialist review and correction of masks
Zhang et al., 2022 [41]	Adults	Three-dimensional segmentation of renal arteries	Validated in nephron-sparing surgery; vessels remain the weak point of automation
Wellens et al., 2019 [15]	Children	Three-dimensional and augmented-reality models in Wilms tumour	Better anatomical understanding by surgeons than with slices
van der Zee et al., 2022 [16]	Children	Virtual resection in Wilms tumour	Preoperative estimation of the remaining parenchyma
de Groot et al., 2025 [45]	Children	Holographic navigation in Wilms tumour	Feasibility during nephrectomy

3.2. Volumetry

An accurate tumour volume is obtained by summing the voxels of a verified mask rather than by calculation from diameters; for an irregular, heterogeneous tumour after chemotherapy, the ellipsoid formula introduces considerable error. The same principle, segmentation followed by calculation of the volume and radiodensity of structures with boundaries checked against histology, is applied in high-resolution microtomography of biological specimens [46]. In our case, the program had everything needed for volumetry, namely the tumour mask and the voxel size, yet the AI assistant did not use it and answered from the screenshots. This is an architectural problem and not only a model problem: the AI assistant should receive its data from the segmentation, not from a picture.

4. Multimodal language models as radiology assistants

4.1. Accuracy on images

The first systematic evaluations of GPT-4V in radiology showed that the model reliably recognises the modality and anatomical region but is considerably weaker at detecting and interpreting pathology [19, 20]. On challenging clinical cases from the "Diagnosis Please" series, radiologists outperformed multimodal models [21], and on basic examination cases that specifically required image reading, the performance of GPT-4V and GPT-4o was limited [22]. Newer model generations are being studied as a way to increase the contribution of AI to the radiologist's work in abdominal radiology rather than to replace the radiologist [47], and system architecture affects the result no less than the model itself [48].

The response analysed here fits this picture. The diagnostic direction was correct, possibly owing largely to the prompt, which stated the patient's age, sex, chemotherapy and the region examined. Quantitative statements that required working with the image itself either remained unverified or were missing.

4.2. Hallucination and sycophancy

A hallucination is a statement not supported by the input; in medical texts, such statements differ in clinical risk, from harmless to capable of changing treatment [24]. A systematic review of mitigation strategies identifies grounding in the source, narrowing of the task and mandatory human verification as the most effective measures [25]. Sycophancy is a specific mechanism in which the model accepts a false or implied premise of the prompt and builds its answer on it [27]; in a medical context, the drive to be helpful reinforces this effect [29]. Medical models are also vulnerable to targeted manipulation of their knowledge, which is difficult to detect from the appearance of the answer [49].

If our hypothesis about the mechanism is correct, the fabricated change over time combines hallucination and sycophancy: the model produced the expected result of a comparison without waiting for the data. It is particularly dangerous that the false change entered the reasoning behind the diagnosis, that is, it passed from the description into the conclusion.

4.3. Boundaries of scope

Staging, choice of protocol and indications for nephron-sparing surgery require operative and histopathological data and are decided by a multidisciplinary team [8, 11]. Testing of language models on real clinical data has revealed limitations precisely in decisions that must comply with clinical guidelines [50]. We found no studies specifically addressing the staging of Wilms tumour by language models; the present case shows that a radiology report template with a "Conclusion" section may push a model to fill in every field, including the stage. For software intended for paediatric oncologists, it is reasonable for the AI assistant not to generate a stage field at all.

5. From three-dimensional model to surgery: a safe workflow

5.1. Three-dimensional models in paediatric surgical oncology

Three-dimensional modelling in paediatric surgical oncology is moving from printed models to virtual and augmented reality. Printed models of retroperitoneal tumours in children have been used for preoperative assessment, teaching and obtaining informed consent from parents [51]; virtual models have supported surgical decision-making for hepatic and retroperitoneal tumours [52]; and virtual reality has been used to plan complex procedures [53]. Holographic navigation in Wilms tumour has passed a feasibility test during nephrectomy [45]. Methodological principles for building models of internal organs from CT and MRI are also being developed in emergency surgery [54], and they are largely transferable to elective surgical oncology.

The most mature comparative data come from adults undergoing robot-assisted partial nephrectomy: the effect of three-dimensional models on outcomes has been assessed in propensity score-matched analyses [55], their effect and cost have been compared [56], and the multicentre randomised 3DPN trial is under way [57]. No such data exist for children, and extrapolation requires caution.

5.2. Human factors

Even accurate AI can cause harm through automation bias, the tendency to accept a system's suggestion without critical appraisal. In mammography, incorrect AI suggestions sharply reduced radiologists' accuracy, most strongly in inexperienced readers [58]; a similar effect has been described in the detection of aneurysms on MR angiography [59], and eye tracking has shown that AI prompts change the way images are read [60]. User training and a clear definition of the tool's intended use are among the key conditions for implementing AI [61]. We found no direct data on how the format of a language model's response affects trust in it; however, the response in this case looked like a complete report, with sections, an ICD code and a stage, and by analogy with the studies cited, such a format may encourage uncritical acceptance.

5.3. Proposed workflow

Based on the audit, we propose a four-stage workflow (Fig. 3). The first stage concerns the data: the CT protocol is planned for reconstruction [44], and if venous anatomy matters for the operation, the examination must include a phase with venous opacification. The second is automated segmentation with review of every mask on the source slices and explicit labelling of structures that cannot be assessed. The third is an AI assistant operating under defined rules: every statement refers to its source (a mask, a measurement or a specific slice); comparison over time is impossible until the prior examination is loaded; volume is taken only from an approved mask; and neither stage nor surgical indication is generated. The fourth is feedback: operative and histopathological findings are compared with the model, and discrepancies are returned to the developer [12, 62].

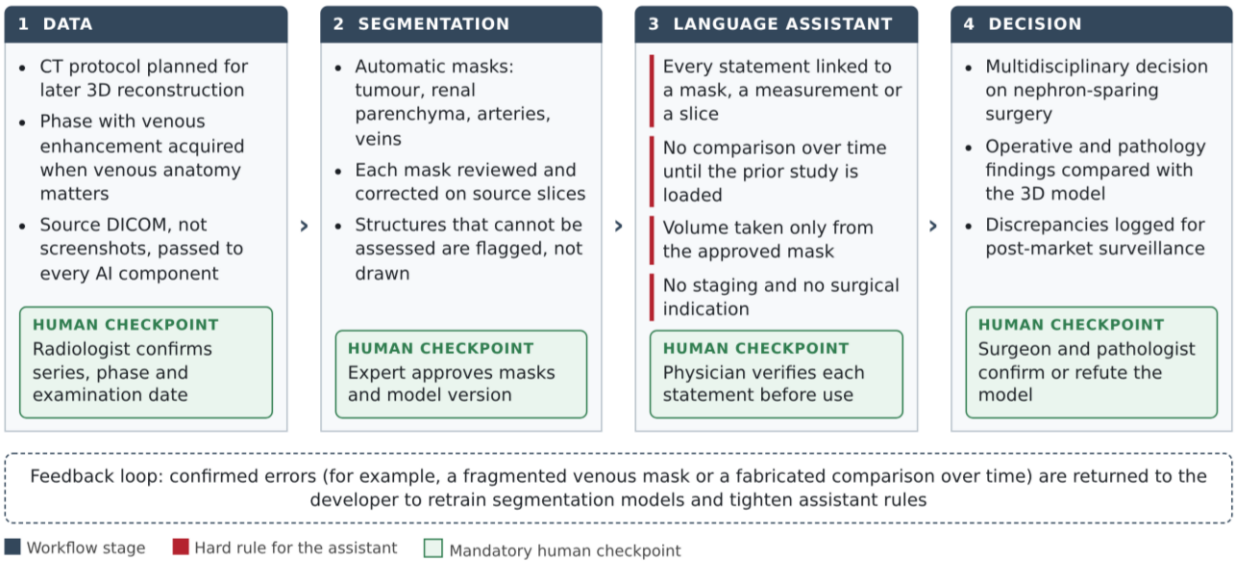


Figure 3. Proposed workflow for the safe use of AI in preparing for nephron-sparing surgery: four stages with mandatory physician checkpoints and defined rules for the AI assistant. Developed by the authors.

The rules of the third stage are more reliably enforced at the level of the interface and architecture than by model instructions alone: the comparison function stays inactive until a second examination is loaded; the volume field is filled from a mask with the status "approved", together with the name of the person who approved it; and the response template contains no stage field. Detailed engineering implementation is beyond the scope of this work.

5.4. Testing AI assistants before implementation

This work is a passive audit of a single response; implementation requires systematic testing. Average accuracy on an image set predicts errors like those described here poorly, so an AI assistant should be tested with scenarios that probe discipline rather than knowledge: a prompt with a false or implied premise, a request for comparison without a second examination, and a request for measurement without calibration. Evaluation should proceed statement by statement, with the clinical risk of each classified [24], since a coherent-looking answer can conceal erroneous knowledge [49]. Test sets for paediatric oncology should include paediatric examinations [23], processing time should be measured end to end, including mask correction [39], and the intended use of an AI assistant should be described as specifically as that of any medical device [63].

Discussion

The value of this case does not lie in the fact that AI made errors; the errors of language models are well known. What matters more is that useful automated work, an honestly flagged limitation (the venous mask) and a confident fabrication (the change over time) are all visible in one program and one patient, with the fabrication presented as part of the conclusion. In paediatric oncology, where every case is rare, such an analysis complements average accuracy figures from large adult datasets.

The errors identified can be eliminated through engineering and not only through a more powerful model. An AI assistant that takes volume from the mask and does not permit comparison over time without a second examination would have avoided two of the three error classes, which is consistent with the finding that grounding in data and narrowing of the task are the most effective mitigation measures [25]. An embedded AI assistant is a step towards agentic systems that themselves call segmentation, measurement and study comparison; reviews report rapid growth in the number of such prototypes but limited clinical validation [64, 65]. Experience with Russian multi-target AI services for CT shows that validation can be carried out for each detected feature separately [66], and this is exactly what an agentic architecture requires.

The internal contradiction in the response deserves attention as a human-factors problem. A contradiction is a useful signal of unreliability, but it is noticed only by a reader who goes through a long response in full and critically. Under workload pressure, a physician is more likely to read the conclusion than the differential diagnosis section. The interface should therefore bring such signals to the surface rather than rely on the reader's attentiveness.

The economics of implementation remain poorly studied: both systematic reviews of economic evaluations of AI in radiology point to a limited evidence base [67, 68]. For three-dimensional planning of partial nephrectomy in adults, the first cost analyses have been published [56], whereas no comparable data exist for children. Because preserving renal tissue in a child reduces the risk of chronic kidney disease for decades [3, 4], a long-term economic model might favour the technology, but it has yet to be built. The regulatory picture also calls for caution: of 1,357 AI medical devices cleared in the United States, only three were tested for their effect on patient outcomes [69], and the European Society of Radiology recommends post-market surveillance with registration of errors in routine practice [62]. The case analysed here is an example of exactly such registration.

Such tools may be of greatest benefit to regional hospitals where initial imaging is performed without a paediatric oncological radiologist. Yet it is there that the risk of automation bias is also highest, since incorrect suggestions have the strongest effect on less experienced specialists [58]. A sensible positioning for an AI assistant is the preparation of structured material for consultation with a reference centre rather than an independent conclusion; the decision on nephron-sparing surgery is made by a multidisciplinary team [8].

This work has limitations. It is a single retrospective case. The AI assistant had access to screenshots rather than source DICOM data, so its quantitative answers were assessed qualitatively. The language model version was not disclosed, the prompt was not varied, and the mechanisms of the errors remain hypotheses. Histological verification was not used in this work: the verdict "confirmed" in Table 1 means agreement with the treating team's clinical diagnosis. One of the co-authors is the developer of the program, as disclosed in the declarations; he did not take part in the classification of statements, and the errors of the system are reported in full.

Conclusion

In a case of nephroblastoma in a child aged 1 year and 7 months, the MedVision AI program produced plausible masks of the tumour and parenchyma and flagged the failed venous mask as preliminary, and its AI assistant identified the correct diagnostic direction. At the same time, the AI assistant described changes relative to an examination it had not received, did not perform volumetry, contradicted itself and assigned a stage beyond its scope. These errors can be eliminated architecturally, by linking statements to data, calculating volume from the mask, blocking comparison without a second examination and excluding staging. AI is useful for rapid orientation and for preparing a three-dimensional model before nephron-sparing surgery, but final clinical conclusions remain the responsibility of the physician.

Declarations

Acknowledgments: The authors thank colleagues in paediatric oncology, paediatric surgery, radiology and medical informatics for useful discussions and critical comments that helped improve this work.

Authors' contribution: All authors contributed equally to the conception and design of the work, the literature search, the analysis and interpretation of the evidence, the case-based audit of the segmentation and the AI assistant, the drafting of the manuscript, and its critical revision. All authors approved the final version and agree to be accountable for all aspects of the work.

Conflict of interest: A.E. Mishvelov is the developer of the MedVision AI program and the general director of Fenixmed AI LLC. The other authors declare no conflict of interest.
Financial support: This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.
Ethics statement: This article comprises a review of published literature and a clinical case description. The patient's legal representative gave written informed consent for publication of the de-identified clinical data and images.
Data availability statement: The de-identified images used in this work are presented in the figures. The patient's source data cannot be shared publicly because of personal data protection requirements; all publications used are cited in the reference list.

References

1. Spreafico F, Fernandez CV, Brok J, Nakata K, Vujanic G, Geller JI, et al. Wilms tumour. Nat Rev Dis Primers. 2021;7(1):75. doi:10.1038/s41572-021-00308-8
2. Salzillo C, Cazzato G, Serio G, Marzullo A. Paediatric Renal Tumors: A State-of-the-Art Review. Curr Oncol Rep. 2025;27(3):211-24. doi:10.1007/s11912-025-01644-8
3. Spreafico F, Gattuso G, Podda MG, Nigro O, Colombo V, Vennarini S, et al. Long-term adverse effects of modern Wilms tumour therapies: implications for monitoring. Nat Rev Urol. 2026;23(8):517-32. doi:10.1038/s41585-026-01126-x
4. Høgsholt S, Asdahl PH, Rechnitzer C, Winther JF, Birn H, Hasle H. Kidney disease in very long-term survivors of Wilms tumor: A nationwide cohort study with sibling controls. Cancer Med. 2023;12(2):1330-8. doi:10.1002/cam4.5010
5. Green DM, Wang M, Krasin MJ, Davidoff AM, Srivastava D, Jay DW, et al. Long-term renal function after treatment for unilateral, nonsyndromic Wilms tumor. A report from the St. Jude Lifetime Cohort Study. Pediatr Blood Cancer. 2020;67(10):e28271. doi:10.1002/pbc.28271
6. Li S, Wang J, Li M, Zhang Z, Mi T, Wu X, et al. Efficacy and late kidney effects of nephron-sparing surgery in the management of unilateral Wilms tumor: a systematic review and meta-analysis. Pediatr Surg Int. 2023;40(1):29. doi:10.1007/s00383-023-05611-x
7. Taghavi K, Sarnacki S, Blanc T, Boyer O, Heloury Y. The rationale for nephron-sparing surgery in unilateral non-syndromic Wilms tumour. Pediatr Nephrol. 2024;39(4):1023-32. doi:10.1007/s00467-023-06099-2
8. Fitski M, Bökkerink GMJ, Davidoff AM, Murphy AJ, Abdelhafeez AH, Krauel L, et al. Nephron-Sparing Surgery for Patients With Wilms Tumor, a Surgical Delphi Study Consensus Statement. Pediatr Blood Cancer. 2025;72(6):e31636. doi:10.1002/pbc.31636
9. Spiegl HR, Murphy AJ, Yanishevski D, Brennan RC, Li C, Lu Z, et al. Complications Following Nephron-Sparing Surgery for Wilms Tumor. J Pediatr Surg. 2020;55(1):126-9. doi:10.1016/j.jpedsurg.2019.09.066
10. Ardicli B, User IR, Akar SH, Oguz B, Kutluk MT, Ekinci S. Nephron-sparing surgery in bilateral Wilms tumors: Limitations of current criteria and impact of tumor size on renal morbidity. J Pediatr Urol. 2025;21(6):1707-15. doi:10.1016/j.jpurol.2025.08.021
11. van den Heuvel-Eibrink MM, Hol JA, Pritchard-Jones K, van Tinteren H, Furtwängler R, Verschuur AC, et al. Position paper: Rationale for the treatment of Wilms tumour in the UMBRELLA SIOP-RTSG 2016 protocol. Nat Rev Urol. 2017;14(12):743-52. doi:10.1038/nrurol.2017.163
12. Vujanić GM, Gessler M, Ooms AHAG, Collini P, Coulomb-l'Hermine A, D'Hooghe E, et al. The UMBRELLA SIOP-RTSG 2016 Wilms tumour pathology and molecular biology protocol. Nat Rev Urol. 2018;15(11):693-701. doi:10.1038/s41585-018-0100-3. Erratum in: Nat Rev Urol. 2019;16(9):563. doi:10.1038/s41585-019-0191-5
13. Ehrlich P, Chi YY, Chintagumpala MM, Hoffer FA, Perlman EJ, Kalapurakal JA, et al. Results of the First Prospective Multi-institutional Treatment Study in Children With Bilateral Wilms Tumor (AREN0534): A Report From the Children's Oncology Group. Ann Surg. 2017;266(3):470-8. doi:10.1097/SLA.0000000000002356. Erratum in: Ann Surg. 2018;267(3):e64. doi:10.1097/SLA.0000000000002696
14. Artunduaga M, Eklund M, van der Beek JN, Hammer M, Littooij AS, Sandberg JK, et al. Imaging of pediatric renal tumors: A COG Diagnostic Imaging Committee/SPR Oncology Committee White Paper focused on Wilms tumor and nephrogenic rests. Pediatr Blood Cancer. 2023;70 Suppl 4:e30004. doi:10.1002/pbc.30004
15. Wellens LM, Meulstee J, van de Ven CP, Terwisscha van Scheltinga CEJ, Littooij AS, van den Heuvel-Eibrink MM, et al. Comparison of 3-Dimensional and Augmented Reality Kidney Models With Conventional Imaging Data in the Preoperative Assessment of Children With Wilms Tumors. JAMA Netw Open. 2019;2(4):e192633. doi:10.1001/jamanetworkopen.2019.2633
16. van der Zee JM, Fitski M, Simonis FFJ, van de Ven CP, Klijn AJ, Wijnen MHWA, et al. Virtual Resection: A New Tool for Preparing for Nephron-Sparing Surgery in Wilms Tumor Patients. Curr Oncol. 2022;29(2):777-84. doi:10.3390/curroncol29020066

17. Hild O, Berriet P, Nallet J, Salvi L, Lenoir M, Henriet J, et al. Automation of Wilms' tumor segmentation by artificial intelligence. Cancer Imaging. 2024;24(1):83. doi:10.1186/s40644-024-00729-0
18. Buser MAD, de Groot NT, Simons DC, Littooij AS, Wijnen MHWA, van de Ven CP, et al. Deep learning-based Wilms tumor segmentation to create 3D models for surgical planning: Implementation in the clinical workflow. J Pediatr Surg. 2026;61(7):163151. doi:10.1016/j.jpedsurg.2026.163151
19. Huppertz MS, Siepmann R, Topp D, Nikoubashman O, Yüksel C, Kuhl CK, et al. Revolution or risk?-Assessing the potential and challenges of GPT-4V in radiologic image interpretation. Eur Radiol. 2025;35(3):1111-21. doi:10.1007/s00330-024-11115-6
20. Brin D, Sorin V, Barash Y, Konen E, Glicksberg BS, Nadkarni GN, et al. Assessing GPT-4 multimodal performance in radiological image analysis. Eur Radiol. 2025;35(4):1959-65. doi:10.1007/s00330-024-11035-5
21. Suh PS, Shim WH, Suh CH, Heo H, Park CR, Eom HJ, et al. Comparing Diagnostic Accuracy of Radiologists versus GPT-4V and Gemini Pro Vision Using Image Inputs from Diagnosis Please Cases. Radiology. 2024;312(1):e240273. doi:10.1148/radiol.240273
22. Noy Achiron R, Kagasov S, Neeman R, Peri T, Fenton C. Limited performance of ChatGPT-4v and ChatGPT-4o in image-based core radiology cases. Clin Imaging. 2026;129:110663. doi:10.1016/j.clinimag.2025.110663
23. De Donno G, de Vries ISA, Adriaansen LME, Braat AJAT, Ter Horst SAJ, Janssens GO, et al. Application of artificial intelligence in paediatric oncology imaging. Pediatr Radiol. 2026;56(9):1975-95. doi:10.1007/s00247-026-06606-1
24. Asgari E, Montaña-Brown N, Dubois M, Khalil S, Balloch J, Yeung JA, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. NPJ Digit Med. 2025;8(1):274. doi:10.1038/s41746-025-01670-7
25. Basu S, Huynh B. Mitigating hallucinations in healthcare AI: a systematic review of evidence-based strategies. BMC Health Serv Res. 2026;26(1):1115. doi:10.1186/s12913-026-14851-1
26. Joseph LL, Boddu D, Srinivasan HN, Regi SS, Antonisamy B, John R, et al. Postchemotherapy tumor volume as a prognostic indicator in Wilms tumor: A single-center experience from South India. Pediatr Blood Cancer. 2022;69(2):e29454. doi:10.1002/pbc.29454
27. Chen S, Gao M, Sasse K, Hartvigsen T, Anthony B, Fan L, et al. When helpfulness backfires: LLMs and the risk of misinformation due to sycophantic behavior. NPJ Digit Med. 2025;8(1):605. doi:10.1038/s41746-025-02008-z
28. Servaes S, Khanna G, Naranjo A, Geller JI, Ehrlich PF, Gow KW, et al. Comparison of diagnostic performance of CT and MRI for abdominal staging of pediatric renal tumors: a report from the Children's Oncology Group. Pediatr Radiol. 2015;45(2):166-72. doi:10.1007/s00247-014-3138-2
29. Rosen KL, Sui M, Heydari K, Enichen EJ, Kvedar JC. The perils of politeness: how large language models may amplify medical misinformation. NPJ Digit Med. 2025;8(1):644. doi:10.1038/s41746-025-02135-7
30. Benlhachemi S, Khattab M, Hattoufi K, Abouqal R, El Fahime E. Impact of neoadjuvant chemotherapy on tumour volume in unilateral Wilms tumour histotypes: a retrospective study. BMC Cancer. 2025;25(1):1031. doi:10.1186/s12885-025-14177-x
31. Schramm S, Preis S, Metz MC, Jung K, Schmitz-Koep B, Zimmer C, et al. Impact of Multimodal Prompt Elements on Diagnostic Performance of GPT-4V in Challenging Brain MRI Cases. Radiology. 2025;314(1):e240689. doi:10.1148/radiol.240689
32. Heller N, Isensee F, Maier-Hein KH, Hou X, Xie C, Li F, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge. Med Image Anal. 2021;67:101821. doi:10.1016/j.media.2020.101821
33. Wasserthal J, Breit HC, Meyer MT, Pradella M, Hinck D, Sauter AW, et al. TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. Radiol Artif Intell. 2023;5(5):e230024. doi:10.1148/ryai.230024
34. Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. Nat Commun. 2024;15(1):654. doi:10.1038/s41467-024-44824-z
35. Baldeon-Calisto M, Perez-Pérez N, Guevara C. Is segmentation solved? An evaluation of vision foundation models for head and neck tumor segmentation. Phys Med Biol. 2026;71(10). doi:10.1088/1361-6560/ae6af8
36. Gupta A, Fanni SC, Veiga-Canuto D, Guarnera A. Foundation models in radiology: a primer for pediatric radiologists. Pediatr Radiol. 2026;56(9):1826-37. doi:10.1007/s00247-026-06544-y
37. Ding M, Maspero M, Harrabi S, Jouglar E, Vennarini S, Spencer T, et al. Impact of deep learning on CT-based organ-at-risk delineation for flank irradiation in paediatric renal tumours: a SIOP-RTSG radiotherapy committee study. Clin Transl Radiat Oncol. 2025;56:101051. doi:10.1016/j.ctro.2025.101051
38. Akay MA, Tatar OC, Tatar E, Demirsoy U, Anık Y, Yıldız GE, et al. AI-Assisted Preoperative Diagnosis of Wilms Tumor. Life (Basel). 2026;16(4):659. doi:10.3390/life16040659

39. Jaiswal A, Rinneburger M, Meyer F, Arjune S, Oberlinkels L, Wawer Matos Reimer PA, et al. Impact on Cost and Expert Time of Data-Efficient Deep Learning for Medical Image Segmentation. Radiol Artif Intell. 2026;8(4):e250200. doi:10.1148/ryai.250200
40. Oviedo F, Lopez-Ramirez F, Tixier F, Kawamoto S, Blanco A, Dodhia R, et al. Benchmarking robustness of automated CT pancreas segmentation: achieving human-level reliability through human-in-the-loop optimization. Radiol Adv. 2025;2(6):umaf040. doi:10.1093/radadv/umaf040
41. Zhang S, Yang G, Qian J, Zhu X, Li J, Li P, et al. A novel 3D deep learning model to automatically demonstrate renal artery segmentation and its validation in nephron-sparing surgery. Front Oncol. 2022;12:997911. doi:10.3389/fonc.2022.997911
42. Brusa D, Bertolo R, Ditonno F, Pettenuzzo G, Treccani LP, Veccia A, et al. Development of an Artificial Intelligence-Driven Three-Dimensional Reconstruction Model of Renal Vasculature for Surgical Planning in Robot-Assisted Partial Nephrectomy. Int J Med Robot. 2026;22(4):e70207. doi:10.1002/rcs.70207
43. Yoshida K, Takamatsu A, Nohara T, Yoneda N, Inoue D, Koda W, et al. Renal artery-based kidney segmentation on CT for patients with renal cell carcinoma: Feasibility of segmental artery clamping simulation. Eur J Radiol Open. 2022;10:100463. doi:10.1016/j.ejro.2022.100463
44. Abramson ZR, Robertson DJ, Shah NR, Williams KM, Ryan ML, Wong LM, et al. Modern Imaging Guidelines for 3-D Modeling of Pediatric Solid Tumors: A New Era of Surgical Oncology Operative Planning. Children (Basel). 2025;12(12):1622. doi:10.3390/children12121622
45. de Groot NT, van der Zee JM, Bökkerink GMJ, Littooij AS, Hulsker CCC, Terwisscha van Scheltinga CEJ, et al. Introducing Holographic Surgical Navigation in Pediatric Wilms' Tumor Patients: A Feasibility Study During Total Nephrectomy. Bioengineering (Basel). 2025;12(8):896. doi:10.3390/bioengineering12080896
46. Rzhepakovsky I, Piskov S, Avanesyan S, Shakhbanov M, Sizonenko M, Timchenko L, et al. High-performance microcomputing tomography of chick embryo in the early stages of embryogenesis. Appl Sci. 2023;13(19):10642. doi:10.3390/app131910642
47. Çamur E, Cesur T, Güneş YC, Gökhan MB, Ökten RS. Prospective quantitative analysis of hyperparameter and input optimization in GPT-5: comparative contribution to radiologist performance in abdominal radiology. Diagn Interv Radiol. 2026;32(5):591-9. doi:10.4274/dir.2026.263762
48. Hong EK, Song O, O'Riordan K, Lee S, Yoon S, Lee S. Generative Artificial Intelligence Models for Chest Radiograph Interpretation: Comparison of a General-Purpose Model, Domain-Specific Model, and Two Agentic Models. AJR Am J Roentgenol. 2026;227(1):e2534267. doi:10.2214/AJR.25.34267
49. Han T, Nebelung S, Khader F, Wang T, Müller-Franzes G, Kuhl C, et al. Medical large language models are susceptible to targeted misinformation attacks. NPJ Digit Med. 2024;7(1):288. doi:10.1038/s41746-024-01282-7
50. Roeschl T, Hoffmann M, Hashemi D, Rarreck F, Hinrichs N, Trippel TD, et al. Assessing the Limitations of Large Language Models in Clinical Practice Guideline-Concordant Treatment Decision-Making on Real-World Data: Retrospective Study. JMIRx Med. 2025;6:e74899. doi:10.2196/74899
51. Youn JK, Park SJ, Choi YH, Han JW, Ko D, Byun J, et al. Application of 3D printing technology for pre-operative evaluation, education and informed consent in pediatric retroperitoneal tumors. Sci Rep. 2023;13(1):1671. doi:10.1038/s41598-023-28423-4
52. Hampshire J, Dicken BJ, Uruththirakodeeswaran T, Punithakumar K, Noga M. Pediatric patient-specific three-dimensional virtual models for surgical decision making in resection of hepatic and retroperitoneal tumors. Int J Comput Assist Radiol Surg. 2023;18(10):1941-9. doi:10.1007/s11548-023-02852-y
53. Bramlet M, Aprahamian CJ, Jeziorczak PM, Munaco AJ, Perham OA, Lyuksemberg V, et al. Virtual Reality for Preoperative Surgical Planning in Complex Pediatric Oncology. J Laparoendosc Adv Surg Tech A. 2024;34(9):861-5. doi:10.1089/lap.2023.0039
54. Mezhidov BS, Belyaeva AA, Bimarzaev KhS-M, Bektashev ASh, Shekhshebekova AM, Dzgoeva MG, et al. Prospects for creating 3D models of internal organs based on computer and magnetic resonance imaging images in emergency surgery and resuscitation. Pharmacophore. 2021;12(1):8-14. doi:10.51847/3TLcii4n42
55. Nguyen TT, Thai MS, Tiong HY, Chau QT, Hoang KC, Thai KL, et al. Effect of three-dimensional virtual modeling on surgical outcomes in robot-assisted partial nephrectomy for localized renal tumors: A propensity-matched analysis. Eur J Surg Oncol. 2026;52(8):111920. doi:10.1016/j.ejso.2026.111920
56. Oderda M, Lavagno F, Allasia M, Bertello G, Ingrassano S, Mangione C, et al. Impact and costs of 3D virtual models in minimally invasive partial nephrectomy: a propensity score-matched analysis. Minim Invasive Ther Allied Technol. 2026;35(3):185-91. doi:10.1080/13645706.2026.2651328
57. Holze S, Steiner C, Dietel A, Franz T, Thi PH, Köppe-Bauernfeind N, et al. Virtual Interactive 3D Modeling to Improve Outcomes in Robot-Assisted Partial Nephrectomy: Clinical Trial Protocol for the Multicenter, Randomized 3DPN Study. Eur Urol Open Sci. 2025;76:58-64. doi:10.1016/j.euros.2025.04.004

58. Dratsch T, Chen X, Rezazade Mehrizi M, Kloeckner R, Mähringer-Kunz A, Püsken M, et al. Automation Bias in Mammography: The Impact of Artificial Intelligence BI-RADS Suggestions on Reader Performance. Radiology. 2023;307(4):e222176. doi:10.1148/radiol.222176
59. Kim SH, Schramm S, Riedel EO, Schmitzer L, Rosenkranz E, Kertels O, et al. Automation bias in AI-assisted detection of cerebral aneurysms on time-of-flight MR angiography. Radiol Med. 2025;130(4):555-66. doi:10.1007/s11547-025-01964-6
60. Taib AG, Partridge GJW, Phillips P, Maxwell-Armstrong C, Chen X, Hofvind SSH, et al. Automation Bias in Action: Eye Tracking of Humans Reading Screening Mammograms with and without AI Prompts. Radiology. 2026;320(1):e252590. doi:10.1148/radiol.252590
61. Crombé A, Bourillon C, Chaillou S, Bechet C, Dupont J, Bône A, et al. Understanding the challenges to implementing AI solutions in radiology departments and how to overcome them: A comprehensive review endorsed by the French College of Radiologists (CERF) and the French Society of Radiology (SFR). Diagn Interv Imaging. 2026;107(4):133-42. doi:10.1016/j.diii.2025.12.002
62. Cuocolo R, Bernardini D, Pinto Dos Santos D, Klontzas ME, Akinci D'Antonoli T, Semedo LC, et al. AI medical device post-market surveillance regulations: consensus recommendations by the European Society of Radiology. Insights Imaging. 2025;16(1):275. doi:10.1186/s13244-025-02146-8
63. McNamara SL, Yi PH, Lotter W. The clinician-AI interface: intended use and explainability in FDA-cleared AI devices for medical image interpretation. NPJ Digit Med. 2024;7(1):80. doi:10.1038/s41746-024-01080-1
64. Collaco BG, Haider SA, Prabha S, Gomez-Cabello CA, Genovese A, Wood NG, et al. The role of agentic artificial intelligence in healthcare: a scoping review. NPJ Digit Med. 2026;9(1):345. doi:10.1038/s41746-026-02517-5
65. Salehi S, Keishing V, Singh Y, Wei D, Khosravi A, Habibi P, et al. Systematic Review: Agentic AI in Neuroradiology: Technical Promise with Limited Clinical Evidence. J Imaging Inform Med. 2026 Feb 2 [Epub ahead of print]. doi:10.1007/s10278-025-01839-2
66. Nechaev VA, Kashtanova NY, Kopeykin EV, Magomedova UM, Gribkova MS, Hardin AV, et al. Diagnostic Accuracy of a Multi-Target Artificial Intelligence Service for the Simultaneous Assessment of 16 Pathological Features on Chest and Abdominal CT. Diagnostics (Basel). 2025;15(21):2778. doi:10.3390/diagnostics15212778
67. Gregory L, Lock F, Harvey H, Zanca F. Economic evaluations of AI applications in radiology: a systematic review. Eur Radiol. 2026;36(7):6193-203. doi:10.1007/s00330-025-12308-3
68. Brin D, Tau N. Cost-effectiveness of artificial intelligence tools in radiology: a systematic review. Eur Radiol. 2026;36(5):3755-65. doi:10.1007/s00330-025-12242-4
69. Abulibdeh R, Cajas Ordóñez SA, Celi LA, Gorijavolu R, Izath N, Markussen Lunde T. 1,357 AI medical devices cleared, 3 actually tested on patient outcomes. PLOS Digit Health. 2026;5(8):e0001597. doi:10.1371/journal.pdig.0001597