communications medicine

A Nature Portfolio journal

Article

https://doi.org/10.1038/s43856-025-01264-0

A clinically validated AI framework for kidney cancer detection and characterization

Bohdan Petryshak[1,2,3], Mikhail Iljin[2], Alina Denissova[2,4], Joonas Ariva[1,5], Toomas Häide[2], Veljo Lasn[2], Martin Reim[2,4], Ihor Ivaniv[2], Toomas Lillsaar[4], Hardi Vilt[4], Anu Leht[4], Anti Karask[4], Gitana Kiudma[4], Rait Käpp[4], Priit Salumaa[2], Pilvi Ilves[4] & Dmytro Fishman[1,2,5]

Abstract

Background Renal cell carcinoma is one of the most common cancers of the urinary tract and is usually diagnosed by interpreting contrast-enhanced computed tomography scans. Rising demand for radiological services, combined with a shortage of radiologists, makes timely and accurate diagnosis increasingly challenging. Automated approaches may help radiologists improve efficiency and accuracy.

Methods We developed BMVision, a deep learning-based tool for detecting and characterizing kidney cancer. The tool integrates with a web-based viewer designed to provide an intuitive interface for radiologists. Its performance was evaluated in a two-stage retrospective reader study. Six radiologists independently reviewed 200 scans across both AI-assisted and unaided workflows, allowing comparison of diagnostic performance and workflow efficiency with and without support from the tool. Statistical analysis compared AI-aided and unaided workflows across predefined clinical criteria, including diagnostic sensitivity, lesion measurement, reporting efficiency, and inter-radiologist agreement, using non-parametric tests and bootstrapping.

Results Here we show that BMVision reduces radiologists' reporting time by an average of 33%, up to 52%. The tool provides structured auto-generated reports, minimizing the need for manual dictation or typing. In addition, BMVision improves sensitivity for detecting benign renal lesions (from 79.9 to 86.3%) and leads to a significant increase in inter-radiologist agreement.

Conclusions To the best of our knowledge, BMVision is the first clinically validated commercial artificial intelligence tool for kidney cancer detection and characterization. By improving diagnostic accuracy and reporting efficiency, it has the potential to enhance patient care and help radiologists meet the growing demand for high-quality cancer diagnostics.

Plain language summary

Kidney cancer is one of the most common cancers of the urinary system. It is typically identified using specialized medical images called computed tomography (CT) scans, which are carefully reviewed by radiologists. However, there are not enough radiologists, and the demand for scans is growing. This makes it more challenging to provide patients with fast and accurate results. In this study, we developed a computer tool called BMVision, which utilizes artificial intelligence to analyze CT scans and assist radiologists in diagnosing kidney cancer. We tested BMVision with six radiologists, who reviewed a large number of scans with and without help from the tool. We found that BMVision enables radiologists to work more efficiently and consistently. Tools like BMVision can help patients by making cancer diagnosis faster, more reliable, and more widely available.

Renal cell carcinoma (RCC) represents around 3% of all cancers[1]. In 2020, there were an estimated 431,288 new cases of RCC globally[1]. Early detection of kidney cancer is critical for improving patient outcomes by enabling timely intervention and treatment. The primary diagnostic tool for kidney cancer is contrast-enhanced computed tomography (CT), which requires precise and efficient interpretation[2]. However, the increasing demand for radiological examinations[3–5], compounded by a shrinking workforce of radiologists[6], presents significant challenges to maintaining diagnostic accuracy and timely reporting.

Given these challenges, recent advances in deep learning have emerged as promising solutions for improving medical image analysis, particularly in the diagnosis of diseases across various imaging modalities[7–13]. Early models,

[1]Institute of Computer Science, Tartu University, Tartu, Estonia. [2]Better Medicine OÜ, Tallinn, Estonia. [3]Ukrainian Catholic University, Lviv, Ukraine. [4]Tartu University Hospital, Tartu, Estonia. [5]STACC OÜ, Tartu, Estonia. e-mail: dmytro.fishman@bettermedicine.ai

such as U-Net[7] focused on segmentation – that is, identifying pixels corresponding to specific anatomical structures or anomalies, but lacked interactive features and integration into clinical workflows. More recent models, such as the Segment Anything Model (SAM)[14] and its medical derivatives, including MedSAM[10], MedSAM-2[12], and ESP-MedSAM[11], introduced interactive segmentation through user-provided inputs, such as points, bounding boxes, and text prompts. These methods significantly improved segmentation accuracy, particularly in cases involving small or irregular lesions.

Building on this, nnInteractive[13] combines the robustness of nnU-Net[9] with SAM-style interaction to enable accurate 3D segmentations, and has been further extended to a 3D-Slicer plugin[15], making it accessible within a widely used medical imaging platform[16]. However, tools like nnInteractive primarily address universal lesion segmentation alone and stop short of providing end-to-end integration into the radiological workflow. They do not automate structured reporting or support full clinical reader studies, both of which are essential for real-world adoption in radiology departments that rely on certified viewers and standardized reporting systems.

To address these limitations, we introduce BMVision, a specialized AI framework for kidney cancer detection and characterization that goes beyond segmentation. While BMVision is based on the nnU-Net model[9], an advanced iteration of U-Net[7], it incorporates optimized postprocessing and characterization modules that are custom-built for the specific challenges of kidney cancer diagnosis. Additionally, BMVision includes a web-based viewer based on Open Health Imaging Foundation (OHIF) V3 web imaging platform for medical imaging, which enables a streamlined and user-friendly interaction with the tool, improving workflow efficiency for radiologists. BMVision is developed specifically for kidney cancer diagnostics and integrates segmentation, post-processing, characterization, and structured reporting within an OHIF-based viewer.

BMVision was validated in a two-stage retrospective reader study involving six radiologists using 200 CT scans from Tartu University Hospital (TUH). The dataset was carefully curated to minimize bias and ensure applicability to real-world clinical environments. Six radiologists reviewed 200 scans across both AI-assisted and unaided workflows, resulting in 2400 individual reads for comparative analysis. Statistical analysis compared these workflows across predefined clinical criteria, including diagnostic sensitivity, lesion measurement, reporting efficiency, and inter-radiologist agreement, using non-parametric tests and bootstrapping.

The study demonstrates that BMVision enhances radiologists' sensitivity in detecting malignant tumors, improves the detection of benign lesions, and reduces the time required to detect, measure, and report lesions by 33%. It also increases inter-radiologist agreement, supporting more consistent clinical decision-making. BMVision is a commercial AI framework specifically designed and clinically validated for the diagnosis of kidney cancer. The work offers three key contributions: the development of a dedicated AI tool for kidney cancer, its validation in a clinical workflow through a multi-radiologist reader study, and the creation of a large, carefully annotated dataset of kidney cancer cases paired with relevant controls. While the dataset is not publicly available due to patient privacy regulations, it plays a crucial role in demonstrating the model's real-world applicability. Together, these results demonstrate that AI assistance can enhance diagnostic accuracy, efficiency, and reliability, thereby advancing the standard of care for patients with kidney cancer.

Methods

BMVision

BMVision is built upon a three-module architecture: segmentation, post-processing, and characterization. At its core is a 3D U-Net architecture adapted from the nnU-Net model[9], which has been trained using a dataset of 612 CTs, a combination of the TUH dataset and publicly available data, including C4KC-KiTS[17,18] and TCGA-KiRC[19,20]. These datasets were carefully curated to provide diverse and representative training data, with the TUH dataset specifically assembled to minimize gender and age-related biases and to include both healthy and cancerous patients. The nnU-Net model was trained to perform four-class semantic segmentation, classifying each voxel into one of four categories: background, benign lesions, malignant lesions, or healthy kidney tissue. See Fig. 1 for more details on BMVision workflow and Supplementary Methods for implementation details.

The post-processing module refines the segmentation output by removing false positives from areas outside the kidney region and resolving inconsistencies in predictions within the anatomical boundaries of the kidney. The characterization module converts the post-processed segmentation mask into an instance segmentation map, distinguishing each individual object, such as the left and right kidney, malignant, and benign lesions. The instance segmentation map is then used to compute key metrics essential for radiologists, such as lesion size, volume in cubic millimeters, and the largest diameter.

Data overview

To develop BMVision kidney AI, we constructed a dataset comprising CT scans from three independent sources. The primary dataset includes 291 histology-proven renal cancer cases and 300 controls diagnosed with appendicitis, which were randomly selected from patients undergoing treatment at TUH between 2010 and 2020. During the data acquisition phase, 100 controls and 100 cancer cases were reserved for a test set. The remaining 391 CT scans were allocated to the development set. Potential confounding factors, such as age and gender, were mitigated by employing a stratified randomized split, ensuring balanced distributions of these variables between the development and test sets. This stratification also ensured an equal representation of controls and cases in both sets. To minimize potential confounding from appendicitis-related findings, the appendix region was removed from test CT scans, both controls and cancer cases, prior to model evaluation.

To enhance the heterogeneity of the training data, we included additional scans from two publicly available online databases: 41 cases from the

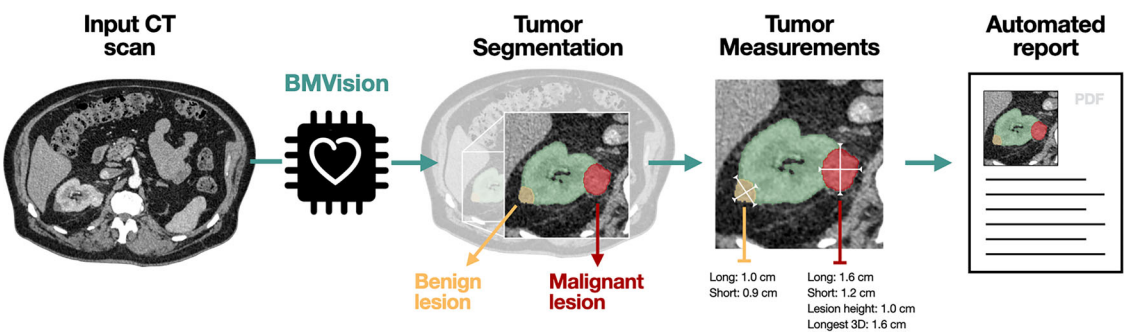

Fig. 1 | BMVision workflow. First, the model segments the scan into four classes: background, healthy kidney tissue, malignant and bening kidney lesions. Then the predictions are postprocessed and converted into indiviual instances, where key metrics are calculated. Finally, based on all the findings, automated report is generated, which the radiologist can use for reporting.

C4KC-KiTS dataset and 180 cases from the TCGA-KIRC dataset. These supplementary datasets provided diversity in terms of scanner manufacturers, acquisition protocols, and patient populations. In total, the training set comprised 612 CT volumes, while the test set contained 200 volumes. All CT scans were annotated in-house by a team of radiology residents and board-certified radiologists, with the details of the annotation process described in Section 5.

Both the development and test sets exclusively included contrast-enhanced CT volumes captured in the corticomedullary, nephrogenic, or portal-venous contrast phases. Only images reconstructed using a soft CT kernel were used. The slice thickness of the selected images ranged from 0.625 to 5.0 mm, and the peak kilo voltage ranged from 90 to 150.

As part of our data acquisition efforts, we have excluded pregnant individuals, subjects younger than 18 years, patients with anatomical kidney abnormalities, polycystic kidney disease, or severe hydronephrosis, as well as individuals with kidney transplants. Scans were also excluded if they contained only native or excretory contrast phases, or if they lacked full representation of the kidneys or any of the kidney-related benign or malignant lesions. All the aforementioned exclusion criteria were applied uniformly across all three sources to ensure dataset quality and consistency.

Pre-clinical evaluation

Before exposing BMVision to radiologists in the reader study, the framework was thoroughly validated internally. For its validation, BMVision was tested on a separate dataset of 200 CT scans from TUH. This dataset was specifically designed to be akin to real-world clinical settings where kidney tumors are rather rare. Therefore, half of the scans (100 CTs) were from subjects without kidney tumors, while the other half (100 CTs) represented patients with pathologically confirmed kidney tumors. Additionally, the dataset was carefully balanced by both age and gender. The subjects were equally distributed, with 100 male and 100 female participants, and there was an equal number of male and female patients within the cancer and non-cancer groups. Age matching was also performed between the groups to ensure that any differences in model performance could not be attributed to demographic discrepancies.

Following the evaluation on our internal dataset, we compared BMVision's performance to the top segmentation models from the 2021 Kidney and Kidney Tumor Segmentation Challenge (KiTS21), a popular benchmark for kidney tumor segmentation from CT scans[17,18]. Specifically, we evaluated BMVision's pixel-level DICE score, which measured its segmentation accuracy, and found it to be comparable with the top models from KiTS21, which achieved DICE scores in the range of 81%–86% for malignant lesions. While this comparison provides general context, we emphasize that potential differences in imaging protocols, lesion characteristics, and test populations limit direct comparison across datasets. Therefore, we include this reference as a rough point of comparison rather than a formal benchmark. In addition to the DICE score, commonly used in segmentation tasks, we focused on object-level sensitivity, a metric that more directly reflects the model's practical ability to detect malignant lesions. BMVision's object-level sensitivity for malignant lesions reached 93.4%, and its patient-level sensitivity was 96%, comparable to radiologist performance reported in the literature[21,22]. These results confirmed that BMVision's performance was on par with the top models in the field, supporting its readiness for the reader study.

Reader study

The primary objective of the reader study was to validate BMVision in the hands of radiologists. Specifically, the study aimed to evaluate the potential speed-up BMVision could offer in radiological reporting of renal cancer, assess frameworks' performance in sensitivity and specificity at both the object and patient levels when used in combination with radiologists, and determine whether the use of BMVision influenced inter-radiologist agreement.

To address these goals, the study was designed with two independent arms and employed a washout period to mitigate carryover (memorization) effects. Figure 2 presents the structure of the reader study. Extensive randomization was performed at every stage to eliminate potential biases.

In the first arm, radiologists followed the standard clinical workflow: interpreting CT images by manually identifying and measuring suspicious malignant and benign renal lesions, followed by manually typing a corresponding report. In the second arm, radiologists performed the same tasks but with the assistance of BMVision, which helped in detecting and measuring renal findings. Radiologists then used the model's refined measurements to semi-automatically generate a report based on the identified lesions.

To enable a precise assessment of the time required for detecting, measuring, and reporting malignant and benign lesions, each CT volume was analyzed in two stages: one for malignant lesions and another for benign lesions. In the malignant stage, radiologists measured the longest tumor diameter in 3D, followed by three orthogonal measurements—two in the axial plane and one in the coronal or sagittal plane. These measurements are essential for tumor staging using the TNM classification system[23] and for

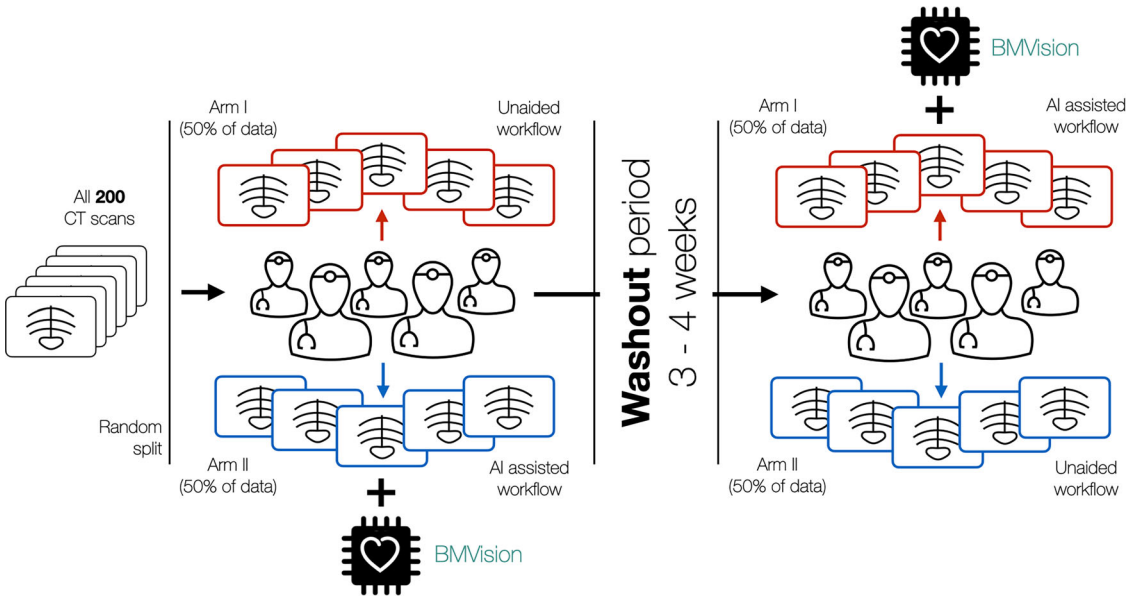

Fig. 2 | Overview of the reader study flow. During the first phase, every radiologist should report 50% of the test data following standard clinical practices and 50% with the assistance of the BMVision AI algorithm. After a 3–4 week washout period, AI-aided and non-aided test data points should be swapped, and the procedure should be repeated.

determining the nephrometry score[24]. In the benign stage, radiologists marked benign lesions and recorded two measurements in the axial plane for each lesion. In routine clinical practice, benign lesions are often assessed qualitatively, without detailed measurement, to evaluate their nature.

The study was conducted in two phases, separated by a 3-4-weeks washout period. In the first phase, radiologists analyzed and reported on 50 cases and 50 controls, both with and without the assistance of BMVision. The order in which CT scans were assigned to each radiologist, as well as whether scans would come with or without assistance, was randomized, and each radiologist's order was unique. During the washout period, radiologists continued with their regular duties to avoid any memory bias from the first phase. In the second phase, the cases from the first arm were swapped, meaning radiologists reported the same studies but under the opposite condition (i.e., with or without AI assistance). This fully crossed design maximized the utility of the dataset and ensured each radiologist reviewed every case under both conditions (see Fig. 2 for an overview).

A total of six practicing radiologists participated in the study, each with between 4 and 26 years of experience, covering various subspecialties including oncology, abdominal imaging, musculoskeletal radiology, and interventional radiology. To minimize learning bias and ensure consistency, the radiologists underwent a training session prior to the experiment, during which they familiarized themselves with the study protocol. Each radiologist reviewed ten training cases, both with and without BMVision assistance, to get comfortable with the viewer and the workflow. These training cases were selected from the development set and were not included in the final test set.

All radiologists in both arms of the study utilized the same custom-developed image viewer based on OHIF to eliminate any bias related to software interfaces and ensure that any observed differences in reporting time were attributable to the use of BMVision. The viewer was extended with a precise time-tracking system to measure the time spent on malignant, benign, and reporting tasks. This time-tracking system recorded every radiologist's interaction within the environment, enabling an accurate comparison of efficiency between AI-assisted and unaided reporting.

The primary endpoint of the study was the time required to complete a diagnostic interpretation, measured from the moment a CT series was received until a final diagnostic report was generated for the renal region. Secondary endpoints included the sensitivity of detecting kidney lesions from CT scans, sensitivity and specificity for detecting patients with malignant renal lesions, and the inter-radiologist agreement rates in identifying kidney lesions. These endpoints were calculated for both groups: radiologists assisted by BMVision and those following the standard of care without AI. Where relevant, BMVision's standalone performance, independent of the reader study, is also reported for comparison.

Reference standard

The annotation team, comprising five members, including board-certified radiologists and radiology residents, created high-quality annotations for the CT scans in the development and test sets. They used annotation tools integrated into the BMVision custom viewer and followed detailed annotation guidelines to ensure consistency and accuracy of resulting annotations. The annotation process consisted of two main stages: localization and classification of the lesions and lesion segmentation.

Initially, radiologists identified and classified all renal findings, including both benign and malignant lesions. Localization was performed using the "ruler" tool, which enables radiologists to draw visible lines on the scans, indicating the longest and shortest axes in the axial plane for each suspicious object. The "Benign lesion" category included Bosniak categories I, II, and IIF, while the "Malignant lesion" category encompassed Bosniak categories III–IV and histologically confirmed solid renal tumors. Annotations included all lesions, regardless of size, including small benign and malignant findings as small as 1–2 mm.

During the second stage, pre-segmentations generated by a continuously refined segmentation model were presented after the completion of the localization and classification. The radiologists verified that their initial markings accurately corresponded to the segmented objects produced by the model, ensuring consistency in classification and localization. Then, they were invited to enhance the pre-segmentations. During this stage, radiologists also segmented the kidney tissue itself, which included benign conditions, such as infarctions, inflammatory changes, and stones.

To ensure that the annotation team could make well-informed decisions and produce high-quality ground truth data, in addition to contrast-enhanced CT volumes, they were provided with native-phase CT images and, for cancerous samples, patient histological reports. The histological report included detailed descriptions of the morphology and location of the renal lesions. Native-phase images were essential for distinguishing enhancing lesions from cysts containing dense substances.

Each CT volume was annotated based on the consensus of at least two radiologists. The resulting segmentation masks were used as a reference standard. For malignant lesions, if both radiologists annotated an object as malignant, the union of their boundaries was used. If only one radiologist marked the object as malignant, the radiologists discussed the case and reached a consensus to finalize the annotation. For benign lesions and kidney tissue categories, the union of annotations from both radiologists was applied to create the final ground truth.

Statistics and reproducibility

All statistical analysis was performed to compare AI-aided and unaided radiologist workflows across multiple predefined criteria, including diagnostic sensitivity, segmentation accuracy, lesion measurement, reporting efficiency, and inter-radiologist agreement. Each test addressed a distinct clinical question. Therefore, we did not apply a correction for multiple comparisons, as recommended when testing non-overlapping hypotheses. For all hypothesis tests, we used a two-sided significance threshold of $\alpha = 0.05$. Statistical analysis was performed in Python and R using standard libraries.

To assess statistical differences between AI-aided and unaided workflows, we first tested all continuous variables for normality using the Shapiro-Wilk test. As most distributions were non-normal, we primarily used the Wilcoxon signed-rank test for paired comparisons. For patient-level and object-level sensitivity and specificity, we applied non-parametric bootstrapping to estimate confidence intervals and p-values, due to the limited number of paired samples. Object-level agreement (i.e., the proportion of lesions identified by all radiologists) was compared using the chi-squared test. Inter-radiologist agreement in lesion classification was quantified using Cohen's Kappa statistic; results were averaged across all reader pairs and compared using the Wilcoxon rank-sum test. Additional details on the Kappa analysis are provided in the Supplementary Methods.

Sample size calculations were performed using reporting time as the primary endpoint, defined as a continuous outcome. Assuming a conservative 20% reduction in reporting time with AI support, equal allocation between AI-assisted and unaided workflows, no drop-outs due to retrospective recruitment, $\alpha = 0.025$, and 80% power, we estimated that at least 266 scans would be required to detect a statistically significant difference. The present clinical investigation recruited 600 subjects, of which 391 were used for fine-tuning of the AI model. The remaining 200 subjects, comprising 100 kidney cancer cases and 100 age- and gender-matched controls, were included in the reader study. With a fully crossed design, each of the six radiologists reviewed all cases in both AI-assisted and unaided conditions, yielding 2400 individual reads. In this context, biological replicates correspond to patient cases, and technical replicates correspond to independent reads by different radiologists. This design ensures reproducibility of results and satisfies the required sample size for adequate statistical power.

Performance metrics

One of the primary objectives of this study was to evaluate how AI affects both the efficiency and accuracy of radiologists. To this end, we utilized the following metrics to capture reporting time, diagnostic performance, and inter-radiologist agreement.

Time measurement. Radiologists' task times were recorded using a custom time-tracking system built into the BMVision viewer. The system logged all

user interactions and measured active working time for each task. If no interaction was detected for more than 15 s, the timer paused until further activity resumed. This ensured that only active working time was recorded. Each CT volume was analyzed in two stages: a malignant stage, in which radiologists measured the longest tumor diameter in 3D along with three orthogonal measurements (two axial, one coronal or sagittal), and a benign stage, in which benign lesions were identified and measured in two axial planes.

Sensitivity and specificity. Diagnostic performance was evaluated using sensitivity and specificity at both the object and patient levels. Object-level sensitivity was defined as the proportion of lesions of a given class (benign or malignant) that were correctly identified. Patient-level sensitivity was defined as the proportion of patients with at least one lesion of a given class for whom at least one lesion of that class was correctly identified. Specificity was defined as the proportion of patients without lesions of a given class who were correctly identified as lesion-free. Patients with both benign and malignant lesions were included in both sensitivity calculations.

Inter-radiologist agreement. Consistency in radiological interpretation was assessed using three complementary measures: object-level agreement, disagreement rates, and Cohen's Kappa coefficient. Object-level agreement represents the proportion of lesions consistently identified by all six radiologists. Disagreement measures the variation in the number of lesions reported per scan relative to the group mean. Cohen's Kappa coefficient quantifies the pairwise agreement between radiologists beyond chance. Kappa values were averaged across all reader pairs. Details of the Kappa analysis are provided in the Supplementary Methods.

Ethics committee approval

Clinical investigation presented in this work has been approved by the Research Ethics Committee of the University of Tartu on 12/06/2023. As a retrospective clinical investigation using only anonymized clinical data, informed consent was not required by the Ethics Committee.

Results

Time efficiency in reporting

Across all radiologists, AI-aided workflow led to a mean reduction of 33% in reporting time compared to unaided workflow. Individual radiologists exhibited varying degrees of reduction, ranging from 18 to 52%. The reduction in reporting time when using BMVision was statistically significant ($p < 0.00001$). The impact of BMVision on radiologist reporting time is summarized in Table 1.

In addition to the overall workflow time, the time required to generate a diagnostic report showed a significant reduction. Without AI assistance, the average time to prepare a report was 48.9 s. With the integration of BMVision, this time was reduced to 9.4 s on average, representing an 81% reduction, which was also statistically significant ($p < 0.00001$). Individual radiologists experienced varying levels of improvement, ranging from 71 to 87%.

Diagnostic performance: sensitivity and specificity

The sensitivity of unaided radiologists in detecting benign renal lesions (object-level sensitivity) was 79.9%, compared to 86.3% with AI assistance. For detecting malignant renal lesions, unaided radiologists achieved a sensitivity of 95.6%, while AI-assisted radiologists achieved 96.7%. Statistical testing via bootstrapping revealed that the use of BMVision significantly improved sensitivity for benign lesion detection ($p < 0.00001$), while its impact on malignant lesion detection was non-inferior ($p = 0.41$). Sensitivity and specificity results discussed in this section are presented in Table 2.

At the patient level, unaided radiologists achieved a sensitivity of 89.7% for detecting patients with benign renal lesions, which increased to 95.0% with AI assistance. For detecting patients with malignant renal lesions, sensitivity was 98.0% without AI and 99.16% with AI. Bootstrapping indicated that AI significantly increased patient-level sensitivity for benign lesions ($p < 0.00001$) but had no statistically significant effect on malignant lesions ($p = 0.13$).

Table 1 | Mean reporting times in minutes for unaided and AI-assisted radiologists

	Unaided	AI-assisted	Relative speed-up
Controls	1.95 min	1.32 min	-32%
Cases	4.99 min	3.32 min	-33%
All scans	3.47 min	2.32 min	-33%

BMVision helps the radiologists to work 33% faster on average.

Table 2 | Test set sensitivities and specificities calculated for the AI model, unaided radiologists, and AI-assisted radiologists

		Object-level	Patient-level	
		Sensitivity (%)	Sensitivity (%)	Specificity (%)
Benign	AI only	82.4	94.4	89.3
	Unaided	79.9	89.7	89.1
	AI-assisted	86.3	95.0	91.0
Malignant	AI only	93.4	96.0	95.0
	Unaided	95.6	98.0	99.0
	AI-assisted	96.7	99.2	98.2

Metrics are calculated separately on benign and malignant lesions. AI assistance significantly helps to detect benign lesions and gives comparable results for malignant lesions. Object level specificities are not included since there are no true negative objects.

Table 3 | Inter-radiologists agreement results

	Object-level agreement	Kappa coefficient
Unaided	59.7%	0.68
AI-assisted	82.3%	0.88

Metrics are calculated on all lesions together.

Table 4 | Inter-radiologists agreement results

		Disagreement	Avg # of found lesions
Benign	Unaided	0.902	2.828
	AI-assisted	0.496	3.026
Malignant	Unaided	0.06	0.541
	AI-assisted	0.041	0.563

Metrics are separately calculated on benign and malignant lesions. The use of AI decreases disagreement among radiologists.

The specificity for identifying control patients—those with only benign lesions or no lesions, was 89.1% without AI and 91.1% with AI. For detecting patients with malignant lesions, specificity was 99.0% in the unaided group and 98.2% in the AI-assisted group. Bootstrapping analysis suggested that BMVision had no significant impact on specificity for control patients ($p = 0.37$), nor did it lead to any noticeable reduction in specificity for identifying patients with malignant lesions ($p = 0.16$).

Diagnostic performance: inter-radiologists agreement

BMVision led to a 22.6% absolute increase in object-level agreement compared to unaided radiologists, with the chi-squared test confirming statistical significance ($p < 0.00001$). Similarly, the kappa score improved from 0.68 in the unaided workflow to 0.88 with AI assistance, as confirmed by the Wilcoxon rank-sum test ($p < 0.00001$), indicating a substantial increase in inter-radiologist agreement with the use of BMVision. Results for object-level agreement and kappa scores are summarized in Table 3.

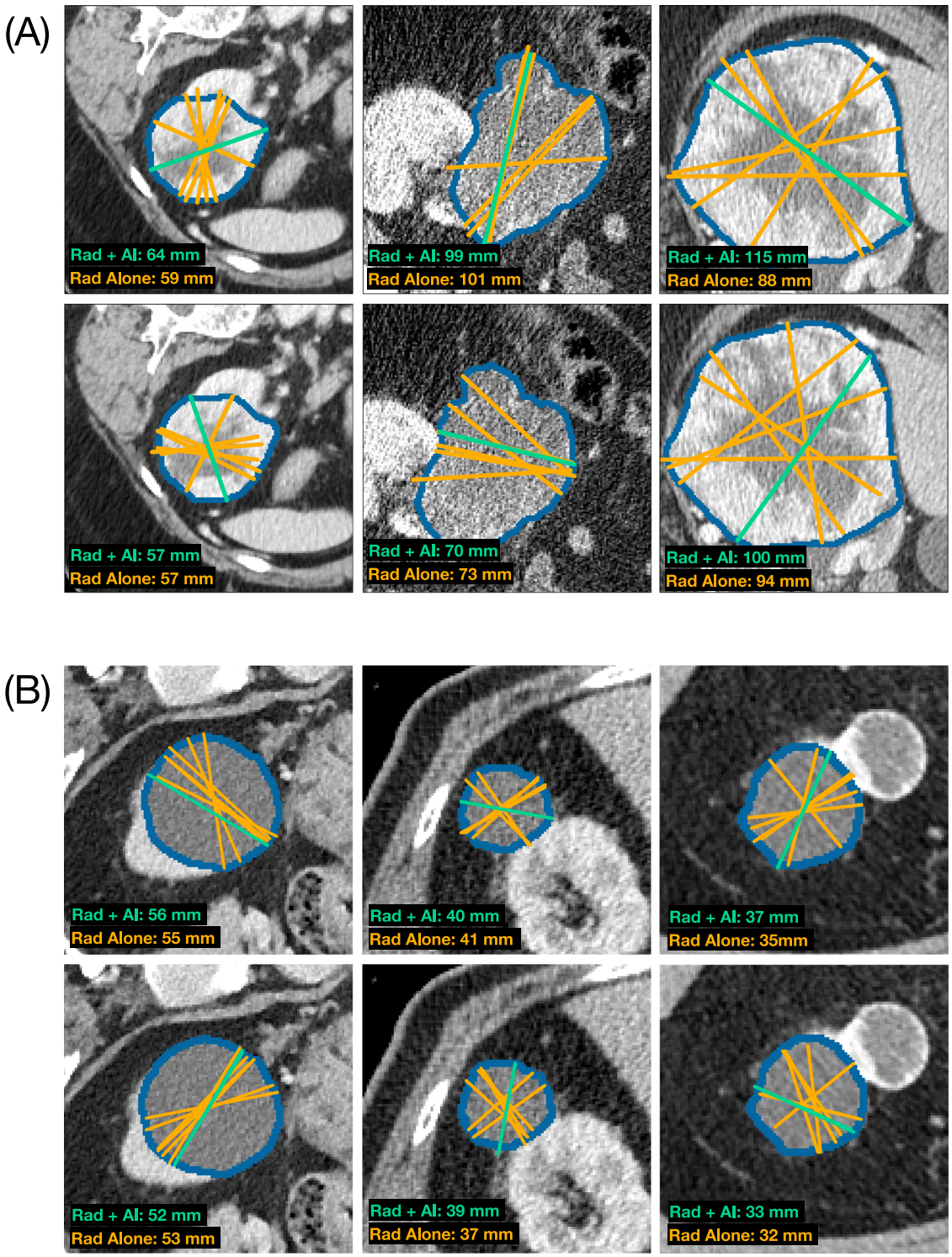

Fig. 3 | Variability between individual radiologists' measurements. Orange lines represent measurements made independently by radiologists, who consistently chose different axes to define either the longest or shortest diameter of the malignant (**A**) and benign (**B**) lesions. However, when assisted by the AI, radiologists uniformly preferred the same measurement proposed by the model, which is denoted by the green line. The length in mm of the longest axis selected by either group is illustrated in the legend.

For benign lesions, the use of AI resulted in a 48.6% relative reduction in disagreement, a statistically significant improvement ($p < 0.00001$). For malignant lesions, the disagreement rate was reduced by 34.3% with BMVision ($p = 0.0008$), further demonstrating the positive impact of AI on consistency in lesion detection. Disagreement rates for benign and malignant lesions are detailed in Table 4. Figure 3 shows an example of variability in radiologist measurements and the consistent lines proposed with AI assistance.

Discussion

Our study demonstrates that AI assistance can significantly enhance efficiency and consistency in kidney cancer reporting workflows. BMVision reduced reporting time by approximately one-third, increased sensitivity for benign lesions, and improved inter-radiologist agreement while maintaining high sensitivity for malignant lesions. These improvements have practical implications: shorter reporting times can ease the workload of radiologists[25], potentially shorten patient waiting times[26,27], help to alleviate patient anxiety[28],

and enable more timely treatment decisions[29]. These gains are relevant across both traditional and voice-enabled reporting setups, as BMVision generates structured reports that reduce the need for manual dictation or typing[30].

Beyond efficiency, diagnostic accuracy also improved, most notably for benign renal lesions. Increased sensitivity for benign findings may help reduce unnecessary biopsies and interventions[31], contributing to safer and more cost-effective care. Specificity remained stable across conditions, indicating that these gains in sensitivity do not come at the expense of increased false positives. By contrast, sensitivity for malignant lesions remained high in both conditions, with no significant difference between AI-assisted and unaided workflows. This outcome likely reflects the study design: radiologists were specifically instructed to focus on kidney cancer, which contributed to the high overall malignant sensitivity of 98%, compared with the 84% sensitivity reported in the literature for incidental detection of malignant kidney lesions on CT scans[32]. In more open-ended settings, where radiologists are asked to evaluate a broader range of organs, AI assistance may reveal more pronounced effects on the detection of malignant lesions.

Equally important is the observed improvement in inter-radiologist agreement. Variability in lesion detection and measurement has long been recognized as a barrier to reliable diagnosis and consistent patient management[33,34]. In our study, AI support led radiologists to adopt more consistent measurement strategies, as illustrated in Fig. 3, where they converged on the same axes proposed by the model. Greater agreement not only strengthens reproducibility in routine care but also supports more reliable treatment planning in multidisciplinary settings, such as tumor boards, where standardized imaging assessments are essential[35,36].

The above findings align with prior studies demonstrating that AI can improve both the accuracy and efficiency of radiological workflows[37]. While some research has validated deep learning models in detecting kidney lesions[38], BMVision, to the best of our knowledge, is the first AI tool for kidney cancer diagnosis to undergo validation within a clinical diagnostic workflow specific to kidney cancer. This validation supports its potential role in enhancing radiological accuracy and efficiency within kidney cancer diagnostics.

While this study demonstrates the potential of BMVision to enhance efficiency and diagnostic consistency in kidney cancer workflows, it is important to acknowledge that the study was conducted at a single site, TUH. This single-center design may limit the generalizability of our findings to other clinical settings. As a next step, a multi-center study employing a more open-ended design, in which radiologists receive less directed instructions, could provide further insights. Such an approach may better reflect routine clinical practice and allow for a more comprehensive evaluation of BMVision's utility in detecting malignant renal lesions beyond study settings.

Another limitation relates to the composition of the control cohort, which consisted of patients with suspected appendicitis. While this selection provided access to abdominal CT scans without renal cancer, it may not fully reflect the diversity of cases encountered in routine clinical practice. To minimize the potential for bias, the appendix region was removed from all scans–both controls and cancer cases–prior to algorithm evaluation. Future work will aim to include a broader range of control cases, such as a random sample of CTs without renal pathology, to better represent the clinical scenarios in which BMVision is expected to operate. Taken together, these limitations highlight the need for broader multi-center validation in real-world, heterogeneous populations.

Conclusion

In this study, we validated BMVision, a decision-support AI tool for kidney cancer diagnosis from CT scans, through a reader study that showed a reduction in reporting time of approximately one-third, with individual improvements of up to 52%. In addition, the tool enhanced sensitivity for detecting benign renal lesions, increasing from 79.9 to 86.3%. The sensitivity for malignant lesions remained high, with no statistically significant difference between AI-assisted and unaided workflows. Notably, BMVision contributed to a significant increase in inter-radiologist agreement, suggesting that AI assistance may enhance consistency in clinical decision-making. With further validation in diverse clinical settings, BMVision has the potential to become a valuable tool in enhancing diagnostic accuracy and efficiency in kidney cancer care, ultimately supporting radiologists in delivering high-quality, timely care to patients.

Data availability

This study used both publicly available and private clinical imaging data. The public datasets (C4KC-KiTS[17,18] and TCGA-KiRC[19,20]) are available through their respective repositories under the conditions described in their data-use agreements. The internal dataset of abdominal CT scans collected at TUH was used under ethical approval granted by the Research Ethics Committee of the University of Tartu and cannot be shared publicly due to patient-privacy and institutional data-protection restrictions (GDPR compliance). The data can be made available to qualified researchers for non-commercial research purposes upon reasonable request to the corresponding author (dmytro.fishman@bettermedicine.ai), subject to approval by the institutional data controller and completion of a data-use agreement. Requests will be reviewed within approximately four weeks. All code necessary to reproduce the model training and evaluation is incorporated into the commercial BMVision platform (Better Medicine OÜ) and can be made available for academic research collaborations under a research agreement.

Code availability

BMVision is a proprietary commercial software developed by Better Medicine OÜ and cannot be publicly shared due to intellectual property and licensing restrictions. Statistical analysis was performed in Python and R using standard libraries. Access to the software for research validation can be provided to qualified investigators under a collaboration agreement with Better Medicine OÜ. Inquiries should be directed to the corresponding author.

Received: 21 April 2025; Accepted: 13 November 2025;
Published online: 27 November 2025

References

1. Ljungberg, B. et al. Eau guidelines for renal cell carcinoma. https://d56bochluxqnz.cloudfront.net/documents/full-guideline/EAU-Guidelines-on-Renal-Cell-Carcinoma-2024.pdf.
2. Powles, T. et al. Renal cell carcinoma: ESMO clinical practice guideline for diagnosis, treatment and follow-up. *Ann. Oncol.* **35**, 692–706 (2024).
3. Winder, M., Owczarek, A. J., Chudek, J., Pilch-Kowalczyk, J. & Baron, J. Are we overdoing it? changes in diagnostic imaging workload during the years 2010–2020 including the impact of the sars-cov-2 pandemic. In *Healthcare*, vol. **9**, 1557 (MDPI, 2021).
4. Kasalak, Ö. et al. Work overload and diagnostic errors in radiology. *Eur. J. Radiol.* **167**, 111032 (2023).
5. Lantsman, C. D. et al. Trend in radiologist workload compared to number of admissions in the emergency department. *Eur. J. Radiol.* **149**, 110195 (2022).
6. Clinical radiology workforce census 2023. https://www.rcr.ac.uk/media/5befglss/rcr-census-clinical-radiology-workforce-census-2023.pdf.
7. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241 (Springer, 2015).
8. Chan, H.-P., Hadjiiski, L. M. & Samala, R. K. Computer-aided diagnosis in the era of deep learning. *Med. Phys.* **47**, e218–e227 (2020).
9. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211 (2020).
10. Ma, J. et al. Segment anything in medical images. *Nat. Commun.* **15** https://doi.org/10.1038/s41467-024-44824-z (2024).
11. Xu, Q. et al. De-lightsam: Modality-decoupled lightweight sam for generalizable medical segmentation arXiv: 2407.14153. (2024).

12. Zhu, J., Hamdi, A., Qi, Y., Jin, Y. & Wu, J. Medical sam 2: Segment medical images as video via segment anything model 2 arXiv: 2408.00874 (2024).
13. Isensee, F. et al. nninteractive: Redefining 3d promptable segmentation arXiv: 2503.08373 (2025).
14. Kirillov, A. et al. Segment anything arXiv: 2304.02643 (2023).
15. de Vente, C., Venkadesh, K. V., van Ginneken, B. & Sánchez, C. I. Slicernninteractive: A 3d slicer extension for nninteractive arXiv: 2504.07991 (2025).
16. Pieper, S., Halle, M. & Kikinis, R. 3d Slicer. In *2004 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro (IEEE Cat No. 04EX821)*, Vol. **1**, 632–635 (2004).
17. Heller, N. et al. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes arXiv: 1904.00445 (2019).
18. Heller, N. et al. Data from C4KC-KiTS: data from the training set of the 2019 Kidney and Kidney Tumor Segmentation Challenge. The Cancer Imaging Archive (TCIA) https://www.cancerimagingarchive.net/collection/c4kc-kits/. Version updated 2020-06-18; 210 subjects; CC BY 3.0 license (2019).
19. Akin, O. et al. The Cancer Genome Atlas Kidney Renal Clear Cell Carcinoma Collection (TCGA-KIRC), Version 3. The Cancer Imaging Archive (TCIA) https://www.cancerimagingarchive.net/collection/tcga-kirc/. Version updated 2020-05-29; 267 subjects, CC BY 3.0 license (2016).
20. Network, C. G. A. R. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499**, 43–49 (2013).
21. Vogel, C. et al. Imaging in suspected renal-cell carcinoma: systematic review. *Clin. Genitourin. Cancer* **17**, e345–e355 (2019).
22. Akram, F. et al. Diagnostic accuracy of contrast enhanced ct for detection of renal cell carcinoma taking histopathology as gold standard. *J Ayub Med. Col. Abbottabad-Pakistan* **35** (2023).
23. Brierley, J. D., Gospodarowicz, M. K. & Wittekind, C.*TNM Classification of Malignant Tumours* (Wiley, 2017).
24. Alsaikhan, N., Alshehri, W., Cassidy, F., Aganovic, L. & Vahdat, N. Renal tumor structured reporting including nephrometry score and beyond: what the urologist and interventional radiologist need to know. *Abdom. Radiol.* **44**, 190–200 (2019).
25. Krupinski, E. A., Berbaum, K. S., Caldwell, R. T., Schartz, K. M. & Kim, J. Long radiology workdays reduce detection and accommodation accuracy. *J. Am. Coll. Radiol.* **7**, 698–704 (2010).
26. Pierre, K. et al. Applications of artificial intelligence in the radiology roundtrip: process streamlining, workflow optimization, and beyond. *Semin. Roentgenol.* **58**, 158–169 (2023).
27. Li, X. et al. Artificial intelligence-assisted reduction in patients' waiting time for outpatient process: a retrospective cohort study. *BMC Health Serv. Res.* **21** https://doi.org/10.1186/s12913-021-06248-z. (2021).
28. Gagliardi, A. R. et al. The psychological burden of waiting for procedures and patient-centred strategies that could support the mental health of wait-listed patients and caregivers during the covid-19 pandemic: a scoping review. *Health Expect.* **24**, 978–990 (2021).
29. Reichert, A. & Jacobs, R. The impact of waiting time on patient outcomes: evidence from early intervention in psychosis services in england. *Health Econ.* **27**, 1772–1787 (2018).
30. Koivikko, M. P., Kauppinen, T. & Ahovuo, J. Improvement of report workflow and productivity using speech recognition-a follow-up study. *J. Digit. Imaging* **21**, 378–382 (2008).
31. Krishna, S., Leckie, A., Kielar, A., Hartman, R. & Khandelwal, A. Imaging of renal cancer. *Semin. Ultrasound CT MRI* **41**, 152–169 (2020).
32. Bradley, A., Maskell, G., Mannava, A., Pollard, A. & Welsh, T. Routes to diagnosis and missed opportunities in the detection of renal cancer. *Clin. Radiol.* **76**, 129–134 (2021).
33. Gierada, D. S., Rydzak, C. E., Zei, M. & Rhea, L. Improved interobserver agreement on lung-rads classification of solid nodules using semiautomated CT volumetry. *Radiology* **297**, 675–684 (2020).
34. Ekpo, E. U., Ujong, U. P., Mello-Thoms, C. & McEntee, M. F. Assessment of interradiologist agreement regarding mammographic breast density classification using the fifth edition of the BI-RADS atlas. *Am. J. Roentgenol.* **206**, 1119–1123 (2016).
35. Benchoufi, M., Matzner-Lober, E., Molinari, N., Jannot, A.-S. & Soyer, P. Interobserver agreement issues in radiology. *Diagn. Interven. Imaging* **101**, 639–641 (2020).
36. Hasnain, M., Onishi, H. & Elstein, A. S. Inter-rater agreement in judging errors in diagnostic reasoning. *Med. Educ.* **38**, 609–616 (2004).
37. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H. & Aerts, H. J. W. L. Artificial intelligence in radiology. *Nat. Rev. Cancer* **18**, 500–510 (2018).
38. ETEM, T. & TEKE, M. Enhanced deep learning based decision support system for kidney tumour detection. *BenchCouncil Trans. Benchmarks Stand. Eval.* **4**, 100174 (2024).

Acknowledgements

We would like to thank Reinis Zarinš and Rauno Pihlak for their assistance with the labeling of training data. We also acknowledge the support from wider Better Medicine, the University of Tartu, and the TUH teams, which made this research possible.

Author contributions

Bohdan Petryshak, Dmytro Fishman and Priit Salumaa designed the study. Bohdan Petryshak developed and trained the deep learning model that underpins BMVision. Bohdan Petryshak, Mikhail Iljin, and Joonas Ariva analyzed the results. Alina Denissova, Martin Reim, and Ihor Ivaniv labeled the training data. Toomas Häide and Veljo Lasn developed the BMVision tool. Toomas Lillsaar, Hardi Vilt, Anu Leht, Anti Karask, Gitana Kiudma, and Pilvi Ilves participated as test radiologists in the study. Rait Käpp supervised data acquisition. Dmytro Fishman and Priit Salumaa supervised the overall project. Dmytro Fishman, Bohdan Petryshak, Joonas Ariva, Mikhail Iljin, and Alina Denissova wrote the manuscript, with input from all co-authors.

Competing interests

Bohdan Petryshak, Dmytro Fishman, Martin Reim, and Priit Salumaa are employees and co-founders, and therefore shareholders, of Better Medicine, the company that sponsored this research. Three out of these four (except Dmytro Fishman) have received salaries from Better Medicine. Mikhail Iljin, Toomas Häide, Veljo Lasn, Alina Denissova, and Ihor Ivaniv are employees or ex-employees of Better Medicine. Rait Käpp is a consultant for Better Medicine. Joonas Ariva is a PhD student at the University of Tartu, working on a research project sponsored by Better Medicine. Martin Reim, Alina Denissova, Toomas Lillsaar, Hardi Vilt, Anu Leht, Anti Karask, Gitana Kiudma, and Pilvi Ilves are employees of TUK, which has received research funding from Better Medicine for this project.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43856-025-01264-0.

Correspondence and requests for materials should be addressed to Dmytro Fishman.

Peer review information *Communications Medicine* thanks Zhenjie Cao and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.