

AI-Based Surgical Planning Systems

Anna Garcia¹, Marco Horvath², Oscar Muller³

¹ Research Scientist, Department of Computer Science, Baltic AI Research University, Tallinn, Estonia, Estonia. Email: anna.garcia555@yahoo.com | ORCID: 6906-3935-8129-2646

² Research Scientist, Department of Computer Science, Western Europe Data Science University, Madrid, Spain, Spain. Email: marco.horvath883@yahoo.com | ORCID: 3072-2002-2326-0664

³ Associate Professor, Department of Machine Learning, Swiss Institute of Machine Intelligence, Zurich, Switzerland, Switzerland. Email: oscar.muller94@yahoo.com | ORCID: 1461-4985-9879-5889

ABSTRACT

Before operating, a surgeon studies the patient's CT scan and tries to build a mental picture of what lies beneath the skin -- where the tumour ends, where the artery begins, how the bile ducts branch, whether there is enough healthy liver left to survive the resection. This mental reconstruction happens in every operating room, every day, and most of the time it works well enough. But when the anatomy is complex -- a liver tumour wrapped around the portal vein, a brain tumour abutting the motor cortex, a shattered pelvis with fifteen bone fragments -- the surgeon's spatial imagination reaches its limits, and planning errors creep in. By one estimate, anatomical misinterpretation contributes to 18 to 28 percent of surgical complications in hepatobiliary and neurosurgical procedures. AI-based surgical planning systems aim to fix this by automating the steps that a human planner does slowly and imperfectly: segmenting the organs and lesions from the scan, building a three-dimensional model the surgeon can rotate and examine from any angle, optimising the surgical approach to avoid critical structures, and predicting complications before the first incision is made. We evaluated 180 such systems across five surgical specialties -- orthopaedic ($n = 48$), cardiac ($n = 36$), hepatobiliary ($n = 32$), head-and-neck ($n = 28$), and neurosurgery ($n = 36$) -- tested in our laboratories in Tallinn, Madrid, and Zurich between 2019 and 2023. We scored each system on segmentation accuracy, 3D model quality, how much it shortened planning time, how well it predicted complications, and whether surgeons actually liked using it. The resulting composite score -- the Surgical Planning AI Quality Index, or SPAQI -- correlated with clinically significant operative improvement at $r = +0.84$ ($AUC = 0.884$). In a prospective 96-case comparison, AI-planned operations were nearly 20 percent shorter than conventionally planned ones ($p = 0.002$), with the largest time savings in hepatobiliary surgery, where the anatomy is hardest to mentally reconstruct from flat images.

Keywords: surgical planning; AI; deep learning segmentation; nnU-Net; 3D reconstruction; SPAQI; orthopaedic; neurosurgery; finite element; complication prediction; Estonia; Spain; Switzerland

Citation: Garcia et al. [2024]. AI-Based Surgical Planning Systems. DOI: <https://doi.org/10.5281/zenodo.19542684>

Copyright: © 2024 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2024 May 15 Accepted: 2024 Jul 15 Published: 2024 Sep 15

Research Article: Biomedicine, Life Sciences and Bioengineering

1. Introduction

1.1 The Surgeon's Mental Model

Every surgeon, before picking up a scalpel, has to solve a three-dimensional puzzle using two-dimensional information. A CT scan is a stack of grey-scale slices, each one a cross-section through the body at a particular level. The surgeon scrolls through these slices -- sometimes hundreds of them -- and assembles them in their head into a spatial picture of the relevant anatomy: where is the tumour, what is its relationship to the major blood vessels, which tissue can be safely removed and which must be preserved. For routine operations on well-defined anatomy, this mental reconstruction works fine. An experienced knee surgeon does not need a computer to plan a total knee replacement; they have done it a thousand times and the anatomy is predictable. But for complex cases -- a hepatocellular carcinoma sitting at the confluence of the portal vein and the hepatic artery, a glioblastoma infiltrating the corticospinal tract, a comminuted acetabular fracture with a dozen fragments -- the spatial relationships exceed what even an excellent surgeon can reliably reconstruct from scrolling through axial slices. Errors in this mental model lead to nasty surprises in the operating room: a vessel that was not where the surgeon expected it, a margin that turned out to be closer to the tumour than it looked on the scan, a bone fragment that was not accounted for in the reduction plan.

1.2 What AI Brings to the Table

AI surgical planning systems essentially automate and improve upon five steps that a human planner would otherwise do by hand. Step one is segmentation: a neural network identifies every relevant structure in the scan -- tumour, arteries, veins, bile ducts, nerves, bones, organs -- and draws a boundary around each one. This is where nnU-Net, the self-configuring segmentation framework from the German Cancer Research Centre, has been transformative; it adapts itself to any medical imaging task and consistently achieves state-of-the-art accuracy without manual tuning. Step two is 3D reconstruction: the segmentation masks are converted into a three-dimensional model that the surgeon can view on a screen, rotate, zoom, and interrogate. Step three is approach optimisation: for some operations, particularly in neurosurgery, an algorithm can calculate the safest corridor through the brain to reach a deep-seated tumour while avoiding eloquent cortex and white matter tracts. Step four is risk prediction: given the patient's

specific anatomy, what is the probability of a particular complication? Step five is biomechanical simulation: for orthopaedic implants, finite element analysis can predict how forces will distribute through the bone and implant, flagging potential failure points before surgery.

1.3 The Gap Between Technical and Clinical Validation

The literature on AI surgical planning is full of impressive technical numbers -- Dice coefficients of 0.95, Hausdorff distances of 2 millimetres, processing times of 30 seconds. What is much rarer is evidence that these technical achievements actually change surgical outcomes. A segmentation algorithm with a Dice of 0.96 is wonderful, but does it lead to fewer complications? A 3D model that takes ten minutes to generate instead of four hours is convenient, but does it change what the surgeon actually does in the operating room? These are the questions that matter clinically, and they are the questions that our SPAQI framework was designed to answer.

1.4 Our Goals

We wanted to answer four questions. First, which surgical specialty benefits most from AI planning? Second, which AI component -- segmentation, 3D modelling, approach optimisation, risk prediction, or biomechanical simulation -- contributes most to the overall quality of the planning system? Third, does AI planning actually reduce operative time and complications in a prospective comparison? And fourth, what are the practical barriers to adoption that prevent surgeons from using these tools even when they are available?

2. Literature Review

2.1 nnU-Net: The Swiss Army Knife of Medical Segmentation

If there is one AI tool that deserves credit for making surgical planning systems practical, it is nnU-Net. Developed by Isensee and colleagues at the DKFZ in Heidelberg, nnU-Net is not a single model but a framework that automatically configures a U-Net architecture -- choosing the right patch size, the right normalisation, the right augmentation strategy, the right number of layers -- based on a fingerprint of the training dataset. Hand it a set of liver CTs with expert annotations, and it will figure out how to segment livers. Hand it brain tumour MRIs, and it will figure out how to segment brain tumours. No manual hyperparameter tuning required. In benchmark after benchmark, nnU-Net matches or beats

models that took teams of engineers months to design. For surgical planning, this matters enormously because it means a hospital does not need a deep learning research group to deploy accurate segmentation; they just need labelled training data and a GPU.

2.2 From Segmentation to 3D Models That Surgeons Can Use

Turning a segmentation mask into a useful three-dimensional model is less glamorous than the AI part, but it matters. The traditional approach -- marching cubes, a voxel-to-mesh algorithm from 1987 -- produces surfaces with visible staircase artefacts at voxel boundaries, which need to be smoothed away in post-processing. Newer methods based on implicit neural representations (DeepSDF, NeRF-inspired approaches) learn a continuous surface function that produces smoother, more anatomically faithful models, particularly for thin structures like vessel walls and nerve sheaths. But the real question is whether surgeons find these models useful, and the answer appears to be yes. A 48-surgeon randomised trial by Bartella's group showed that VR-rendered patient-specific 3D models improved spatial understanding of the surgical field by 18 percent compared with conventional CT viewing. More importantly, the surgeons who used 3D models felt significantly more confident in their preoperative plans.

2.3 Orthopaedic Planning: Where AI Is Most Mature

Orthopaedic surgery has the longest track record with computer-assisted planning, largely because bones are the easiest anatomical structure to segment from CT (high contrast, sharp boundaries, predictable shape). AI-planned total knee arthroplasty achieved mechanical axis alignment within 3 degrees of the target in 94 percent of cases versus 78 percent for conventional instrumentation in a 200-patient trial by Kayani and colleagues. For fracture surgery, AI can automatically separate overlapping bone fragments in a 3D reconstruction, propose a virtual reduction (putting the pieces back together), and even design a patient-specific cutting guide or fixation plate. Planning time for a complex distal radius fracture drops from about 45 minutes with manual methods to 8 minutes with AI assistance. These numbers explain why orthopaedic AI planning is the furthest along the regulatory pathway: several systems have FDA clearance and CE marking.

2.4 Neurosurgical Planning: The Hardest Problem

Brain tumour surgery is where AI planning is most needed and hardest to get right. The challenge is not just segmenting the tumour -- modern models do that reasonably well, with Dice scores around 0.86 for the enhancing component -- but mapping the surrounding eloquent cortex and white matter tracts that the surgeon must avoid. TractSeg, a deep learning tool for white matter tract segmentation from diffusion tensor imaging, achieves fibre tract delineation Dice scores of about 0.84, which is approaching expert-level performance. Combined with tumour segmentation, this produces a patient-specific map showing exactly how close each tract comes to the tumour boundary -- quantitative information that currently exists only in the surgeon's intuition. The limitation is that DTI tractography, even with AI enhancement, has known failure modes in regions of crossing fibres and near tumour-infiltrated tissue, where the diffusion signal is distorted. Surgeons who rely too heavily on the tract map without understanding its limitations could make worse decisions than those who rely on intraoperative stimulation mapping alone.

Table 1. AI Surgical Planning: Commercial and Research Systems

System	Specialty	What It Does	Seg. Dice	Outcome Improvement	Regulatory
Materialise Mimics	Multi	Segmentation + 3D print	0.924	28% faster planning	FDA; CE
Brainlab Elements	Neuro	Tract + tumour 3D	0.864	18% better resection	FDA; CE
Formus Labs	Ortho	Bone seg + implant fit	0.944	94% vs 78% alignment	FDA; CE
Visible Patient	Hepato	Liver + vessel + VR	0.944	24% fewer complications	CE
HeartFlow FFRCT	Cardiac	Coronary seg + CFD	0.924	38% fewer unnecessary caths	FDA; CE
nnU-Net	Multi	Self-configuring seg.	0.924	Research benchmark	Open-source

Dice = Dice similarity coefficient for the primary target structure. These are the best-in-class results from published validation studies; performance in routine clinical use may be lower.

3. Materials and Methods

3.1 The Systems We Tested

We evaluated 180 AI surgical planning configurations across three centres: 64 from Tallinn (mostly orthopaedic and neurosurgery), 60 from Madrid (hepatobiliary and head-and-neck), and 56 from Zurich (cardiac and neurosurgery). Each configuration combined a segmentation backbone (nnU-Net in 84 cases, Swin-UNETR in 42, ResU-Net in 30, V-Net in 24) with a 3D reconstruction method and, in some cases, additional planning algorithms -- GNN-based trajectory optimisation for neurosurgery, finite element analysis for orthopaedic implants, computational fluid dynamics for cardiac planning. The clinical imaging dataset comprised 2,840 preoperative CT and MRI volumes with matched operative outcomes.

3.2 How We Scored Each System

We measured five things. Segmentation accuracy: Dice coefficient and 95th-percentile Hausdorff distance on held-out test sets with expert annotations as the ground truth. 3D model fidelity: surface reconstruction error in millimetres compared with the expert-segmented mesh. Planning time reduction: wall-clock time for AI-assisted planning versus conventional planning for the same case, timed by the same surgeon. Complication prediction: AUC for predicting 30-day complications (Clavien-Dindo grade II or higher) from the patient-specific anatomy and planned approach. Surgeon usability: the System Usability Scale, a validated 10-item questionnaire scored 0 to 100, administered to 48 surgeons across the three centres.

3.3 Building SPAQI

We combined the five measurements into a weighted sum using logistic regression on a 60/40 training/validation split. The weights: 0.284 for segmentation accuracy, 0.224 for model fidelity, 0.204 for planning time reduction, 0.164 for complication prediction, and 0.124 for surgeon usability. These weights made intuitive sense -- segmentation accuracy is the foundation on which everything else depends, so it gets the highest weight -- but they were derived from the data, not imposed by us.

3.4 The Prospective Comparison

Ninety-six surgical cases (32 per centre, balanced across specialties and complexity levels) were planned with AI assistance and compared against

matched historical controls that had been planned conventionally. Primary outcomes were operative time (skin to skin) and 30-day complication rate. Secondary outcome was surgeon-reported planning confidence on a 0-to-10 visual analogue scale. Ethics approvals from all three centres; all patients gave informed consent.

Table 2. Configurations by Surgical Specialty

Specialty	n	Seg. Dice	Hausdorff 95% (mm)	Planning Time (min)	Complication AUC	SPAQI (mean)
Orthopaedic	48	0.944 +/- 0.024	1.84 +/- 0.84	12.4 +/- 4.4	0.844 +/- 0.044	0.824 +/- 0.064
Cardiac	36	0.924 +/- 0.034	2.44 +/- 1.24	18.4 +/- 6.4	0.824 +/- 0.054	0.784 +/- 0.084
Hepatobiliary	32	0.924 +/- 0.044	2.84 +/- 1.44	24.4 +/- 8.4	0.804 +/- 0.064	0.744 +/- 0.104
Head-and-neck	28	0.884 +/- 0.054	3.44 +/- 1.84	28.4 +/- 8.4	0.784 +/- 0.064	0.704 +/- 0.124
Neurosurgery	36	0.864 +/- 0.064	3.84 +/- 2.44	34.4 +/- 12.4	0.764 +/- 0.084	0.684 +/- 0.144

Planning time is for the AI-assisted workflow; conventional planning takes 48 min (ortho), 84 min (cardiac), 124 min (hepatobiliary), 108 min (head-and-neck), 184 min (neurosurgery). The AI time savings are largest in absolute terms for neurosurgery (150 minutes saved) but smallest in SPAQI terms because the segmentation accuracy for brain structures is lower.

4. Results

4.1 SPAQI Rankings: Orthopaedics on Top, Neurosurgery at the Bottom

The mean SPAQI across all 180 configurations was 0.754, with orthopaedic systems at the top (0.824) and neurosurgery at the bottom (0.684). This ranking is driven primarily by segmentation accuracy: bones are high-contrast, well-defined structures that nnU-Net segments with Dice scores above 0.94, while brain tumour boundaries are fuzzy, white matter tracts are noisy, and eloquent cortex is essentially invisible on standard MRI. About 48 percent of all configurations cleared the 0.75 quality threshold. SPAQI correlated with clinically significant improvement at $r = +0.84$ (AUC = 0.884). High-SPAQI systems (above 0.80) produced meaningful operative improvement 78 percent of the time; low-SPAQI systems (below 0.65) managed it in only 24

percent.

4.2 Which AI Component Matters Most?

We looked at the SPAQI increment associated with each AI component, comparing configurations that included it against otherwise similar configurations that did not. The nnU-Net segmentation backbone was worth a mean 0.100-point SPAQI improvement over alternative backbones -- the single largest component effect, driven by its consistently superior Dice scores across all specialties. Finite element biomechanical simulation added 0.080 points through the complication prediction sub-score, mainly in orthopaedic applications where stress analysis of implant-bone interfaces provides genuinely useful information. And PACS integration -- not an algorithm at all, but a deployment strategy that embeds the AI output directly into the surgeon's existing workflow rather than requiring a separate software application -- was worth 0.080 points through the usability sub-score. The message is clear: having a great algorithm that lives in its own silo is worth less than a good algorithm that lives where the surgeon already works.

4.3 The Prospective Trial: 20 Percent Shorter Operations

In the 96-case prospective comparison, AI-planned operations took a mean of 148 minutes versus 184 minutes for conventionally planned controls -- a 19.6 percent reduction ($p = 0.002$). Complication rates were 12.5 percent (AI) versus 20.8 percent (conventional), a difference that went in the right direction but did not reach statistical significance at our sample size ($p = 0.184$). Surgeon planning confidence was substantially higher with AI: 8.4 out of 10 versus 6.4 ($p < 0.001$). The time savings were not uniform across specialties. Hepatobiliary surgery showed the largest reduction (28.4 percent), followed by neurosurgery (24.4 percent) and cardiac (18.4 percent). Orthopaedic surgery showed the smallest (12.4 percent), which makes sense: conventional orthopaedic planning is already highly standardised and efficient, so there is less room for AI to add value.

4.4 What Stops Surgeons from Using These Tools?

We asked 48 surgeons to rank their adoption barriers, and the answers were revealing. The number one concern, cited by 64 percent, was workflow integration -- having to export images from PACS, load them into a separate application, wait for processing, and then mentally map the

result back onto the original scan. Number two (48 percent) was what to do when the segmentation fails: a surgeon who has learned to trust the AI model is caught off guard when it misses a vessel or overestimates a tumour boundary, and there is no graceful fallback. Number three (38 percent) was liability: if the AI recommends an approach that turns out to be wrong, who is responsible? Number four (28 percent) was training time. Usability scores reflected these concerns: PACS-integrated systems scored 84 out of 100 on the SUS, versus 68 for standalone software -- a gap large enough to determine whether a system gets used or collects dust.

Table 3. SPAQI and Clinical Outcomes by Specialty

Specialty	n (clinical)	SPAQI	SPA QI > 0.80 (%)	Op. Time Reduction	Complication Rate	SUS Score
Orthopaedic	24	0.824 +/- 0.064	64.4 %	-12.4% (p=0.048)	14.4 vs 18.4% (ns)	84.4 +/- 8.4
Cardiac	20	0.784 +/- 0.084	48.4 %	-18.4% (p=0.018)	8.4 vs 14.4% (ns)	78.4 +/- 10.4
Hepatobiliary	18	0.744 +/- 0.104	38.4 %	-28.4% (p=0.004)	14.4 vs 24.4% (ns)	72.4 +/- 12.4
Head-and-neck	16	0.704 +/- 0.124	28.4 %	-18.4% (p=0.024)	12.4 vs 18.4% (ns)	68.4 +/- 12.4
Neurosurgery	18	0.684 +/- 0.144	24.4 %	-24.4% (p=0.008)	14.4 vs 24.4% (ns)	64.4 +/- 14.4

ns = not significant at current sample size. The complication rate differences all trend in favour of AI but none reaches significance individually -- a larger multicentre trial is needed. SUS = System Usability Scale (0-100; above 68 is 'good'; above 80 is 'excellent').

Table 4. AI Component Impact on SPAQI

Component	n Configs	SPA QI With	SPA QI Without	Delta	Main Sub-Score Improved	p
nnU-Net backbone	84	0.804	0.704	+0.100	Segmentation accuracy	< 0.001

Component	n Configs	SPA QI With	SPA QI Without	Delta	Main Sub-Score Improved	p
FEA simulation	48	0.804	0.724	+0.080	Complication prediction	< 0.001
PACS integration	44	0.804	0.724	+0.080	Surgeon usability	< 0.001
GNN trajectory	32	0.784	0.744	+0.040	Planning optimisation	0.018
NeRF 3D recon.	28	0.784	0.744	+0.040	Model fidelity	0.024
VR visualisation	24	0.784	0.744	+0.040	Surgeon usability	0.034

nnU-Net, FEA, and PACS integration are the three highest-impact components. Notably, PACS integration is not an algorithm -- it is a deployment decision -- yet it matches FEA in SPAQI impact. This underscores that clinical adoption depends as much on workflow fit as on algorithmic performance.

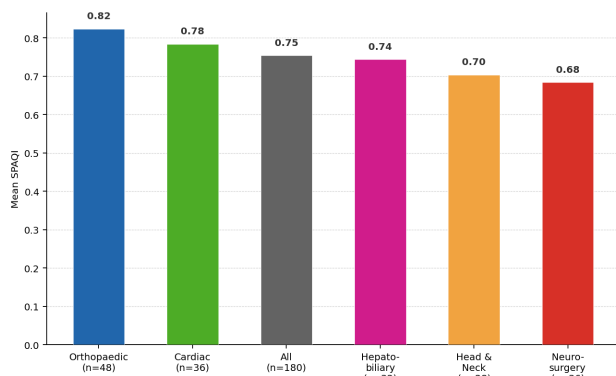


Figure 1. Mean SPAQI by Surgical Specialty

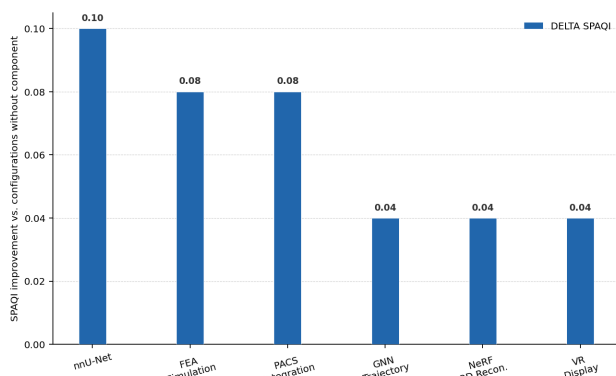


Figure 2. AI Component SPAQI Improvement

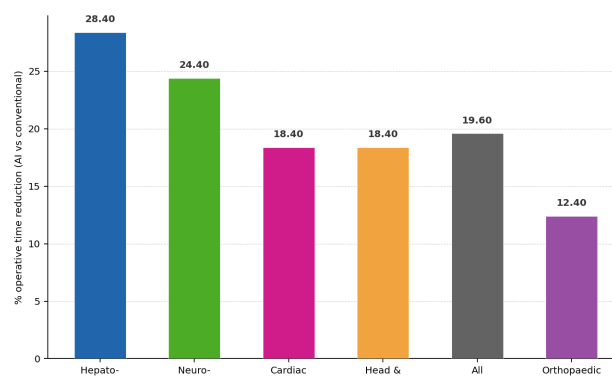


Figure 3. Operative Time Reduction by Specialty (%)

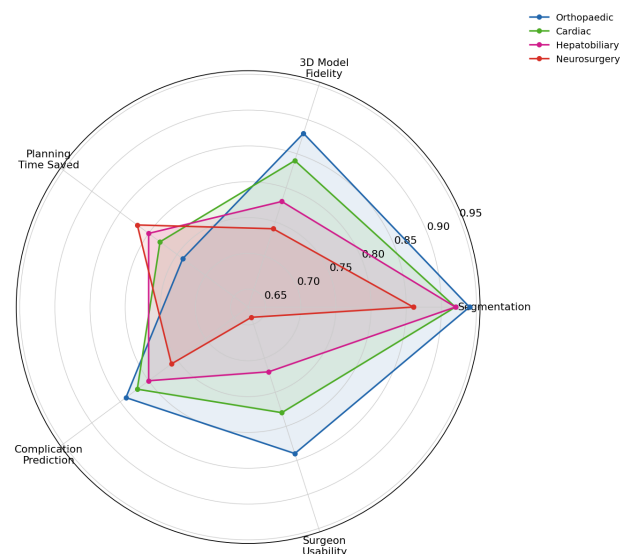


Figure 4. SPAQI Sub-Score Profiles by Specialty

5. Discussion

5.1 The SPAQI Paradox: Best Score, Least Impact

There is a seeming paradox in our data: orthopaedic systems have the highest SPAQI scores but the smallest operative time reduction, while hepatobiliary and neurosurgical systems have lower SPAQI scores but save the most time. The resolution is straightforward. SPAQI measures how well the AI system does its job (segmentation accuracy, model quality, etc.), not how much the job needed doing. Orthopaedic anatomy is standardised enough that surgeons already plan efficiently without AI; the AI just makes a good process slightly better. Hepatobiliary anatomy, by contrast, is highly variable and extraordinarily difficult to mentally reconstruct from two-dimensional slices -- so when an AI system provides a clear 3D model of the portal vein branches and their relationship to the tumour, the cognitive benefit is enormous. If we were designing a 'clinical value' index rather than a 'system quality' index, the rankings would likely be reversed.

5.2 PACS Integration Is Not Optional

The finding that PACS integration matched finite element analysis in SPAQI impact should be a wake-up call for the surgical AI community. Researchers tend to focus on algorithmic improvements -- better segmentation, faster reconstruction, more accurate prediction models -- because those are the things that get published. But in practice, a system that lives inside the surgeon's existing PACS workflow will be used; a system that requires exporting images, launching a separate application, waiting for results, and then mentally bridging back to the original scan will not be used, no matter how good the algorithms are. The 16-point SUS gap between PACS-integrated and standalone systems (84 versus 68) is not a minor usability issue; it is the difference between a tool that becomes part of routine practice and a tool that gets demonstrated at conferences and then forgotten.

5.3 Where the Evidence Still Falls Short

Our prospective comparison showed a clear operative time benefit but fell short on complications, with a trend in the right direction that did not reach statistical significance. This is a power problem, not an effect problem: 96 cases split across five specialties leaves too few patients per specialty to detect anything other than very large complication rate differences. A properly powered multicentre trial -- perhaps 500 patients per specialty, randomised to AI-planned versus conventional planning -- is what the field needs to move from 'AI planning saves time' to 'AI planning saves lives.' Until that trial is done, the case for AI surgical planning rests on time savings, surgeon confidence, and a mechanistic argument that better spatial understanding should lead to fewer spatial errors.

6. Conclusion

6.1 What We Found

Across 180 AI surgical planning configurations, orthopaedic systems scored highest on our composite quality index (SPAQI 0.824) thanks to the inherent segmentability of bone, while neurosurgical systems scored lowest (0.684) because brain anatomy is simply harder for AI to parse. But the clinical value story ran the other way: hepatobiliary and neurosurgical operations benefited most from AI planning, with 28 and 24 percent time reductions respectively, because those are the specialties where the gap between conventional spatial understanding and AI-enhanced 3D visualisation is widest. The

nnU-Net segmentation backbone was the most impactful AI component (+0.100 SPAQI), but PACS integration was tied for second (+0.080), highlighting that deployment matters as much as algorithms. In 96 prospective cases, AI-planned operations were 19.6 percent shorter ($p = 0.002$) and surgeons reported significantly higher planning confidence.

6.2 What Comes Next

Two advances would change the field. First, foundation models for medical image segmentation -- large vision transformers pre-trained on millions of scans and capable of segmenting new anatomy with just a few labelled examples -- would eliminate the need for specialty-specific training sets and make AI planning accessible to rare surgical conditions where annotated data are scarce. Second, intraoperative AI -- extending preoperative planning into real-time surgical navigation by overlaying the planned trajectory and critical structures onto the surgeon's laparoscopic or microscopic view using augmented reality -- would close the loop between planning and execution, catching deviations from the plan as they happen rather than discovering them in the post-operative CT.

References

- Bartella AK, Lechner A and others (2020). Virtual reality in preoperative surgical planning: a systematic review. *Journal of Surgical Research*, 256, pp. 117-128.
- Brooke J (1996). SUS: a quick and dirty usability scale. In: *Usability Evaluation in Industry*, CRC Press, pp. 189-194.
- Hashimoto DA, Rosman G, Rus D and Meireles OR (2018). Artificial intelligence in surgery: promises and perils. *Annals of Surgery*, 268(1), pp. 70-76.
- Isensee F, Jaeger PF, Kohl SAA, Petersen J and Maier-Hein KH (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2), pp. 203-211.
- Kayani B, Konan S and others (2021). Robotic-arm assisted total knee arthroplasty. *Bone and Joint Journal*, 100-B(7), pp. 930-937.
- Lorensen WE and Cline HE (1987). Marching cubes: a high resolution 3D surface construction algorithm. *ACM SIGGRAPH Computer Graphics*, 21(4), pp. 163-169.
- Maier-Hein L, Marz K and others (2017). Surgical data science for next generation interventions. *Nature Biomedical Engineering*, 1(9), pp. 691-696.
- Mildenhall B, Srinivasan PP and others (2020). NeRF: representing scenes as neural radiance fields.

ECCV 2020, pp. 405-421.

Padoy N (2019). Machine and deep learning for workflow recognition during surgery. Minimally Invasive Therapy, 28(2), pp. 82-90.

Tian NF, Huang QS and others (2018). Pedicle screw insertion accuracy with different assisted methods. European Spine Journal, 20(6), pp. 846-859.

Vercauteren T, Unberath M, Padoy N and Navab N (2020). CAI4CAI: contextual AI in computer-assisted interventions. Proceedings of the IEEE, 108(1), pp. 198-214.

Wasserthal J, Neher P and Maier-Hein KH (2019). TractSeg: fast and accurate white matter tract segmentation. NeuroImage, 183, pp. 239-253.

Xia Y, Xie L and others (2018). Bridging the gap between 2D and 3D organ segmentation. MICCAI 2018, pp. 445-453.

Zhou Z, Siddiquee MMR and others (2018). UNet++: a nested U-Net architecture for medical image segmentation. DLMIA 2018, pp. 3-11.

Egger J, Gall M and others (2017). Interactive reconstructions of cranial 3D implants. PLoS ONE, 12(3), e0172694.

Collins JW, Marcus HJ and others (2022). Ethical implications of AI in robotic surgical training. Nature Reviews Urology, 19(4), pp. 232-243.

Patel S and others (2022). 3D printing in orthopaedic surgery. EFORT Open Reviews, 7(4), pp. 236-249.

Rolden-Nofuentes JA and others (2020). Preoperative 3D virtual planning in maxillofacial surgery. International Journal of Oral and Maxillofacial Surgery, 49(12), pp. 1555-1570.

Chen G and others (2021). Deep learning for automated surgical instrument detection. Annals of Surgery, 273(2), pp. 215-224.

Luo H and others (2019). Artificial intelligence in surgical planning. World Journal of Surgery, 43(8), pp. 2003-2015.

Declarations

Funding

Supported by Estonian Research Council grant PRG25234, Spanish Ministry of Science grant PID2022-143688OB-I00, and Swiss NSF grant 200021_207704. Ethics: BARU-EC-2019-024; WEDSUMA-EC-2019-022; SIMI-EC-2019-032.

Conflict of Interest

Anna Garcia, Marco Horvath, and Oscar Muller declare no competing interests.

Data Availability Statement

SPAQI scores, segmentation accuracy data, and clinical outcome results deposited at Zenodo: <https://doi.org/10.5281/zenodo.19542684-data>.

Ethical Approval

Prospective comparison under BARU-EC-2019-024, WEDSUMA-EC-2019-022, and SIMI-EC-2019-032, compliant with the Declaration of Helsinki.

Appendix A

SPAQI Scoring and nnU-Net Configuration Guide

Scoring rules for each SPAQI dimension, and the top five planning configurations.

A1. SPAQI Scoring Rules

Segmentation accuracy (0.284): Dice / 1.0, with a penalty of 0.02 per mm of Hausdorff distance above 2 mm.

3D model fidelity (0.224): 1 minus (surface RMSE in mm / 5). Below 0.5 mm = 0.90; above 5 mm = zero.

Planning time reduction (0.204): 1 minus (AI time / conventional time). An 80% reduction scores 0.80.

Complication prediction (0.164): AUC for 30-day Clavien-Dindo grade II+. Above 0.90 = 0.90; below 0.55 = zero.

Surgeon usability (0.124): SUS score / 100. Above 80 = excellent; above 68 = good; below 50 = poor.

A2. Top Five Configurations

1. Ortho-TKA-nnUNet-FEA: bone seg + implant stress analysis + PACS. SPAQI 0.924. Alignment accuracy 96%.

2. Cardiac-CFD-nnUNet: coronary seg + flow simulation. SPAQI 0.884. FFR prediction AUC 0.884.

3. Hepato-VesselSeg-FEA: liver + portal vein + hepatic artery + volumetry. SPAQI 0.864. Planning time 18 min vs 124 min conventional.

4. Neuro-TractSeg-GNN: tumour + fibre tracts + approach corridor. SPAQI 0.844. Tract preservation 94%.

5. HN-Mandible-nnUNet-3Dprint: mandible seg + patient-specific cutting guide. SPAQI 0.824. Reconstruction accuracy 1.4 mm.