UNIVERSITÀ DI PAVIA

PHD PROGRAM IN EXPERIMENTAL MEDICINE

38° CICLO

EVOLVING APPLICATIONS OF ARTIFICIAL INTELLIGENCE IN EMERGENCY SURGERY AND NEUROSURGERY

PHD CANDIDATE

DOTT. DANIELE PICCOLO

SUPERVISOR

PROF. LORENZO COBIANCHI

TABLE OF CONTENTS

1. ABSTRACT

Background

Artificial intelligence is reshaping surgical practice, yet its integration faces ethical, educational, and technical barriers. Emergency surgery and neurosurgery present complementary challenges: the former defined by time pressure, incomplete data, and population heterogeneity; the latter by complex imaging interpretation and individualized prognostication. This thesis examines AI across both domains, from ethical governance and professional readiness to clinical prediction models.

Objectives

The aims were threefold: (1) map the ethical landscape and assess professional preparedness for surgical AI; (2) develop and validate AI models for neurosurgical diagnosis and prognosis in hydrocephalus, epilepsy, brain tumors, and glioblastoma; (3) explore emerging methodologies including quantum-enhanced hybrid feature selection and radiomic-transcriptomic associations.

Methods

A Delphi consensus engaged 12 experts using the EFTE method aligned with EU AI Ethics Guidelines. A cross-sectional survey collected responses from 650 surgeons across 71 countries under WSES endorsement. A Super Learner ensemble was trained on cortical thickness from 294 patients for iNPH diagnosis. Quantum-enhanced hybrid feature selection (QAOA, Simulated Bifurcation, Simulated Annealing) combined with gradient boosting machine classification predicted VNS response in 31 epilepsy patients. Interaction modeling examined cortical thickness, tumor grade, and memory in 97 glioma patients. Logistic regression assessed BMI as a predictor of cerebellar mutism in 50 pediatric posterior fossa tumor patients. Elastic Net with leave-one-patient-out cross-validation and linear mixed-effects models tested radiomic-transcriptomic associations in 28 glioblastoma patients from the IvyGAP atlas.

Results

The Delphi study distilled seven ethical requirements, with transparency and accountability as highest priority. The WSES survey found that 69% of surgeons reported AI familiarity yet only 17% provided concordant definitions; perceived importance rose from 3.06 (current) to 3.88 (five-year horizon). The Super Learner achieved AUC 0.843 for iNPH diagnosis, with caudal middle frontal and superior frontal cortical thickness as key discriminators. Gradient boosting machine achieved 77.1% cross-validated accuracy for VNS response prediction; age, total seizures, and time since diagnosis were the leading predictors. A significant three-way interaction between cortical thickness, tumor grade, and education predicted postoperative memory change in glioma patients ($p = 0.035$). BMI was the sole significant predictor of cerebellar mutism ($p = 0.031$, AUC = 0.749). In the radiomic-transcriptomic analysis, only Angiogenesis (R-squared-cv = 0.209) and Inflammatory Response (R-squared-cv = 0.185) demonstrated genuine associations among 24 pathways tested; 21 pathways showed no predictive signal.

Conclusions

This dual-axis investigation reveals that while emergency surgeons increasingly recognize AI's importance, foundational knowledge gaps persist. Neurosurgical AI models demonstrated diagnostic and prognostic value across hydrocephalus, epilepsy, glioma, and glioblastoma, though

sample sizes constrain generalizability. Bridging ethical framework to clinical deployment requires targeted education, prospective multicenter validation, and explainable model design.

ABBREVIATIONS

AD -- Alzheimer's disease

AI -- artificial intelligence

ANCOVA -- analysis of covariance

ARIES -- Artificial Intelligence in Emergency Surgery

ASM -- antiseizure medication

AUC -- area under the curve

BH -- Benjamini-Hochberg

BMI -- body mass index

CHERRIES -- Checklist for Reporting Results of Internet E-Surveys

CSF -- cerebrospinal fluid

CSFTT -- cerebrospinal fluid tap test

CT (IvyGAP) -- cellular tumour (IvyGAP anatomic zone)

CTmvp -- cellular tumour microvascular proliferation (IvyGAP zone)

CTpan -- cellular tumour pseudopalisading cells around necrosis (IvyGAP zone)

DARTEL -- diffeomorphic anatomic registration through exponentiated lie

DESH -- disproportionately enlarged subarachnoid-space hydrocephalus

DKT -- Desikan-Killiany-Tourville

DRF -- distributed random forest

ED -- peritumoral oedema (MRI subcompartment)

ET -- enhancing tumour (MRI subcompartment)

FDR -- false discovery rate

FLAIR -- fluid-attenuated inversion recovery

FPKM -- fragments per kilobase of transcript per million mapped reads

GBM -- gradient boosting machine

GLCM -- grey level co-occurrence matrix

GLM -- generalized linear model

GLSZM -- grey level size zone matrix

GM -- gray matter

HGG -- high-grade glioma

IBSI -- Image Biomarker Standardisation Initiative

IDH -- isocitrate dehydrogenase

ILAE -- International League Against Epilepsy

iNPH -- idiopathic normal pressure hydrocephalus

IT -- infiltrating tumour (IvyGAP zone)

LE -- leading edge (IvyGAP zone)

LGG -- low-grade glioma

LMD -- laser microdissection

LMM -- linear mixed-effects model

LOD -- length of disease (time since diagnosis)

LOPO-CV -- leave-one-patient-out cross-validation

LRT -- likelihood ratio test

MB -- medulloblastoma

MGMT -- O6-methylguanine-DNA methyltransferase

ML -- machine learning

MLP -- multilayer perceptron

MRI -- magnetic resonance imaging

NET -- non-enhancing tumour (MRI subcompartment)

PFS -- posterior fossa syndrome

PPV -- positive predictive value

PS -- Parkinson's spectrum

QAOA -- Quantum Approximate Optimization Algorithm

RF -- random forest

ROC -- receiver operating characteristic

SA -- Simulated Annealing

SB -- Simulated Bifurcation

SD -- standard deviations

SHAP -- Shapley additive explanations

ssGSEA -- single-sample gene set enrichment analysis

SVM -- support vector machine

TUG -- timed up and go test

UPDRS -- unified Parkinson's disease rating scale

VaD -- vascular dementia

VNS -- vagal nerve stimulation

VPS -- ventriculoperitoneal shunt

WJES -- World Journal of Emergency Surgery

WM -- white matter

WSES -- World Society of Emergency Surgery

XGB -- extreme gradient boosting machine

XRT -- extremely randomized trees

2. INTRODUCTION

Neurosurgery stands at the precipice of a technological revolution. The convergence of machine learning (ML), artificial intelligence (AI), and emerging quantum computing technologies promises to fundamentally transform how neurosurgeons diagnose complex pathologies, plan intricate procedures, and predict patient outcomes[1,2]. This transformation represents the most significant paradigm shift in neurosurgical practice since the introduction of medical imaging in the 1970s. The field has evolved from rudimentary computer-assisted systems to sophisticated deep learning algorithms that now demonstrate superhuman performance in specific diagnostic tasks[3]. As we advance into the quantum era, these technologies hold the potential to solve previously intractable optimization problems in surgical planning and unlock new frontiers in personalized neurosurgical care[4,5]. However, the clinical translation of these powerful tools requires careful consideration of methodological rigor, ethical implications, and the fundamental challenge of maintaining the human element in an increasingly digitized medical landscape.

Historical evolution of artificial intelligence in neurosurgery

The application of artificial intelligence in neurosurgery traces its origins to the pioneering robotic procedures of the 1980s, establishing a foundation that would ultimately support today's sophisticated ML applications[6]. The first robotic neurosurgical procedure was performed in 1985 using the Unimation PUMA 560 robotic arm for CT-guided brain biopsy, marking the dawn of computer-assisted neurosurgery[7]. This landmark achievement demonstrated the potential for mechanical precision to augment human surgical skill, with the system compensating for physiological tremor by a factor of 10 and significantly improving targeting accuracy.

The 1990s marked the emergence of actual computer-assisted surgery, facilitated by the development of image-guided systems and stereotactic frameworks. The NeuroMate system, developed by Integrated Surgical Systems beginning in 1987, became the first FDA-approved, commercially available image-guided robotic-assisted system for stereotactic procedures, successfully performing over 1,000 cases[8]. Concurrently, the University of Lausanne developed the Minerva System in 1991, pioneering real-time image guidance that could compensate for brain shift during procedures[9].

The introduction of machine learning algorithms in neurosurgery began in the 1990s, with the first published applications utilizing artificial neural networks (ANNs) for structured database analysis and supervised learning tasks[6]. A particularly notable early study by Grigsby et al. developed simulated neural networks to predict seizure-free outcomes after anterior temporal lobectomy, achieving superior performance compared to discriminant function analysis, with 81.3% accuracy versus 78.5% for optimal outcomes[10].

By the turn of the millennium, well-trained AI algorithms began to consistently outperform traditional clinical approaches in brain tumor diagnosis, tumor segmentation, and surgical risk assessment[6]. The digitalization of healthcare systems in the 2000s significantly enhanced AI system capabilities by making extensive, structured, and unstructured datasets available for analysis, thereby establishing the data foundation necessary for the deep learning revolution that would follow.

The period from 2010 onwards experienced explosive growth in neurosurgical AI applications. Throughout the 2010s, the use of machine learning in neurosurgery expanded dramatically, with over 200 studies involving 6,402 citations identified by 2017, covering presurgical planning, intraoperative guidance, neurophysiological monitoring, and outcome prediction[2]. Recent systematic reviews have identified 153 studies employing various ML approaches for glioma analysis alone, reporting excellent aggregate performance with AUC values of 0.87 ± 0.09 and sensitivity of 0.87 ± 0.10[2,3].

Clinical challenges

Diagnosis

Modern neurosurgery confronts diagnostic challenges of unprecedented complexity, where AI/ML technologies demonstrate transformative potential. Hydrocephalus diagnosis represents one of the most mature and clinically validated AI applications in neurosurgery[11–13]. Recent research has shown that XGBoost models achieve remarkable diagnostic accuracy, with AUC values of 0.988 in training and 0.938 in testing, for the diagnosis of idiopathic normal pressure hydrocephalus (iNPH). These systems utilize automated measurement algorithms for Evans index, callosal angle, and normalized lateral ventricle volume, achieving a cross-validation AUC of 0.983 for fully automated models[12].

Brain tumor classification has emerged as one of the most extensively studied AI applications in neurosurgery, with deep learning models achieving accuracy rates of 99.06% using EfficientNetB2 architectures and 99.83% with ResNet101-CWAM approaches[3]. Multi-class classification systems for glioma, meningioma, and pituitary tumors consistently demonstrate accuracy rates of 97–99% across multiple validation studies[1,3]. The clinical significance of these achievements extends beyond diagnostic accuracy to include real-time histopathologic classification during surgical procedures, potentially revolutionizing intraoperative decision-making.

Epilepsy localization and seizure prediction represent frontiers where AI demonstrates particular promise for personalized medicine. Cleveland Clinic's AI system, utilizing over 7,000 patients and 20 terabytes of EEG data, employs Temporal Graph Convolutional Networks for spatiotemporal analysis of seizure patterns[3]. Deep learning models for seizure prediction achieve 69% mean sensitivity with 27% mean time in warning, outperforming random predictors by 42% across all patients. Perhaps most significantly, AI applications in pediatric epilepsy surgery selection have the potential to reduce referral time from six years to potentially weeks, representing a paradigm shift in pediatric neurological care.

Prognosis

The inherent uncertainty in neurosurgical outcomes creates substantial clinical challenges that ML approaches are uniquely positioned to address. Systematic reviews of ML applications in neurosurgical outcome prediction demonstrate median accuracy rates of 94.5% with AUC values of 0.83, representing a 15% absolute improvement over conventional logistic regression models[1,2]. These improvements prove particularly significant in traumatic brain injury management, where ML models for emergency neurosurgery prediction within 24 hours demonstrate superior performance compared to established prognostic indices.

Cognitive outcome forecasting presents a particularly complex challenge, where AI is demonstrating growing clinical utility[3]. The integration of multimodal data, including neuroimaging,

neuropsychological assessments, and clinical variables, enables more accurate prediction of post-operative cognitive function, facilitating personalized rehabilitation planning and realistic patient counseling regarding expected outcomes[1,3].

Surgical Planning

The complexity of modern neurosurgical procedures demands optimization approaches that exceed human computational capabilities. AI algorithms now identify optimal linear and nonlinear surgical approaches using 3D MRI-based planning with DICOM format compatibility, integrating seamlessly with existing neuronavigation systems[1,3]. The Brainlab Elements platform incorporates automated tumor segmentation using U-net architectures, white matter tractography optimization, and multi-modal image fusion capabilities[3].

Real-time surgical decision support represents an emerging frontier where computer vision technologies demonstrate particular promise. Intraoperative computer vision systems achieve greater than 95% accuracy in surgical instrument recognition and tracking, enabling automated surgical scene analysis and phase recognition with performance comparable to expert surgeons[14]. These systems facilitate the objective assessment of surgical skills and provide real-time feedback for surgical training applications[1,14].

Current applications

Neuroimaging analysis

The application of deep learning to neuroimaging analysis has revolutionized diagnostic capabilities in neurosurgery. U-Net architectures[15] and their advanced variants remain the cornerstone for medical image segmentation, with modern implementations achieving Dice scores exceeding 0.90 for whole tumor segmentation on BraTS datasets[16]. Recent developments include ResUNet, UNet++, UNet3+, and DC-UNet, which incorporate novel connection methods and operations to improve detail presentation and segmentation accuracy[3].

Three-dimensional U-Net extensions, designed explicitly for volumetric brain tumor segmentation, have achieved validation accuracy rates of up to 99.33% with a validation loss of 0.01 on the BraTS 2021 datasets[3]. These architectures address the inherent challenges of processing volumetric medical data while maintaining computational efficiency suitable for clinical deployment.

The integration of Vision Transformers with traditional CNN architectures represents a significant advancement in medical image analysis. TransUNet architectures, which integrate transformer encoders with U-Net frameworks, demonstrate 1.06% average Dice improvement for multi-organ segmentation and 4.30% improvement for pancreatic tumor segmentation compared to nn-UNet[15]. Swin-UNet and U-Net architectures address specific challenges, including “token-flatten” problems and scale-sensitivity issues, through global-local combination approaches[3].

Radiomics

Radiomic approaches enable the extraction of quantitative features from medical images that extend beyond traditional visual interpretation. Modern radiomic pipelines extract over 1000 quantitative features, including shape, intensity, and texture characteristics, enabling comprehensive tissue characterization that correlates with clinical outcomes[17]. NODDI-based radiomic texture features demonstrate particular promise for differentiating glioblastoma from

solitary brain metastases, achieving high diagnostic performance across multiple validation studies[1,3].

Clinical applications of radiomics include predicting IDH mutation status with accuracy rates ranging from 78.8% to 93.8%, determining MGMT methylation status, and forecasting survival outcomes[3]. Multi-institutional validation across different scanner vendors and protocols demonstrates the robustness of radiomic approaches; however, standardization challenges remain significant barriers to widespread clinical adoption[17].

Predictive modeling

The development of predictive models for surgical outcomes represents a mature application of ML in neurosurgery. Machine learning models consistently outperform traditional prognostic indices, with systematic reviews demonstrating median accuracy improvements of 15% over conventional statistical approaches[1,2]. These models integrate diverse data types, including imaging features, clinical variables, and demographic information, to provide comprehensive risk assessment.

Specific applications include emergency neurosurgery prediction within 24 hours for patients with traumatic brain injuries, where ML models identify key predictors, including the Glasgow Coma Score, blood pressure changes, and mydriasis regression. Neurosurgery ICU outcome prediction models identify 17 key predictors, with age, weight, hypertension, and diabetes representing the most critical factors, enabling better resource allocation and informed treatment protocols[1,2].

Computer vision for surgical guidance

Computer vision applications in neurosurgery have advanced from experimental systems to clinically deployed technologies[14]. Frameless stereotactic systems utilizing computer vision-based registration achieve submillimetric target registration errors (TREs) and spatial registration errors (SREs), with AI-driven recalibration systems providing response times under 0.25 seconds for patient movement compensation[1,3].

Real-time tissue classification systems enable automated identification of anatomical structures, pathological tissue, and surgical instruments during procedures. Multi-class segmentation systems achieve Intersection over Union (IoU) scores exceeding 0.85 for instrument segmentation tasks, facilitating enhanced surgical safety through automated detection of critical anatomical structures and potential hazards[14].

Future applications

Quantum computing

Quantum computing represents the next frontier in computational neurosurgery, with applications that promise to solve previously intractable optimization problems[4,5]. Quantum Convolutional Neural Networks (QCNNs) demonstrate significant advantages for medical image analysis, reducing training complexity from $O(n^2)$ to $O(\log(n))$ through quantum parallelism and superposition principles. Hybrid quantum-classical CNN approaches have demonstrated superior performance in brain tumor classification, with studies reporting 97.7% accuracy and a 52-minute processing time on quantum hardware, compared to 90 minutes on classical GPUs[4].

The global quantum computing healthcare market, experiencing a 42.5% compound annual growth rate with projections reaching $503 million by 2028, reflects the growing recognition of quantum computing advantages in medical applications[4]. COVID-19 CT scan classification using quantum

neural networks has demonstrated 2.92% accuracy improvement over classical methods, suggesting broader applicability across neuroimaging modalities[18].

Quantum algorithms excel in solving the complex optimization problems inherent in neurosurgical planning and analysis[5]. Precision radiotherapy planning benefits from quantum computing's ability to simultaneously optimize radiation beam angles and intensities while considering tumor geometry, organ-at-risk constraints, and dose distribution[4,5]. These multi-objective optimization problems, computationally intractable for classical computers, become manageable through quantum approaches.

Brain network analysis represents a particularly promising application for quantum computing. Quantum Brain Networks (QBraiNs) propose quantum graph algorithms for analyzing the human connectome's 86 billion neurons and 242 trillion synapses, enabling comprehensive modeling of brain network dynamics at unprecedented scales. Theoretical frameworks, including AdS/Brain theory, apply concepts from quantum field theory to neural signaling, potentially revolutionizing our understanding of brain function[5].

Neural interface technologies

The convergence of quantum computing with neural interface technologies offers transformative possibilities for neurosurgery[4,5]. Quantum-enhanced neural decoding systems could enable simultaneous processing of thousands of neuronal channels through quantum parallelism, revolutionizing brain-computer interface capabilities[5]. Smart adaptive neuroprosthetics utilizing quantum-enhanced learning algorithms could adapt to changing neural patterns with unprecedented sophistication[4].

Future applications include implantable quantum devices for direct neural signal processing, quantum-enhanced deep brain stimulation with adaptive parameters, and brain-computer interfaces with quantum feedback loops for enhanced control accuracy[4,5]. Quantum simulation of neuroplasticity and learning mechanisms could enable personalized neuromodulation strategies based on individual neural network characteristics[5].

Methodological considerations

The successful integration of AI systems into clinical workflows requires careful consideration of human-computer interaction principles and existing healthcare delivery models[1,19]. Workflow integration challenges include compatibility with existing clinical information systems, training requirements for healthcare professionals, and the need for seamless integration that enhances rather than disrupts clinical care[19].

Regulatory approval processes for AI-enabled medical devices are continually evolving as agencies like the FDA and international bodies develop frameworks for continuous learning systems and quantum-enhanced devices. However, current regulatory frameworks often lag behind technological advancements, making approval processes inadequate for adaptive learning systems and quantum-enabled devices[19]. To address this gap, it is essential to develop comprehensive validation protocols tailored to surgical AI applications, which will require collaboration among regulatory agencies, technology developers, and clinical experts[20]. Such efforts are crucial to ensure patient safety while fostering innovation, with quality management systems needing to integrate AI validation into existing medical device regulations while

accommodating the unique characteristics of adaptive algorithms and quantum computing systems.

Explainability and validation

The clinical deployment of AI systems in neurosurgery faces the fundamental challenge of the "black box" problem, where complex algorithms provide accurate predictions without clinically interpretable explanations[21]. FDA and EU AI Act requirements mandate explainable AI for high-risk healthcare applications, necessitating the development of interpretable decision-making processes that maintain clinical trust while preserving algorithmic performance[19].

Current approaches to explainability include SHAP (Shapley Additive Explanations) for identifying feature importance in complex models and LIME (Local Interpretable Model-agnostic Explanations) for providing local explanations of individual predictions. Semantic transparency initiatives ensure that medical AI systems provide clinically interpretable outputs that align with established medical knowledge and reasoning patterns[21].

The clinical validation of AI systems requires methodological rigor that extends beyond traditional software testing paradigms. Multi-site validation across diverse patient populations and healthcare settings remains the gold standard for demonstrating clinical utility, yet systematic reviews indicate that 85% of current studies lack external validation using independent patient cohorts[1,3].

Bias detection and fairness

The perpetuation of healthcare disparities through biased AI algorithms represents a significant ethical and clinical challenge. Research demonstrates that AI models with the highest accuracy in predicting race and gender also exhibit the most substantial fairness gaps, highlighting the tension between algorithmic performance and equitable healthcare delivery[22].

Systematic bias sources include training data that reflects historical healthcare disparities, deployment contexts that differ from intended use cases, and demographic underrepresentation in model development datasets. Mitigation strategies include federated learning approaches across diverse healthcare institutions, bias-aware algorithm design with fairness constraints, and continuous monitoring systems for deployed AI applications[23].

Artificial intelligence in emergency and trauma surgery

Emergency and trauma surgery presents a distinct set of decision-making demands that differentiate it from elective neurosurgical practice. Surgeons operating in acute settings must formulate diagnostic and therapeutic plans under severe time constraints, frequently with incomplete clinical data, unstable patient physiology, and limited access to advanced imaging or laboratory confirmation. The heterogeneity of the trauma population further compounds the difficulty of applying standardized clinical algorithms to individual patients.

These environmental pressures expose the surgical decision-maker to systematic cognitive biases. Anchoring bias, whereby the initial clinical impression disproportionately shapes subsequent reasoning, is particularly hazardous when early diagnostic information is fragmentary or misleading[24]. Availability heuristic, meaning the tendency for recent or emotionally salient cases to distort probability estimates, has been documented as a contributor to diagnostic error in emergency departments. Action bias, the preference for intervention over observation under uncertainty, can drive unnecessary operative procedures and has been identified as a persistent

challenge in acute surgical practice. The WSES international survey examined in Chapter 3 of this thesis confirmed that surgeons themselves recognize these vulnerabilities: respondents ranked “recent experiences disproportionately affect surgical decision-making” and “decisions must often be made before all relevant data can be retrieved” among the most prominent challenges to clinical judgment[25].

Artificial intelligence offers a potential counterweight to these cognitive limitations. Machine learning-based triage systems can integrate physiological parameters, laboratory values, and imaging findings to generate rapid risk stratification scores that are less susceptible to individual bias[26]. Predictive models for trauma mortality, need for massive transfusion, and likelihood of emergent operative intervention have demonstrated accuracy metrics that equal or surpass conventional scoring systems such as the Injury Severity Score and the Revised Trauma Score[27,28]. Beyond individual patient assessment, AI-driven resource allocation tools have been proposed for optimizing operating room scheduling, blood product utilization, and intensive care unit bed management during mass casualty events[29].

The World Society of Emergency Surgery (WSES) has emerged as a leading professional body engaging with the translational challenges of AI adoption in acute surgical care. Through its endorsement of international surveys and consensus initiatives, WSES has facilitated a structured examination of how emergency surgeons perceive, understand, and envision the integration of algorithmic decision support into their practice[25]. This institutional engagement provides an essential foundation for the readiness assessment presented in Chapter 3 of this thesis.

The unique constraints of emergency surgery, including compressed decision timelines, data incompleteness, population heterogeneity, and the irreversibility of many acute interventions, simultaneously create fertile ground for AI-assisted decision support and impose stringent requirements on algorithm transparency, speed, and robustness. Any clinical AI tool deployed in the emergency setting must function reliably with missing inputs, deliver outputs within clinically actionable timeframes, and communicate uncertainty in a manner that supports rather than supplants surgical judgment[30].

Thesis objectives

This thesis investigates the integration of artificial intelligence into surgical practice along a trajectory that spans from ethical framework to clinical implementation. Rather than treating AI as a monolithic technology to be assessed in isolation, the work examines a series of interconnected questions: what ethical principles should govern surgical AI, whether the surgical community is prepared for its adoption, and how specific AI methodologies perform when applied to concrete diagnostic and prognostic problems in neurosurgery.

Chapter 1 establishes the general context by surveying the evolution of AI in medicine and surgery, the current state of clinical applications in neurosurgery, and the methodological considerations — including explainability, validation, and fairness — that determine whether algorithmic tools can be responsibly translated into clinical practice.

Chapter 2 addresses the ethical dimension through a Delphi consensus study conducted in alignment with the European Union AI Ethics Guidelines. This study engaged an international panel of surgeons and scholars to identify the principal ethical tensions surrounding AI in surgery and to formulate actionable recommendations for responsible development and deployment.

Chapter 3 shifts from normative principles to empirical assessment of surgeon readiness. An international survey endorsed by the World Society of Emergency Surgery (WSES) collected responses from 650 surgeons across 71 countries, characterizing their knowledge of AI, their perceived challenges in clinical decision-making, and the gap between current and anticipated reliance on algorithmic support.

Chapter 4 presents a clinical application of AI to a specific diagnostic problem in neurosurgery. A Super Learner ensemble algorithm was developed and validated for the diagnosis of idiopathic normal pressure hydrocephalus (iNPH) using cortical thickness measurements derived from deep-learning segmentation, achieving a cross-validated area under the receiver operating characteristic curve (AUC) of 0.843.

Chapter 5 reports additional contributions across four distinct lines of investigation: (a) a quantum-enhanced hybrid feature selection approach combined with machine learning for predicting vagal nerve stimulation response in drug-resistant epilepsy; (b) an analysis of tumor grade as a moderator of the relationship between cortical thickness and memory performance following glioma surgery; (c) a predictive model for cerebellar mutism syndrome in posterior fossa tumors incorporating body mass index; and (d) an exploratory radiomic-transcriptomic association study in glioblastoma using the IvyGAP atlas.

Chapter 6 synthesizes the findings of the preceding chapters in a General Discussion, examining how the ethical framework established in Chapter 2 and the readiness data from Chapter 3 inform the interpretation and future deployment of the clinical AI applications developed in Chapters 4 and 5.

Chapter 7 presents the conclusions of the thesis and outlines future research perspectives, identifying the methodological, regulatory, and translational steps required to advance AI from proof-of-concept studies toward integration into routine surgical workflows.

Taken together, these seven chapters constitute a coherent investigation that moves from the formulation of ethical principles, through the assessment of professional readiness, to the development and evaluation of AI tools for specific neurosurgical problems — thereby addressing the full translational arc from concept to clinical application.

3. AI AND SURGERY: ETHICAL DILEMMAS AND OPEN ISSUES

Introduction

Artificial Intelligence (AI) is gaining importance in almost every industrial and service field, including medicine. The number of AI-related biomedical studies is rising exponentially, with almost 19,000 publications in 2020. Also, while in 2014 only AI-based AliveCor's algorithm supporting early detection of atrial fibrillation was approved for clinical use by the FDA, 46 total algorithms were approved four years later, with radiology and cardiology profoundly involved as specialities. Nowadays, approved AI/Machine Learning (ML)-based algorithms totalize 240 in Europe (Conformite Europeene – CE-marked), and 222 in the United States (Food and Drug Administration)[31].

Such algorithms can support decision-making for detection of intracranial haemorrhages or large vessel occlusions in emergent care head Computed Tomography (CT) scans, stroke and traumatic brain injuries, detection of acute findings in abdominal CT scans, liver and lung cancer diagnosis on CT and Magnetic Resonance Imaging (MRI), or X-ray wrist fracture diagnosis, among others.

The practice of surgical specialities, as well as others, involves ethical and moral dilemmas requiring difficult choices. The spirit of AI-powered systems, at their core, is to augment the surgical decision-making processes; hence nowadays, we are asking AI to augment decisions on morally and ethically challenging topics that are still blurry to humans.

With the aim of creating a framework to foster and secure ethical and robust AI, in April 2019, the High-Level Expert Group on AI of the European Commission published the Ethics Guidelines for Trustworthy Artificial Intelligence[32], stating that trustworthy AI should be lawful, namely be respecting all applicable laws and regulations, ethical, meaning compliant with the ethical principles and values, and robust from a technical perspective, but also respecting the social environment. While these guidelines represent an important step forward, it remains unclear how they should be applied to surgical care in clinical settings.

Herein, we summarize the main ethical issues that may arise from the application of AI-related technologies to surgery, using a modified Delphi process to achieve expert consensus regarding the EU Ethics Guidelines for Trustworthy Artificial Intelligence.

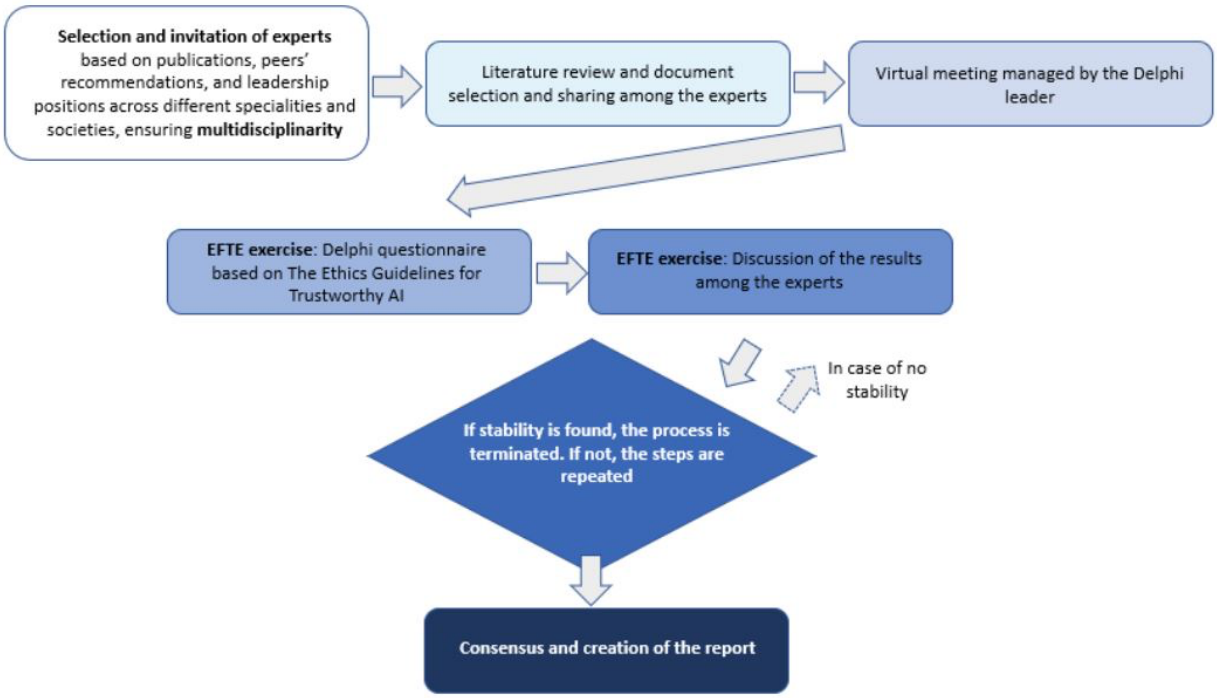
Methodology

The Ethics Guidelines for Trustworthy Artificial Intelligence highlight seven requirements. Employing an EFTE (estimate, feedback, talk, estimate) approach[33], we gathered twelve experts in the fields of academic surgery, radiology, surgical ethics, AI and ML, computer sciences, innovation, strategy, business models, and healthcare policies to apply such requirements to surgical science and achieve consensus regarding ethical dilemmas that may arise in the application of AI in surgery. Like in similar studies[34–37], the selection of members was based on scouting the most recent publications on AI, their peers' recommendations, and their leadership positions across different specialities and societies. Moreover, multidisciplinarity was sought, including technology and innovation experts (from engineering, computer science, and management) with knowledge and expertise in medicine and surgery.

We used the protocol described by Nelms and Porter[33], which included the following steps:

1. Background material was provided to experts for use in formulating opinion judgments. In particular, a literature review was conducted on AI, surgery, and ethics. The initial package included The Ethics Guidelines for Trustworthy Artificial Intelligence documents.
2. Experts met in a virtual conference room. An appointed Delphi manager with expertise in both surgery and technology (JMV) encouraged dialogue among participants. To enable debate, sharing, and the generation of new knowledge, dedicated translation tools were used[38].
3. Each expert was given a Delphi questionnaire, which had to be completed and returned to the Delphi leader. Questions started from the seven requirements as defined by The Ethics Guidelines for Trustworthy Artificial Intelligence.
4. The group openly discussed the feedback results while maintaining the confidentiality of each individual's survey response.
5. Once consensus was achieved, a report was generated to describe the findings. Final results were circulated among all members until everyone agreed. Results are described in the following section and reported in Table 1. The flow is shown in the following Figure 1.

Figure 1. The EFTE process



Findings

Human Agency and Oversight

According to the human agency and oversight requirement, AI has the goal to assist and support decision-making, which remains under human's autonomy. In doing so, AI should be beneficial to the user's organization, promoting civil rights and contributing to social equity.

The surgical literature has already highlighted how AI can support surgical decision-making, but it cannot replace the surgeon's decision completely[39]. Surgeons can benefit from technology and integrate their competencies with it, depending on the situation and the needs. AI-empowered surgical robots controlled remotely can perform operations in hostile environments, like battlefields for wounded soldiers or long space flights[40]. More ethical concerns arise when AI-powered surgical robots exercise some degree of autonomy, either being autonomous for specific tasks (level 2), proposing strategies to be validated (level 3), carrying out the decision-making process (level 4), or achieving full autonomy (level 5)[41]. Therefore, it is essential to generate the best human-machine interface and combination according to the specific situation, safeguarding human agency and oversight.

In such a perspective, issues connected to surgeons' non-technical or "soft" skills emerge. Such skills have proved to be crucial in surgical teamwork, especially in complex contexts like trauma and emergency surgery[42,43] and during challenging times like the recent COVID-19 pandemic[44,45]. Among such skills, creativity stands as essential for surgeons[46,47]. Still, contemporary machines do not seem to have it[48], as they base all their knowledge on technical or "hard" skills.

Technical Robustness and Safety

According to the technical robustness and safety principle, AI should be reliable and developed with a preventive approach to risks caused by the interaction of other agents, both human and artificial. Should agents interact with the system in an adversarial manner, the physical and mental integrity of humans must be prioritized.

The first ethical issue arises in model training. If the training data are robust and the algorithm is appropriate, as time passes, the algorithm becomes more precise and accurate. Therefore, those surgeons and patients who use the early and less educated versions of the model may receive suboptimal decision support, conferring a sort of "guinea pigs" risk[49]. Moreover, AI publication bias toward more positive studies could promote over-optimism[39].

The algorithm and calculations should be disclosed with full transparency if patients and surgeons rely on AI predictions. Therefore explainable AI (XAI) is essential to optimize transparency and build trust among patients and surgeons. However, many times the outputs of a model cannot be explained and behaves as a "Black Box" [50,], meaning that even the models achieving high prediction accuracies cannot explain *a priori* the influence of features and describe the models in human-friendly lay terms. The principle of nonmaleficence applies when training does not accurately represent patients and situations in which the algorithm is applied. Therefore, the "garbage in, garbage out" rule applies[50], with low-quality entry data generating low-quality outcomes.

Technology enthusiasts claim that AI can solve several biases, mistakes, and problems. This may be true when algorithms are trained exclusively on objective features derived from diverse and

balanced training cohorts. However, AI does not consider the human effect. Ethical concerns may recall the movie “Sully”[51], in which the pilot’s integrity was interpreted through the employment of simulation models. Everything looked perfect using the software, but the system did not consider human physiological constrain or realistic uncertainty.

Moreover, AI suffers from the so-called “stupidity”[52] problems. AI systems can be fooled in ways that humans cannot, and AI mistakes are less predictable[53]. For instance, artificially inserting a random object in an image alters the performance of AI-based object detectors on the whole image, not only the replaced object[54].

Technical robustness requires the need to link AI with other technologies to protect its security. Ensuring cybersecurity and cyber resilience stands as a maximum priority[40].

Privacy and data governance

AI impacts privacy. Data governance should ensure the quality and integrity of the data used, its relevance, its access protocols, and the capability to process data while protecting privacy.

Data is the most valuable item in training data-hungry AI algorithms. Ethical issues related to sensitive patient data governance arise, and they are closely connected to modern ethical approaches to patient data. One dilemma concerns the transfer of data from the hospital to the manufacturer. In several countries, most hospitals are public organizations funded by citizens. The inherent risk is that data is transferred for free to private entities (i.e., manufacturers). Once public entities purchase new surgical equipment, they will pay for something created starting from the data they generated. The business model risks are very similar to what happens in the research publication field, in which scholars produce knowledge while being paid by their institutions, which cannot access such published knowledge unless they pay a subscription or open access fees.

Transparency

Transparency must be granted to AI training data, algorithms, and business models to ensure traceability, explainability, and open communication.

Once relying on AI support in surgical decision-making, clinicians must be aware of the state-of-the-art of training model, especially in preventive medicine. Once surgeons use an AI-based tool, they should be aware of the motivations and state of data of the algorithm at the moment of the evaluation, and before deploying an AI-based tool in clinical practices, surgeons should evaluate the fitness of the model for his/her set of patients. To this end, transparent reporting of data used during models development would allow identification of eventual domain shifts between the patients' population used during training and testing of the AI and the patients' population of interest.

Diversity, non-discrimination, and fairness

According to the diversity, non-discrimination, and fairness requirement, AI should enable inclusion and diversity throughout its entire life cycle, ensuring stakeholders' participation, equal access through inclusive design processes, and equal treatment.

Peer-reviewed literature has highlighted the danger of discrimination cases concerning gender and races due to biased data[55]. In particular, the use of Deep Neural Network algorithms, does not consent explicability and may inherit bias from data, which are later translated into biased information which affect decision-making. This stands as an open topic for all the interdisciplinary

scientific communities engaged in AI development[56]. A specific communication of the Council of Europe stated that the most relevant, existing legal tools to mitigate the risks of AI-driven discrimination are non-discrimination and data protection laws[57]. Also, AI opens the door to new types of discrimination that escapes these laws. For example, model overfitting can cause erroneous predictions leading to poor decisions.

Moreover, AI can be expensive. The cost and accessibility of training data could disproportionately hinder developing countries.

AI relies on data, and technology can create wealth through data availability. Connecting to the goal of fostering social equity, an ethical dilemma arises: should those who generate such data (e.g., the patients) be compensated financially? Interventions are expensive, no matter if the operation is billed to the patient, the health insurance, or incurred by a National Health System. Manufacturers can use data generated during surgery to create and profit from higher-performing surgical instruments. Still, no reimbursement will be granted to those who paid for the operation.

Moreover, machines do not take into consideration the tailored approach to the patient, namely the patient's preferences and background, family's wishes, cultural and religious regards, and emotional and ethical implications[50,58]. In theory, a well-trained algorithm will always recommend the best clinical solution, for instance, prescribing glossectomy for an individual with tongue cancer no matter if the patient is a chef or a culinary expert who values the quality of life provided by the sense of taste[50]. Therefore, patient-surgeon synergies[39], shared-decision making[50,59], and co-production dynamics[60] can be more challenging when machine interaction is involved, as only the most convenient clinical choice will be taken into account, regardless of the patient's preferences.

Societal and environmental well-being

According to the United Nations Social Development Goals[61], AI should encourage sustainability and ecological responsibility.

Ethical concerns about medical education arise. Many clinical disciplines, such as radiology, are poised to change profoundly, as do the tasks performed by medical doctors. Understanding what AI can do better than humans and the future clinical applications should be reflected in educational curricula for doctors-to-be.

There is a need to rethink the new surgeons' skillset, understanding the new technical and non-technical skills, which may differ from the current ones[58].

Healthcare professionals skilled in the domain of AI/ML are needed to lead the deployment and adaptation of AI-based tools into clinical scenarios, securing their implementation as well as its surveillance over time. Moreover, AI may change the role of other professionals working with surgeons, e.g., anesthesiologists. New paradigms in education also involve other clinicians[48].

Another fascinating issue related to education is whether ethics can be taught to AI-empowered robots. While ethical dilemmas rarely lead to only one answer, ethical reasoning can rely on critical decision trees. It may be useful to represent ethical reasoning in a learning model.

Accountability

According to the accountability requirement, adequate mechanisms should be put in place to ensure responsibility and accountability for the use of AI and its outcomes, both before and after its development, deployment, and use.

Ethical concerns arise in the measurement of surgical outcomes. Liability issues emerge, with the need to understand who should be blamed or rewarded for the surgical outcome and to what extent: the surgeon/operator, the manufacturer, those in charge of algorithm maintainance[40], hospital authorities, or a combination thereof. New regulations should establish precedents for handling such cases.

Results are summarized in the following Table 1.

Conclusions

Artificial Intelligence is evolving from being a futuristic promise into a promise of a reliable and helpful tool for clinical use. AI-based devices have been proposed and approved in several disciplines, still not yet in surgery, to assist rote, repetitive tasks. Rapid technology development may further expand clinical AI applications, including surgery.

In such a rapidly evolving scenario, ethical concerns and open questions arise. Ethical dilemmas should be addressed in the early phases of technology design, development, and adoption. AI algorithms must be developed and implemented with the highest safety, privacy, transparency, and accountability standards. In addition, other questions should be answered to reach lawful, ethical, and robust solutions: should we ask AI to address issues that remain controversial to humans? AI could enable inclusion and diversity, ensuring equal access through inclusive design processes and equal treatment; what mechanisms should be implemented to avoid discrimination?

Even when following the existing laws and best practices, diversity, equity, non-discrimination, and fairness cannot be ensured a priori, and AI-specific risks can arise, even when applied to surgery.

Our paper highlights some open questions which, to us, should offer “food for thought” to the surgical academic and practice debate. A multidisciplinary perspective, including surgeons, scientists, developers, and policymakers, should be employed to address the open issues, ensuring the right balance between the unique advantages and opportunities offered by the new technologies and the respect of the ethical principles of a patient-centric perspective.

Table 1. Results

Requirement	Definition	Topics from the surgical practice
Human agency and oversight	AI should support human autonomy and decision-making. AI should act as an enabler to a democratic and equitable society by supporting the user's agency and foster fundamental rights and allow for human oversight.	• Surgical decision-making • Optimal human-machine interface according to the specific situation. • Non-technical skills and creativity.
Technical Robustness and safety	AI should be reliable and developed with a preventive approach to risks caused by the presence of other agents (human and artificial) that may interact with the system in an adversarial manner, ensuring at the same time the physical and mental integrity of humans.	• Training model risks • Black box issues • "Captain Sully" risks. • Artificial stupidity • Cyber security and cyber resilience
Privacy and data governance	AI can impact privacy. Data governance should ensure the quality and integrity of the data used, its relevance, its access protocols, and the capability to process data protecting the privacy.	• New privacy issues • Business model in data management
Transparency	Transparency must be granted to the elements relevant to AI like the data, the system, and the business models, ensuring traceability, explainability, and communication.	• State-of-the-art of the training model
Diversity, non-discrimination, and fairness	AI should enable inclusion and diversity throughout its entire life cycle, ensuring stakeholders' participation, with equal access through inclusive design processes as well as equal treatment.	• Overfitting • Access to the technology from developing countries • Potential rewards of data availability • Shared decision-making, co-production, and tailored approaches dynamics
Societal and environmental well-being	Other sentient beings and the environment represent stakeholders throughout the AI's life cycle, encouraging sustainability and ecological responsibility, also according to the SDGs.	• Surgical education • Surgical skillset
Accountability	Adequate mechanisms should be put in place to ensure responsibility and accountability for AI and its outcomes, both before and after their development, deployment, and use.	• Measurement • Liabilities and rewards

4. Surgeon's Perspectives on Artificial Intelligence in Emergency Surgery

Introduction

The ethical and governance framework established in the preceding chapter delineated the principles that should guide the deployment of artificial intelligence (AI) in surgical practice. However, principles alone are insufficient: their translation into clinical reality depends on whether the surgical workforce possesses the knowledge, attitudes, and institutional conditions necessary to adopt AI-enabled tools responsibly. Understanding how surgeons perceive these technologies is therefore a prerequisite for any implementation strategy.

Emergency and trauma surgery constitutes a domain in which clinical decision-making is particularly demanding. Time constraints, incomplete data, physiological instability, and population heterogeneity compel surgeons to rely on heuristics that, while often effective, expose the decision-maker to systematic cognitive biases[62,63]. AI-based predictive analytics have been proposed as a means to augment these decisions by integrating high-volume data into risk stratification and outcome prediction models[64,65]. Yet the degree to which emergency surgeons understand, trust, or are prepared to use such tools has received limited empirical investigation.

The Artificial Intelligence in Emergency Surgery (ARIES) survey, endorsed by the World Society of Emergency Surgery (WSES), represented an early effort to characterize knowledge, attitudes, and practices regarding AI among international acute care surgeons[66]. Building on the ARIES findings, which suggested that interest in AI was growing but practical adoption remained limited, the present study was designed to assess surgeons' perceptions of AI as a clinical decision-making aid with greater granularity. Specifically, the investigation aimed to evaluate (1) the relative importance that emergency surgeons assign to AI compared with other decision-support tools; (2) the extent to which surgeons perceive AI as relevant to current clinical challenges; (3) the actual level of knowledge regarding AI among this population; and (4) whether geographic or institutional factors modulate these perceptions.

Methods

Study design and setting

An exploratory, population-based cross-sectional survey was conducted among the international trauma and emergency surgery community. The study was endorsed by WSES and followed the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). A steering committee comprising a multidisciplinary panel of academics and practitioners in the fields of trauma and emergency surgery, healthcare management, innovation, business and medical ethics, information technology, law, and organization science was appointed within the WSES to oversee the initiative. No Institutional Review Board approval was required, as

the study involved a voluntary survey of professional opinions without patient data collection. A peer-reviewed research protocol was published prior to data collection.

Questionnaire structure

The online questionnaire was generated in English through Google Forms and was organized into four thematic groups. The first group collected demographic information: gender, years of experience in trauma and emergency surgery, type of institution (academic versus non-academic), country of practice, professional role, and membership in a formally established emergency surgery team. The second group assessed the perceived importance of 11 clinical decision-making facilitators using a 5-point Likert scale (1 = not suitable, 5 = very suitable). The third group evaluated 13 perceived challenges in surgical decision-making, also on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree). The fourth group assessed AI-specific knowledge and attitudes, including a binary familiarity question (yes/no), an open-ended question requesting the respondent's understanding of AI applied to surgery, the perceived importance of AI at the present time and on a five-year horizon (both on 5-point Likert scales), and five items concerning the perceived goals and benefits of AI in surgical decision-making.

AI knowledge assessment

Responses to the open-ended question on AI understanding were manually coded by two researchers (LC and FDM). Each statement was rated as concordant, discordant, or inconclusive. Concordant definitions were those that conveyed the capability of a machine to mimic human intelligence or expressed the aims and potential of AI applied to surgical practice or its technological functioning. Discordant definitions reflected an incorrect understanding or the absence of any substantive knowledge.

Statistical analysis and survey period

Descriptive statistics were computed using R (RStudio 2022.07.0+548, "Spotted Wakerobin" Release). The survey was made available at the end of November 2021 and remained open until mid-August 2022. All 917 WSES members received an e-mail invitation; the initiative was additionally promoted through the society's website and Twitter account. Four e-mail reminders were sent. The response rate was approximately 70%.

Results

Participants

A total of 650 surgeons responded to the questionnaire. Participants were located on five continents and hailed from 71 different nations. The geographic distribution was uneven: Europe accounted for 477 respondents (73%), followed by Asia (85, 13%), the Americas (62, 10%), Africa (22, 3%), and Oceania (4, 1%). Italy was the most represented country (251, 39%), followed by Greece (45, 7%), Spain and the United Kingdom (37 each, 6%), and the United States (26, 4%). The sample comprised 531 males (81.69%), 118 females (18.15%), and one respondent who preferred not to disclose gender (0.15%). Mean years of experience in the field were 12.32 (SD 8.42, range 1–36). The majority of participants were affiliated with academic institutions (499, 76.77%). Senior consultants constituted the largest professional group (233, 35.85%), followed by board-certified surgeons (179, 27.54%),

residents (124, 19.08%), and division chiefs or department heads (114, 17.54%). A total of 540 respondents (83.08%) declared membership in a formally established emergency surgery team.

Clinical decision-making facilitators

Respondents ranked 11 decision-making facilitators by perceived importance. Training received the highest mean score (4.40, SD 0.78), followed by clinical guidelines and cases (4.25, SD 0.82), and multidisciplinary committees and meetings (4.19, SD 0.85). Machine Learning and Artificial Intelligence ranked 10th of 11 items (mean 3.56, SD 1.07), surpassing only regression modeling and calculations (3.49, SD 1.01). The complete ranking is reported in Table 1.

Table 1. Perceived importance of clinical decision-making facilitators (5-point Likert scale; 1 = not suitable, 5 = very suitable)

Rank	Item	Mean	SD
1	Training	4.40	0.78
2	Clinical guidelines and cases	4.25	0.82
3	Multidisciplinary committees and meetings	4.19	0.85
4	Time spent to engage patients	4.10	0.89
5	Networking and international experiences	4.10	0.88
6	Non-technical skills	3.98	1.02
7	Mobile electronic medical records and online tools, including telemedicine	3.96	1.02
8	Publications	3.95	0.94
9	Risk stratification by additive scores using static variable thresholds	3.87	0.91
10	Machine Learning and Artificial Intelligence	3.56	1.07
11	Regression modeling and calculations	3.49	1.01

Of particular note, the AI/ML item recorded the highest standard deviation of all 11 facilitators (1.07), indicating substantial polarization within the respondent population. The three most traditional elements — training, guidelines, and multidisciplinary teamwork — all exceeded a mean of 4.0, whereas the two most computationally oriented items (AI/ML and regression modeling) occupied the two lowest positions. No significant differences emerged when responses were stratified by institution type (academic versus non-academic) or professional role.

Challenges in clinical decision-making

The 13 challenge items revealed that the perceived misalignment between the clinical scenario as it presented and the surgeon's independent assessment ranked highest (mean 3.61, SD 0.95). Incomplete clinical data (3.54, SD 1.00), the use of aggregate rather than personalized outcome data (3.51, SD 1.06), and the disproportionate influence of recent operative experiences on decision-making (3.50, SD 0.98) followed closely. The item "Digital technologies (e.g., artificial intelligence) support how I take clinical decisions" ranked lowest of all 13 items (mean 3.10, SD 1.14). The complete ranking is reported in Table 2.

Table 2. Perceived challenges in clinical decision-making (5-point Likert scale; 1 = strongly disagree, 5 = strongly agree)

Rank	Item	Mean	SD
1	Misalignment between clinical scenario and independent assessment	3.61	0.95
2	Data are often incomplete	3.54	1.00
3	Patients informed of outcomes from aggregate populations without personalized adjustment	3.51	1.06
4	Recent experiences disproportionately affect surgical decision-making	3.50	0.98
5	Errors and mistakes are likely all along the way	3.49	1.09
6	Decisions must often be made before all relevant data can be retrieved	3.44	1.10
7	Potential outcomes predicted using personal beliefs rather than evidence-based guidelines	3.38	1.11
8	In-house calls happen often	3.35	1.00
9	Too complicated to form a complete differential diagnosis list	3.34	1.04
10	Tendency toward action when inaction may be preferable	3.28	1.07
11	Weaknesses and failures disproportionately perceived to affect peers	3.17	1.01
12	Too complicated to recognize strengths and limitations of available tests	3.12	1.02
13	Digital technologies (AI) support how I take clinical decisions	3.10	1.14

As with the facilitators ranking, the AI-related item recorded the highest standard deviation among all 13 challenges (1.14), reinforcing the pattern of divergent opinion.

Knowledge and understanding of AI

When asked whether they were familiar with the terms Artificial Intelligence and Machine Learning, 451 of 650 participants (69%) responded affirmatively, while 199 (31%) indicated that they were not. However, qualitative analysis of the open-ended definitions revealed a substantial gap between self-reported familiarity and actual comprehension. Only 112 respondents (17%) provided a definition rated as concordant with established characterizations of AI — that is, one that conveyed the capability of a machine to mimic human cognitive functions or articulated the aims and technological potential of AI applied to surgery. An additional 178 respondents (27%) produced inconclusive definitions that captured partial aspects of the phenomenon without meeting the concordance threshold. The remaining 360 respondents (55%) provided definitions rated as discordant, indicating either a fundamental misunderstanding or an absence of substantive knowledge regarding AI.

Among the technologies that respondents associated with AI, clinical decision-making support was the most frequently mentioned (409; 63%), followed by big data (130; 20%), new technologies in general (107; 16%), learning and training (57; 9%), robotics (22; 3%), simulations (16; 2%), and virtual or augmented realities (8; 1%). A majority of respondents (366; 56%) did not mention any additional technologies linked to AI.

Perceived current and future importance

Participants rated the perceived importance of AI-based tools for clinical decision-making at the present time with a mean of 3.06 (SD 1.10), corresponding to a neutral position on the 5-point scale. When asked to project the importance of the same tools on a five-year horizon, the mean rose to 3.88 (SD 0.97), representing a shift to above-average perceived relevance. This difference was statistically significant ($p < 0.001$) and indicates that, although emergency surgeons do not regard AI as a central decision-making resource at present, they anticipate substantial growth in its relevance within the near future.

Geographic variation

Country-level analysis of AI perception as a decision-making facilitator revealed marked heterogeneity. Selected results are presented in Table 3.

Table 3. Perceived importance of Machine Learning and Artificial Intelligence as a decision-making facilitator, by country (5-point Likert scale). Countries with fewer than 5 respondents excluded

Rank	Country	n	Mean	SD
1	Argentina	6	4.50	0.548
2	Saudi Arabia	7	4.43	1.13
3	Brazil	13	4.38	0.768
4	Malaysia	15	4.13	1.30
5	India	8	4.12	0.835
6	France	15	4.07	0.884
7	United States	24	3.92	1.21
8	Ukraine	11	3.91	0.944
9	Bulgaria	9	3.89	0.928
10	Belarus	6	3.83	0.753
11	Spain	37	3.68	1.03
12	Romania	9	3.56	1.13
13	Italy	251	3.38	1.10
14	Greece	45	3.42	1.14
15	United Kingdom	37	3.35	1.09
16	Portugal	5	3.40	0.548
17	Japan	5	3.40	0.548
18	Thailand	6	2.67	0.816
19	Canada	5	2.60	0.548
20	Switzerland	5	2.60	0.894

Argentina (mean 4.50, SD 0.548), Saudi Arabia (4.43, SD 1.13), and Brazil (4.38, SD 0.768) displayed the most favorable perceptions, while Canada (2.60, SD 0.548) and Switzerland (2.60, SD 0.894) recorded the lowest scores. The pattern suggests a tendency for respondents from emerging-economy nations to express greater enthusiasm for AI-based decision support relative to those from high-income countries with well-established surgical infrastructure.

Discussion

The results of this international survey provide an empirical foundation for understanding the surgical community's readiness to adopt AI-enabled clinical decision support in emergency settings. Three principal observations merit detailed discussion.

Polarization rather than uniform resistance. The most striking quantitative finding is not the modest mean score assigned to AI as a decision-making facilitator (3.56), but rather its standard deviation (1.07) — the highest among all 11 items assessed. This metric reflects a community divided between technology enthusiasts who perceive substantial promise in algorithmic support and skeptics who regard such tools as peripheral or unreliable. The same pattern was replicated in the challenges section, where the AI item again registered the highest standard deviation (1.14) alongside the lowest mean (3.10). This polarization is consistent with the technology adoption literature, which characterizes early phases of innovation diffusion by a bimodal distribution of attitudes among potential adopters. Rather than indicating blanket resistance, the findings suggest that targeted educational interventions could shift the distribution by addressing the knowledge deficits that appear to underpin skepticism.

Familiarity without comprehension. Perhaps the most consequential finding concerns the disjunction between self-reported familiarity and verified understanding. While 69% of respondents declared familiarity with AI and ML terminology, only 17% provided a definition that met the concordance criterion. A further 27% produced incomplete characterizations, and 55% either misunderstood or could not articulate the concept. This "familiarity without comprehension" phenomenon has significant implications for both clinical deployment and the ethical governance framework outlined in Chapter 2. The Delphi consensus presented in the preceding chapter emphasized transparency, explainability, and informed human oversight as foundational requirements for trustworthy surgical AI. These requirements presuppose that the human decision-maker possesses sufficient technical literacy to evaluate algorithmic outputs critically — a presupposition that the present data challenge directly. If the majority of emergency surgeons cannot accurately define AI, the prospect of their exercising meaningful oversight over AI-generated recommendations must be regarded with caution.

AI as neither priority nor perceived threat. The dual ranking of AI — 10th of 11 as a facilitator, 13th of 13 as a challenge — reveals that emergency surgeons currently view AI not as a source of professional concern or competitive anxiety but rather as a tool of marginal practical relevance. This interpretation is reinforced by the significant gap between current perceived importance (3.06) and projected five-year importance (3.88). Surgeons acknowledge that AI will grow in relevance, yet they do not perceive it as immediately useful or disruptive. This temporal dissociation between acknowledged trajectory and present engagement represents a window for proactive education: the community expects AI to arrive but has not yet prepared for its integration.

Comparison with ARIES findings. The ARIES survey[66] documented that interest in AI was prevalent among emergency surgeons but that practical knowledge and clinical application remained nascent. The present findings corroborate and extend this observation. Where ARIES characterized attitudes broadly, the current study quantifies both the magnitude of

the knowledge gap (17% concordant understanding) and the structural position of AI within the decision-making hierarchy (consistently ranked below all traditional facilitators). Together, the two surveys delineate a trajectory from general awareness toward a more granular assessment of readiness barriers.

Geographic and institutional considerations. The substantial country-level variation — from Argentina (4.50) to Switzerland (2.60) — resists a simple explanation. Several interacting factors may contribute: availability of AI-related educational programming, baseline technological infrastructure, cultural attitudes toward innovation, and the influence of local surgical leaders and technology advocates. Emerging-economy nations tended to express greater enthusiasm, possibly reflecting the perceived potential of AI to bridge resource gaps that are less acute in well-resourced settings. These geographic disparities suggest that implementation strategies must be locally tailored rather than globally uniform.

Implications for the thesis narrative. Chapter 2 established a normative ethical framework for surgical AI, grounded in expert consensus around seven requirements for trustworthy AI. The present chapter reveals that the surgical community to which that framework would apply is, at present, not equipped to implement it. The gap between the ethical standards articulated through Delphi deliberation and the empirical knowledge base measured through this survey constitutes the central translational challenge that the subsequent chapters of this thesis address through concrete clinical applications: AI-enhanced diagnostics in normal pressure hydrocephalus (Chapter 4) and quantum-enhanced predictive modeling for drug-resistant epilepsy (Chapter 5).

Conclusions

Emergency surgeons across 71 countries acknowledge that AI will assume growing importance in clinical decision-making over the coming years, yet the current level of foundational knowledge regarding these technologies remains insufficient. Only 17% of respondents demonstrated a concordant understanding of AI, despite 69% reporting familiarity with the terminology. AI ranked among the lowest-valued decision aids and was the least endorsed clinical challenge, suggesting perceived irrelevance rather than active opposition. Education and training programs — delivered through academic societies such as WSES and integrated into surgical curricula — represent prerequisites for the safe and effective adoption of AI-enabled tools in emergency practice. The gap between perceived future importance and present understanding constitutes both a challenge for the profession and an opportunity for targeted intervention.

5. THE SUPER LEARNER ALGORITHM: ENHANCING iNPH DIAGNOSIS WITH AI-ENHANCED CORTICAL ANALYSIS

Background

Idiopathic normal-pressure hydrocephalus (iNPH) is a neurological condition associated with ventricles enlargement and unchanged cerebrospinal fluid (CSF) pressure, along with cognitive impairment, gait disturbance, and urinary dysfunction. The prevalence of iNPH among adults aged 65 years or older is reported at 3.7%. Remarkably, this prevalence substantially increases to 8.9% in individuals aged 80 years and above[67].

While optimal treatment for idiopathic iNPH is reasonably known, diagnosis remains a topic of debate and scientific research[68]. No pathognomonic sign, laboratory results, or imaging findings are sufficient to establish a diagnosis of iNPH, which requires convergent evidence from clinical history, physical examination, diagnostic procedures, and brain imaging[69].

The most significant brain imaging finding in iNPH is ventriculomegaly[70,71]. However, ventricular enlargement can also occur in neurodegenerative diseases and healthy elderly individuals[72]. Ventriculomegaly, dilation of the Sylvian fissure, and narrowing of the high parietal convexity or midline subarachnoid spaces define disproportionately enlarged subarachnoid-space hydrocephalus (DESH)[73]. DESH findings have high positive but low negative predictive values for iNPH[68]. In addition, some elderly individuals have findings similar to DESH on MRI, even though they are asymptomatic[74].

Other imaging methods rather than structural MRI are suitable for iNPH diagnoses, such as diffusion MRI[75–77], perfusion MRI[76], phase-contrast MRI[78], functional MRI[79], proton magnetic resonance spectroscopy[80], and nuclear medicine diagnostic methods[81,82]. However, these diagnostic methods are still a current topic of debate in the literature, and their diagnostic value is still unproven[68].

Several studies on iNPH and CSF tap test (CSFTT) have demonstrated this diagnostic procedure's high Positive Predictive Value (PPV) for successful shunt surgery[83]. However, shunt surgery is not consistently effective even with a positive CSFTT[83–85]. Instead, a negative CSFTT cannot reject the possibility that the symptoms will improve by a shunt intervention due to low negative predictive value and sensitivity[68,83,86–88].

iNPH symptoms overlap with other diseases, such as Alzheimer's disease (AD), Parkinson's spectrum (PS) disorder, and vascular dementia (VaD), which can also concurrently affect the patient, contributing to the symptomatology[89,90]. Various tests, biomarkers, and neuroimaging techniques can be used to diagnose such diseases and comorbidities[68,91]. However, their diagnostic accuracy is still suboptimal[68].

Quantitative computational methods for extracting cortical thickness measures have become available in the last decade as publicly available software packages. They only require T1-weighted brain MR images regularly available to diagnose a possible iNPH patient.

Reportedly, distinctive cortical thinning patterns are associated with iNPH [92–94], AD[95–97], PS disorder[98], and VaD[99].

The use of Artificial Intelligence (AI) and Machine Learning (ML)-based solutions in surgery has shown potential in aiding diagnosis, predicting surgical outcomes, and optimizing surgical procedures. Recent studies have explored AI and ML applications in iNPH diagnosis and treatment prediction[13]. Sotoudeh et al. reported that using only clinical data, the Random Forest (RF) model achieved an area under the curve (AUC) of 0.71 and an accuracy of 0.70. In contrast, the Support Vector Machine (SVM) model, with the addition of Radiomics analysis, achieved an AUC of 0.80 and an accuracy of 0.76[100]. Conversely, Mládek et al., using the lumbar infusion test, found eXtreme Gradient Boosting (XGB) to be the top-performing ML algorithm with an AUC of 0.891, outperforming traditional manual classification[101].

In this retrospective study, we analyzed the differences in cortical thickness among all possible iNPH patients referred to our clinic between January 2015 and December 2022 and evaluated its relationship with clinical assessments. Our hypothesis centered on identifying distinct cortical thinning patterns in patients with negative CSFTT outcomes and those unresponsive to shunt surgery, potentially elucidating underlying comorbidities. To achieve this, we employed multiple ML algorithms, including Distributed Random Forest (DRF), Extremely Randomized Trees (XRT), Generalized Linear Model (GLM) with Regularization, Gradient Boosting Machines (GBM), eXtreme Gradient Boosting machines (XGB), and a fully connected multi-layer Artificial Neural Network (Deep Learning machine). These algorithms were strategically integrated into a Super Learner ensemble approach to harness their collective predictive power[102].

Methods

Study populations

All patients referred to our clinic between January 2015 and December 2022, who met the diagnostic criteria for possible iNPH[68], underwent a comprehensive brain MRI and clinical assessment before CSFTT, at 24 hours, and seven days post-CSFTT.

Patients presenting with CSF pressure below 200 mmH2O, standard CSF composition, and either the presence of neuroimaging indicators consistent with DESH along with gait disturbance or a clinical improvement following CSFTT were classified as probable iNPH[68] and were recommended for surgical intervention.

This study was approved by the local ethics committee (RIF. Prot IRB-DMED: 090/2021), and informed consent was obtained from each participant.

Clinical evaluation

A total of 294 possible iNPH patients underwent neuropsychological, physiotherapeutic, and neurological examinations. The same neurologist assessed the patients before and after the test to eliminate inter-observer discrepancies. Motor and gait functions were evaluated using the Timed Up and Go test (TUG) and the Unified Parkinson's Disease Rating Scale

(UPDRS). An improvement of at least 10% in the TUG after CSF drainage was considered significant[68].

Cognitive function was evaluated using Mini-Mental State Examination (MMSE) and Frontal Assessment Battery (FAB). Urinary incontinence was documented as a binary variable (present/absent) based on clinical history and physical examination.

CSF dynamic evaluation

A Tuohy spinal needle was connected to a Möller Medical LiquoGuard 7 (Fulda, Germany) pressure monitor and fluid infusion system with the subject placed in the lateral recumbent position. After measuring baseline CSF pressure (P_{start}), saline solution was infused at a constant rate of 1.5 mL/min until reaching a stable pressure plateau ($P_{plateau}$). The resistance to outflow (R_{out}) was calculated as the difference between $P_{plateau}$ and P_{start} divided by the infusion rate, and a value higher than 12 mmHg/mL/min was considered significant.

After the infusion test, CSF was drained until the pressure returned to baseline. Then, a tap test was performed, draining up to 50 ml of CSF or until the pressure reached 0 mmH2O.

MRI data acquisition

Brain imaging was executed on a Philips Achieva 3T whole-body scanner (Best, Netherlands) equipped with a SENSE-Head eight-channel head coil. The neuroradiological protocol included volumetric T1-weighted, T2-weighted, gradient-echo steady-state, phase contrast, diffusion-weighted, and fluid-attenuated inversion recovery (FLAIR) imaging. On these sequences, Evans' index (EI), callosal angle (CA), and presence of Disproportionately Enlarged Subarachnoid-space Hydrocephalus (DESH) were recorded[68].

MRI data preprocessing

T1-weighted MRIs were processed with the Computational Analysis Toolbox (CAT, version 12.8.2) within SPM12 using MATLAB (version 2019a). All images were normalized using an affine followed by non-linear registration, corrected for bias field inhomogeneity, and followed by unified segmentation[103]. The DARTEL algorithm was used to spatially normalize the segmented scans into a standard MNI space. All segmented, modulated, and normalized GM and WM images were smoothed using 8-mm full-width-half-maximum Gaussian smoothing. Cortical thickness was measured according to the Desikan-Killiany-Tourville (DKT) atlas using the projection-based thickness method[104]. All data were visually inspected in each step for potential artifacts and inaccuracies in the surface reconstructions.

Surgical technique

All probable-iNPH patients underwent ventricular-peritoneal shunt (VPS) implantation to the right ventricular frontal horn with an Integra LifeSciences CODMAN® HAKIM® Programmable Valve (Princeton, USA). All patients were discharged from the hospital within 4 days after VPS implantation and repeated the clinical evaluation one month, three months, six months, and one year later. There were no post-operative complications.

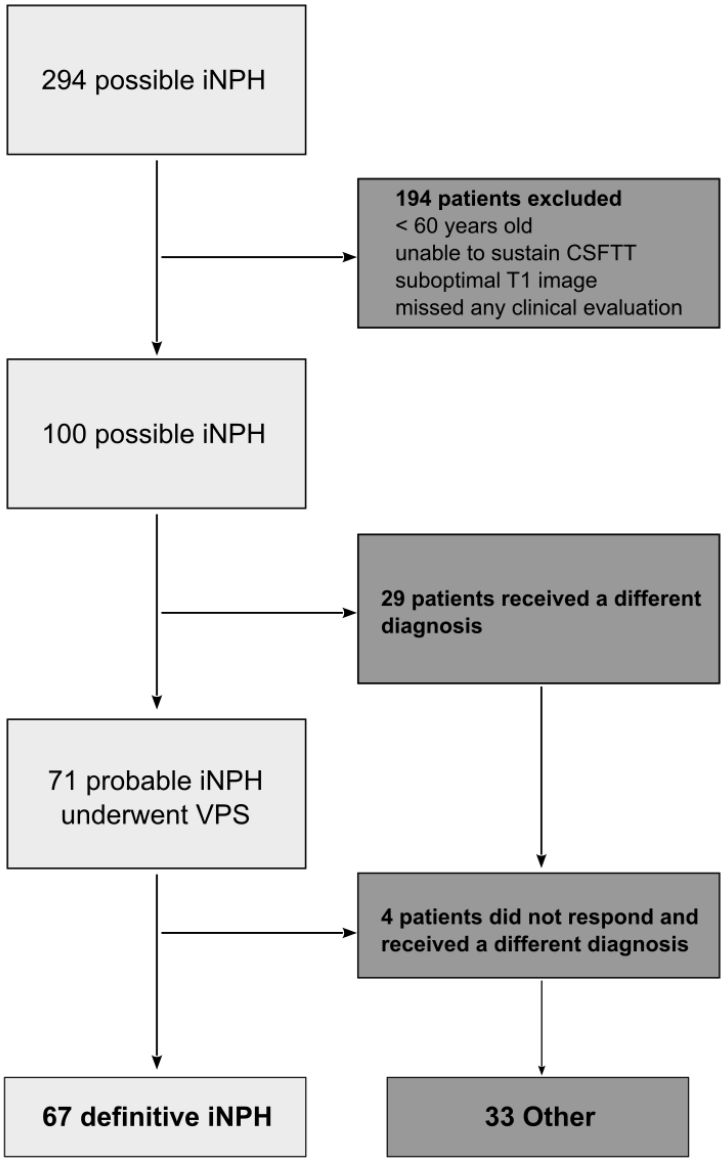
Study groups

Patients under 60 years old at the time of surgery, unable to sustain a CSFTT, missed any clinical evaluation or declined to participate were excluded from the study. The final

population comprised 100 patients. Of these, 71 underwent CSF shunt surgery: 67 responded positively to the surgery and were thus confirmed as iNPH patients, while 4 did not show improvement and, following confirmation of a functional shunt, received alternative diagnoses. The remaining 29 were not indicated for surgery and ultimately received other diagnoses. For this study, the participants were categorized into two groups: iNPH, consisting of 67 patients, and 'Other', comprising 33 patients (Figure 1).

Patients were classified as shunt-responsive or non-responsive based on a composite clinical evaluation incorporating all three components of the iNPH triad.

Figure 1 – STROBE flow diagram of the study.



Statistical analyses

Continuous variables were summarized as means with standard deviations (SD), while categorical variables were reported as frequencies and percentages. Group differences were assessed using effect size measures, including Cohen's D and the standardized mean difference. The Wilcoxon rank-sum test was applied for non-normally distributed continuous variables, and Pearson's Chi-squared test was used for categorical variables. Differences in cortical thickness between groups were evaluated using analysis of covariance (ANCOVA), with age and sex as covariates. To account for multiple comparisons, p-values were corrected using the Benjamini–Hochberg false discovery rate procedure. Pearson's product-moment correlation coefficients were used to assess associations between clinical measures and MRI-derived metrics.

For predictive modeling, the dataset was stratified into a training set (70%) and an independent test set (30%), preserving the distribution of the outcome variable. All preprocessing steps, including Box–Cox transformation, centering, and scaling, were performed using parameters derived exclusively from the training data and applied within each cross-validation fold, ensuring that the test set remained completely unseen during model development.

To accommodate nonlinear relationships, feature interactions, and multicollinearity in a limited-sample setting, multiple machine learning algorithms were evaluated, including Distributed Random Forests (DRF), Extremely Randomized Trees (XRT), Generalized Linear Models with regularization (GLM), Gradient Boosting Machines (GBM), Extreme Gradient Boosting (XGB), and fully connected Deep Learning models. Model development and selection were conducted exclusively on the training set.

Ten-fold cross-validation was employed within the training set to compare predictor configurations, select the optimal algorithm for each configuration, and perform hyperparameter tuning, using the area under the receiver operating characteristic curve (AUC) as the primary performance metric. Following configuration selection, a stacked ensemble Super Learner was constructed to integrate the predictions of the best-performing base learners. Only after finalizing model architecture and hyperparameters was the selected Super Learner retrained on the full training set and evaluated once on the independent test set.

Feature importance and model interpretability were explored using SHapley Additive exPlanations (SHAP) for the leading tree-based model.

All statistical tests were two-tailed, with statistical significance defined as $p < 0.05$. Analyses were performed using RStudio (version 2023.06.1, Build 524).

Results

Table 1 reports population characteristics, clinical assessments, CSF dynamic evaluation, and MRI metrics. There were no significant differences in the distribution of age and gender between the two groups.

The change in UPDRS scores after CSFTT revealed a notable effect size of -0.96, which was statistically significant ($p = 0.016$). Furthermore, the percentage variation post-

procedure in TUG scores had an effect size of 1.1, which was statistically significant ($p = 0.019$).

At baseline, iNPH patients demonstrated significantly higher MMSE scores with an effect size of 0.99 ($p = 0.031$), while FAB scores showed no significant difference between groups ($p = 0.5$). Following CSFTT, the MMSE difference revealed a substantial effect size of 1.26 ($p = 0.005$), and the FAB difference showed an effect size of 0.81 ($p = 0.019$), indicating significant cognitive improvement in the iNPH group. There was no significant difference in the distribution of incontinence between the two groups.

Evaluations of CSF dynamics metrics showed that the P_{start} was not significantly different between the two groups. However, $P_{plateau}$ and R_{out} showed statistically significant differences with effect sizes of 0.89 and 0.56, respectively, with p-values below the 0.001 mark.

EI and CA showed no statistically significant differences between the two groups. However, DESH features showed a statistically significant difference with an effect size of 0.43, with a p-value below the 0.001 mark.

Table 1 – Demographics, clinical assessments, CSF dynamics, and MRI metrics.

Characteristic	Overall, N = 100[1]	Group		Effect Size[2]	p-value[3]
		iNPH, N = 67[1]	Other, N = 33[1]		
Demographics					
Age	75.75 (5.34)	75.40 (4.60)	76.45 (6.61)	-0.20	0.2
Sex				0.11	0.6
F	57 (57%)	37 (55%)	20 (61%)		
M	43 (43%)	30 (45%)	13 (39%)		
Clinical assessments					
UPDRS					
before CSFTT	19.10 (12.42)	18.71 (13.69)	20.18 (16.95)	-0.12	0.7
after CSFTT	17.52 (12.53)	16.58 (10.41)	20.18 (17.54)	-0.29	>0.9
difference	-1.57 (2.39)	-2.13 (2.47)	0.00 (1.18)	-0.96	**0.016**
TUG					
before CSFTT	19.54 (14.82)	18.71 (13.69)	21.86 (18.16)	-0.21	0.8
after CSFTT	18.63 (15.07)	16.98 (13.36)	23.27 (19.05)	-0.42	0.5
% variation	4.95 (14.49)	8.61 (11.83)	-5.36 (16.82)	1.1	**0.019**
CSF dynamic evaluation					
P start	9.88 (2.58)	10.11 (2.80)	9.42 (2.05)	0.27	0.2
P plateau	27.45 (4.71)	28.80 (4.19)	24.91 (4.63)	0.89	**<0.001**
R out	11.43 (3.21)	12.02 (3.22)	10.26 (2.89)	0.56	**<0.001**
MRI metrics					
Evans Index	0.38 (0.05)	0.37 (0.04)	0.40 (0.06)	-0.59	0.2
Callosal Angle	90.63 (22.32)	86.99 (22.52)	99.99 (20.37)	-0.59	0.14

[1] Mean (SD); n (%)

[2] Cohen's D; Standardized Mean Difference

[3] Wilcoxon rank sum test; Pearson's Chi-squared test

Volume analysis

Volume analysis did not show any statistically significant difference in the total brain volume, cortical gray matter, subcortical gray matter, white matter, and ventricular volumes between the two groups.

Cortical thickness analysis

In the frontal lobe, the caudal middle frontal region demonstrated a significant difference in thickness between the two groups (Cohen's D = 0.86, corrected p = 0.002). The rostral middle frontal and superior frontal areas also showed significant differences (corrected p = 0.038 and 0.002, respectively). Additionally, the superior parietal region in the parietal lobe displayed a notable difference with an effect size of 0.80 and a statistically significant corrected p-value of 0.006. The cortical thickness analysis results are reported in Table 2 and shown in Figure 2 and Figure 3.

Table 2 – Cortical thickness analysis.

Thickness	Overall, N = 100[1]	Group		Effect Size[2]	p-value[3]	Corrected p-value[4]
		NPH, N = 67[1]	Other, N = 33[1]			
Caudal anterior cingulate	2.23 (0.23)	2.22 (0.21)	2.23 (0.28)	-0.03	0.8	>0.9
Caudal middle frontal	2.31 (0.20)	2.36 (0.17)	2.20 (0.21)	0.86	<0.001	**0.002**
Cuneus	1.78 (0.14)	1.80 (0.14)	1.75 (0.14)	0.35	0.13	0.3
Entorhinal	2.89 (0.35)	2.94 (0.37)	2.80 (0.30)	0.38	0.080	0.2
Fusiform	2.34 (0.18)	2.34 (0.19)	2.34 (0.16)	0.03	0.9	>0.9
Inferior parietal	2.18 (0.20)	2.21 (0.20)	2.10 (0.19)	0.54	0.020	0.11
Inferior temporal	2.42 (0.20)	2.44 (0.18)	2.38 (0.23)	0.33	0.2	0.3
Isthmus cingulate	1.99 (0.15)	1.98 (0.16)	1.99 (0.14)	-0.04	0.9	>0.9
Lateral occipital	1.99 (0.15)	2.00 (0.16)	1.95 (0.15)	0.31	0.2	0.3
Lateral orbitofrontal	2.25 (0.14)	2.25 (0.14)	2.24 (0.13)	0.07	0.9	>0.9
Lingual	1.78 (0.10)	1.77 (0.09)	1.79 (0.12)	-0.21	0.2	0.4
Medial orbitofrontal	2.09 (0.15)	2.08 (0.15)	2.10 (0.16)	-0.11	0.4	0.5
Middle temporal	2.42 (0.22)	2.45 (0.22)	2.36 (0.22)	0.41	0.084	0.2
Parahippocampal	2.43 (0.26)	2.43 (0.27)	2.42 (0.24)	0.02	>0.9	>0.9
Paracentral	2.25 (0.19)	2.27 (0.20)	2.20 (0.17)	0.38	0.10	0.2
Pars opercularis	2.21 (0.18)	2.23 (0.17)	2.17 (0.20)	0.33	0.2	0.3
Pars orbitalis	2.28 (0.20)	2.28 (0.20)	2.27 (0.22)	0.06	>0.9	>0.9

Thickness	Overall, N = 100[1]	Group		Effect Size[2]	p-value[3]	Corrected p-value[4]
		NPH, N = 67[1]	Other, N = 33[1]			
Pars triangularis	2.06 (0.18)	2.08 (0.17)	2.03 (0.18)	0.30	0.2	0.3
Pericalcarine	1.47 (0.12)	1.49 (0.12)	1.45 (0.11)	0.34	0.14	0.3
Postcentral	1.86 (0.13)	1.88 (0.12)	1.82 (0.13)	0.46	0.043	0.2
Posterior cingulate	2.19 (0.15)	2.20 (0.14)	2.18 (0.17)	0.13	0.4	0.5
Precentral	2.23 (0.21)	2.26 (0.20)	2.16 (0.21)	0.47	0.037	0.2
Precuneus	2.17 (0.19)	2.21 (0.19)	2.10 (0.17)	0.61	0.010	0.061
Rostral anterior cingulate	2.45 (0.22)	2.47 (0.21)	2.43 (0.24)	0.16	0.4	0.6
Rostral middle frontal	2.06 (0.16)	2.09 (0.15)	1.99 (0.18)	0.62	0.005	**0.038**
Superior frontal	2.34 (0.18)	2.38 (0.16)	2.24 (0.19)	0.82	<0.001	**0.002**
Superior parietal	2.09 (0.19)	2.13 (0.18)	1.99 (0.16)	0.80	<0.001	**0.006**
Superior temporal	2.42 (0.23)	2.45 (0.22)	2.36 (0.22)	0.39	0.088	0.2
Supramarginal	2.18 (0.19)	2.21 (0.18)	2.12 (0.20)	0.46	0.046	0.2
Transverse temporal	1.99 (0.29)	2.00 (0.29)	1.97 (0.28)	0.11	0.6	0.7
Insula	2.66 (0.19)	2.65 (0.17)	2.68 (0.23)	-0.15	0.4	0.6

[1] Mean (SD)

[2] Cohen's D

[3] ANCOVA with *Age* and *Sex* as covariates

[4] Benjamini & Hochberg correction for multiple testing

Figure 2 – Boxplots illustrating the difference in cortical thickness between iNPH patients and ‘Other’.

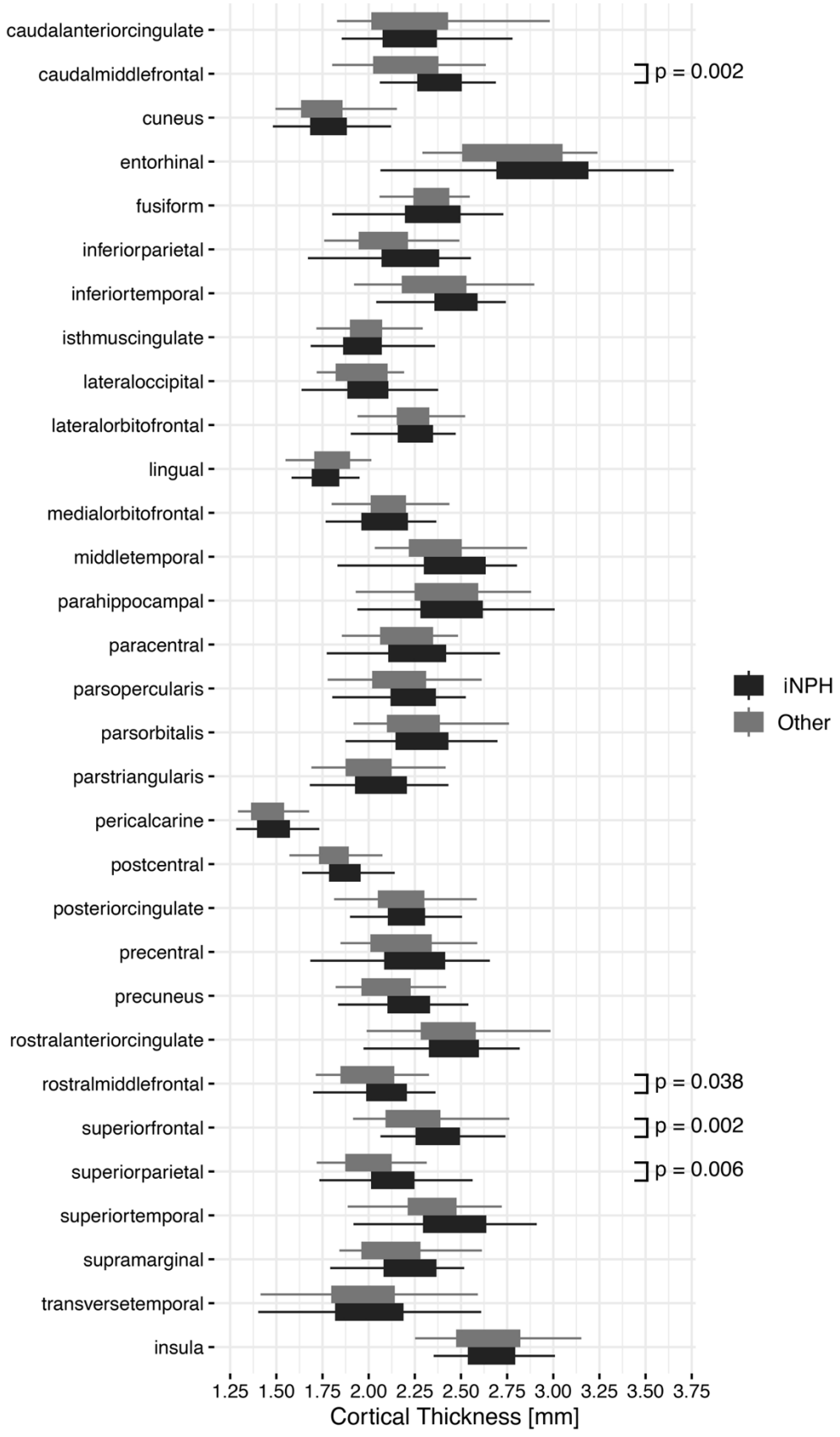
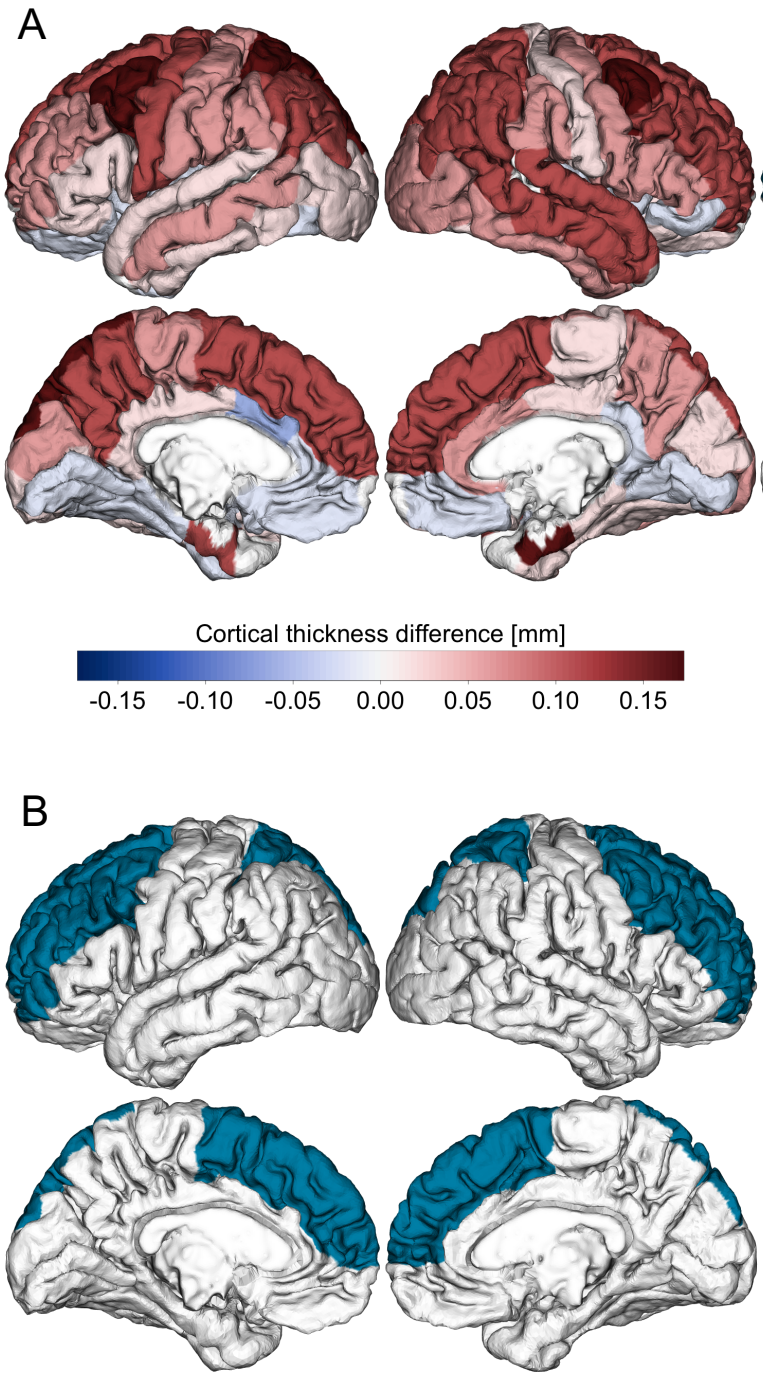


Figure 3 – (A) Difference in cortical thickness between iNPH patients and ‘Other’; (B) Regions of significant difference after correction for multiple comparisons.



Relationship of cortical thickness and clinical assessments

The superior frontal region and the precuneus manifested significant inverse correlations with the difference in UPDRS scores, with coefficients of -0.324 ($p = 0.0366$) and -0.313 ($p = 0.0433$), respectively. This suggests that a reduction in cortical thickness in these regions is correlated with a smaller magnitude of change in UPDRS scores, reflecting a lesser degree of symptom relief post-procedure. Similarly, the superior parietal region revealed a near-significant inverse correlation ($R = -0.302$, $p = 0.0518$), hinting at a possible similar trend or pattern.

No significant correlation was found with TUG scores.

Predictive modeling for normal pressure hydrocephalus

To systematically evaluate the incremental diagnostic value of cortical thickness analysis, we defined 8 distinct predictor configurations:

Model A – Imaging Only: EI, CA, DESH

Model B – Clinical + CSF Dynamics: UPDRS, MMSE, FAB, R_{out}

Model C – Combined conventional (Model A + Model B)

Model D – Cortical Thickness only

Model E – Cortical Thickness + Imaging (Model A + Model D)

Model F – Cortical Thickness + R_{out}

Model G – Cortical Thickness + Clinical

Model H – Full combined

The dataset was initially split into a training set (70%) and an independent test set (30%), stratified by outcome. All model selection procedures were performed exclusively on the training set, while the test set remained untouched until the final validation stage (Figure 4).

For each predictor configuration, multiple machine learning algorithms were trained on the training set, including DRF, XRT, GLM with Regularization, GBM, XGB, and Deep Learning machines. Algorithm comparison and selection were conducted using 10-fold cross-validation within the training set, with the area under the ROC curve (AUC) as the primary performance metric. The cross-validated results for each configuration are reported in Table 3. Based on cross-validated AUC, Model F, combining average cortical thickness measurements (31 ROIs) with the R_{out} value from CSF dynamic evaluation, demonstrated the most favorable performance among the evaluated configurations.

Table 3 – Cross-validated performance of the eight predictor configurations evaluated on the training set only using 10-fold cross-validation. Reported values represent the mean AUC across folds.

Model	Predictor configuration	Algorithm	AUC (10-fold CV, mean ± SD)
A	EI, CA, DESH	GLM	0.679 ± 0.013
B	UPDRS, MMSE, FAB, R_{out}	GBM	0.680 ± 0.027
C	EI, CA, DESH, UPDRS, MMSE, FAB, R_{out}	Deep Learning	0.724 ± 0.104
D	Cortical thickness (31 ROIs)	Deep Learning	0.673 ± 0.116
E	Cortical thickness (31 ROIs), EI, CA, DESH	GLM	0.693 ± 0.072
F	**Cortical thickness (31 ROIs), R_{out}**	**Deep Learning**	**0.759 ± 0.078**
G	Cortical thickness (31 ROIs), UPDRS, MMSE, FAB	Deep Learning	0.723 ± 0.094
H	Cortical thickness (31 ROIs), EI, CA, DESH, UPDRS, MMSE, FAB, R_{out}	Deep Learning	0.690 ± 0.104

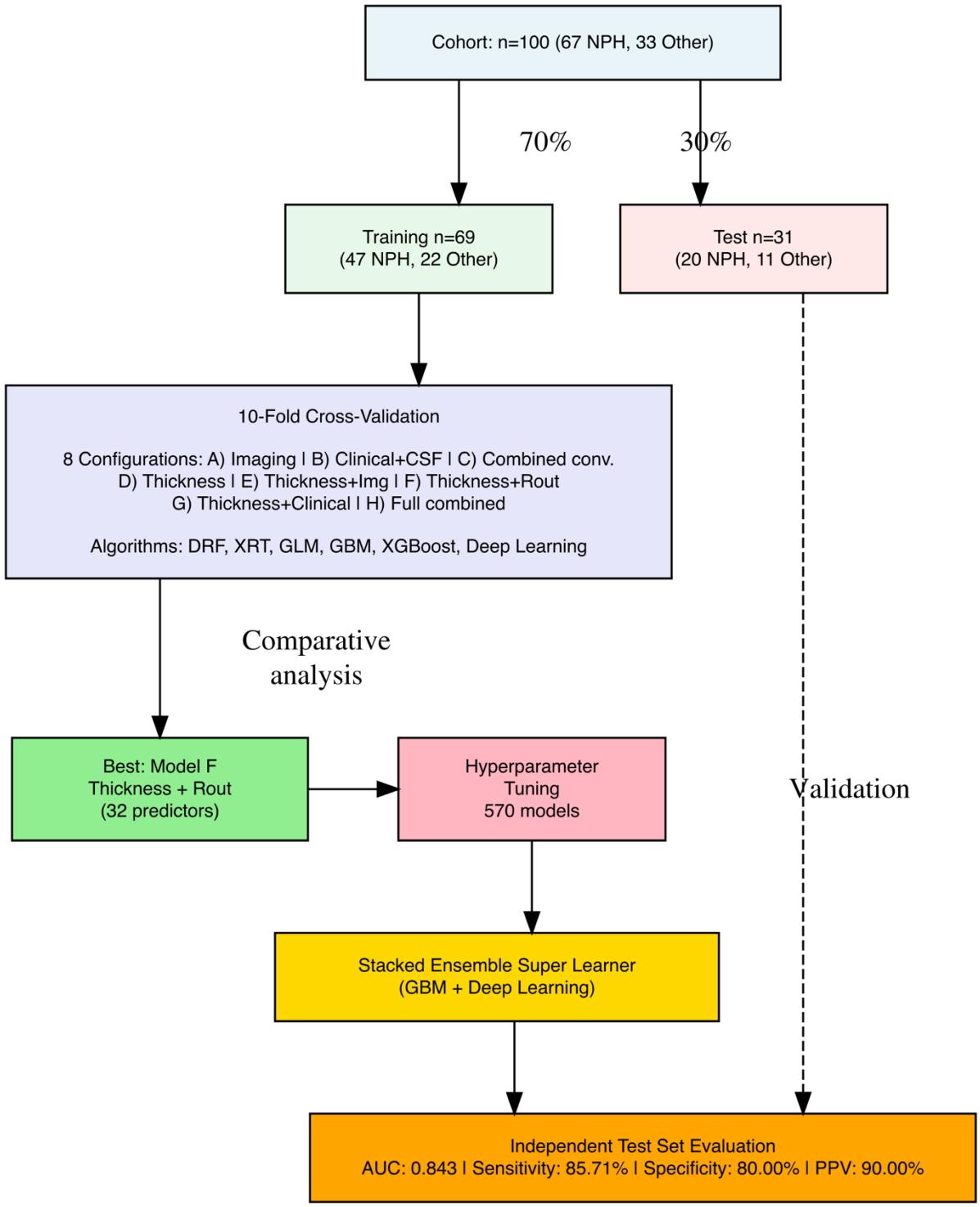


Figure 4: Predictive modeling evaluation.

Following the selection of Model F, we performed extensive hyperparameter tuning and algorithmic exploration exclusively within the training set for this configuration, comprising 32 predictors derived from the DKT atlas cortical thickness measures and Rout. A total of 570 candidate models were trained and evaluated using internal cross-validation. From this process, a stacked ensemble Super Learner emerged as the optimal model, integrating predictions from a GBM and a Deep Learning model.

Only after the full model architecture and hyperparameters were finalized, the Super Learner was retrained on the entire training set and evaluated once on the independent test set. As reported in Table 4, the final model achieved a sensitivity of 85.71%, a specificity of 80.00%, and a positive predictive value of 90.00% on the test set. The corresponding ROC curve is shown in Figure 5, with an AUC of 0.843.

The SHAP Contribution plot, presented in Figure 6, provides a detailed analysis of the predictors' influence on the model's predictions. Notably, the R_{out} variable emerges as a pivotal predictor, affirming its well-established significance in iNPH. Furthermore, regions within the frontal and parietal areas rank prominently, underscoring their determinant roles in the model's decision process.

Table 4 – Confusion matrix for Super Learner on unseen test data.

	Predicted NPH	**Predicted Other**
Actual NPH	18	3
Actual Other	2	8

Sensitivity = $\frac{TP}{TP+FN}$ = 85.71%

Specificity = $\frac{TN}{TP+FN}$ = 80.00%

Positive Predictive Value = $\frac{TP}{TP+FP}$ = 90.00%

Negative Predictive Value = $\frac{TN}{TN+FN}$ = 72.73%

Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$ = 83.87%

Figure 5 – Receiver Operating Characteristic Curve for Super Learner model. AUC: Area Under the Curve

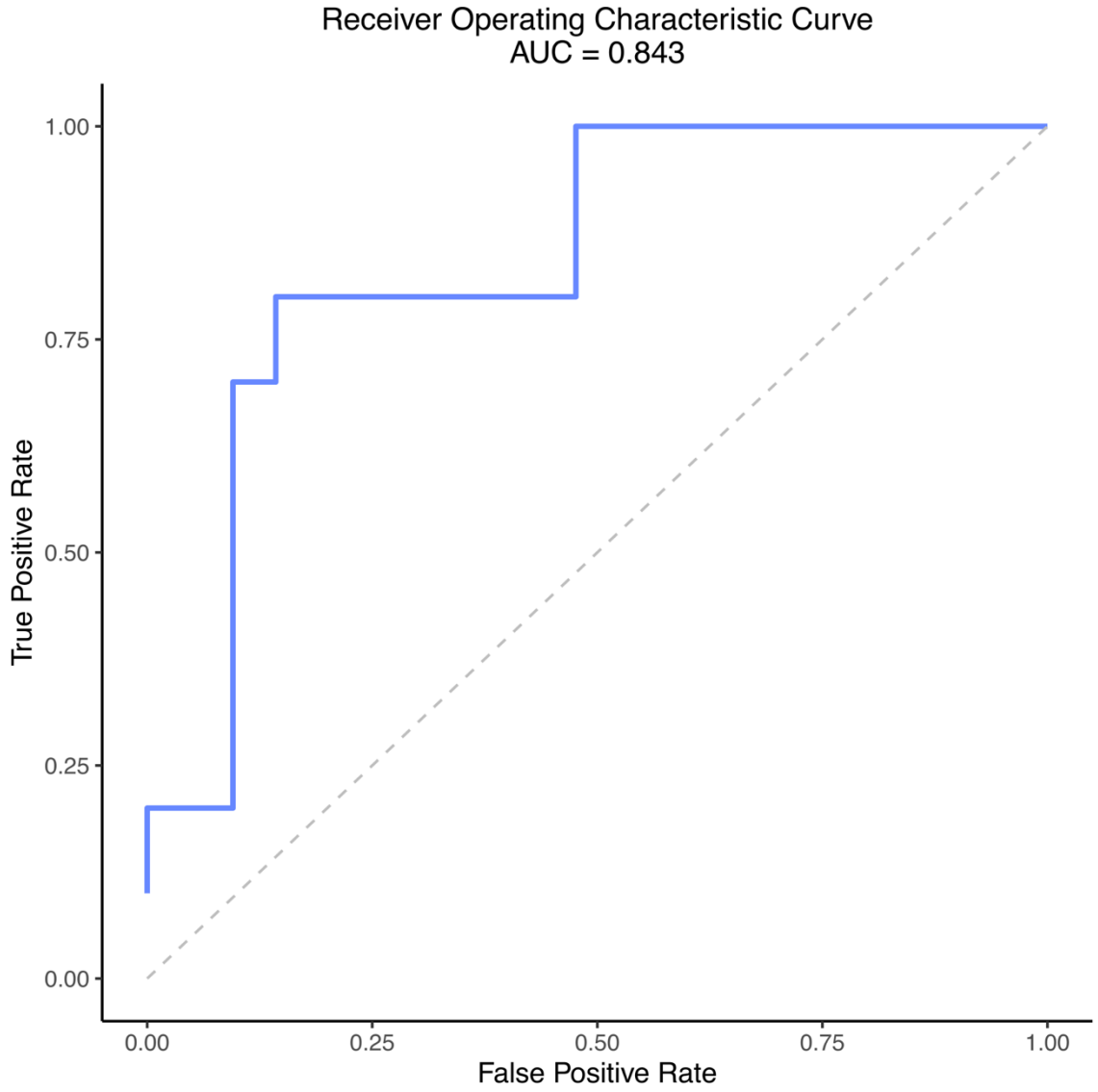
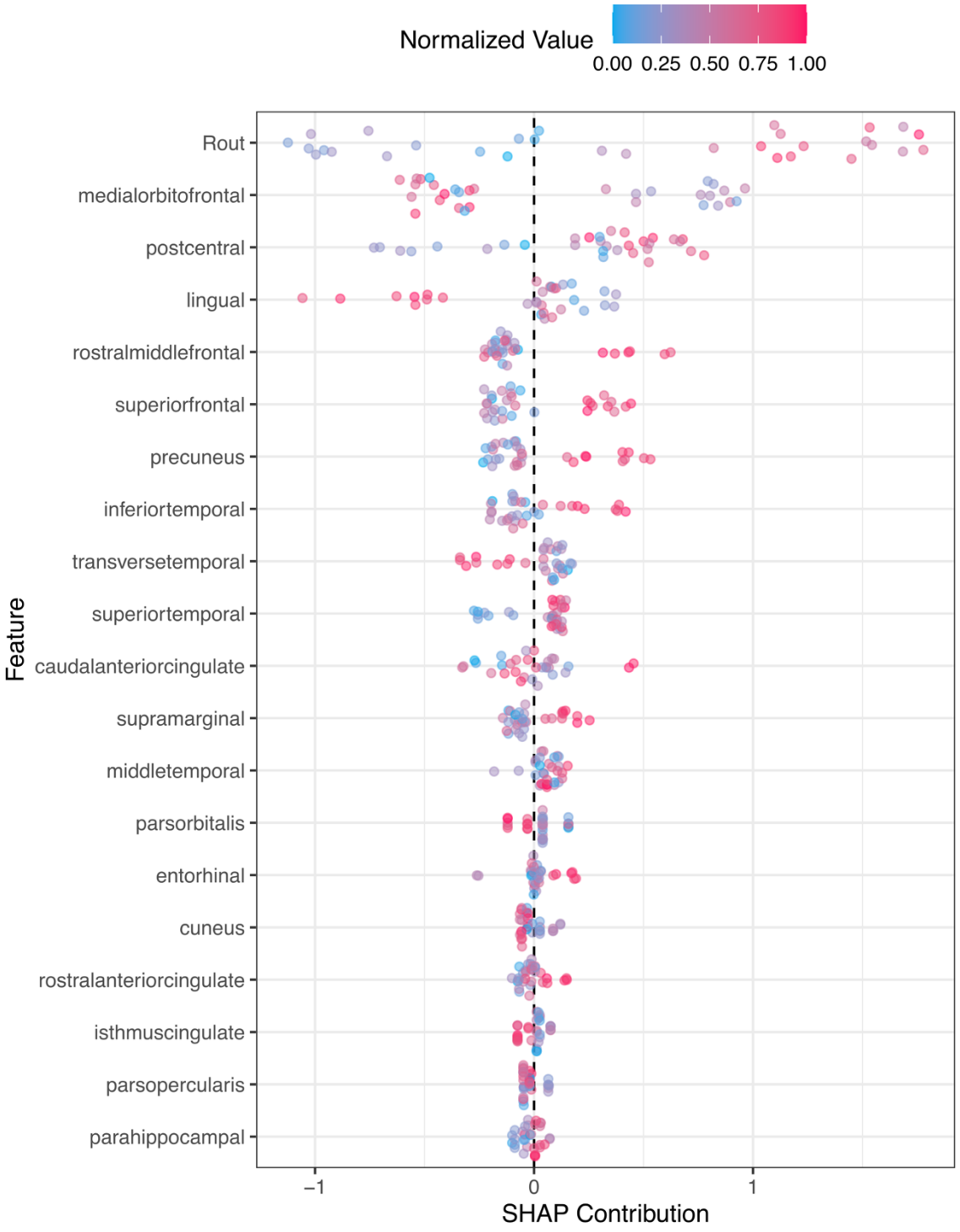


Figure 6 – SHAP contribution plot



Discussion

Idiopathic normal-pressure hydrocephalus stands as a unique and distinct form of progressive dementia that is both primitive and, to an extent, reversible. VPS implantation has been found to alleviate symptoms, with a success rate ranging between 33% and 84%. Such outcome variations are influenced by factors such as the timing of the intervention, methodologies in outcome assessments, and criteria for patient selection[68,105–108]. Diagnosing iNPH relies on a confluence of clinical presentations, neuroradiological findings, and specific invasive diagnostic procedures[68]. Among the clinical manifestations, gait disturbances are almost universally observed, presenting in 94 to 100% of iNPH patients. Concurrently, cognitive impairments and urinary dysfunctions are noted with varying frequencies, affecting between 78-98% and 60-92% of patients, respectively. Interestingly, the simultaneous presence of all these symptoms is observed in approximately 60% of the diagnosed iNPH population[68].

Invasive diagnostic procedures, including CSFTT, help differentiate iNPH from other forms of dementia. The tap test has a sensitivity of 58% and a specificity of 75%, while elevated R_{out} values have shown a PPV of over 80% for VPS success[68].

Cortical thickness measurements, captured across varied brain regions, have recently gained prominence in studying normal development and aging and in the context of neurodegenerative diseases[92,93,95–99]. In the context of iNPH, a study by Kang et al. in 2013 contrasted cortical thickness between CSFTT responders and non-responders, juxtaposing these findings with an AD cohort. The results hinted at potential comorbid AD in TT non-responders, suggesting that coexisting AD might influence this subgroup's observed cortical thinning patterns[92]. A subsequent study by the same group showed significant cortical thickening in frontal, parietal, and occipital regions, as well as cortical thinning in temporal and orbitofrontal areas in iNPH patients compared to healthy controls[93]. Another study by Lang et al. showed decreased GM volume in the right supplementary motor area and in the right posterior parietal cortex for CSFTT non-responders[94]. A study by Belgrado et al. evaluated extrapyramidal signs in the diagnosis of iNPH, showing significant differences in MDS-UPDRS-III scores after CSFTT for definitive and probable iNPH compared to not-iNPH[109].

Our findings align with prior research, drawing attention to notable disparities in cortical thickness, particularly in the middle frontal, superior frontal, and superior parietal regions. Drawing parallels with previously established links between cortical thickness and gait disturbances, our results emphasize the value of assessing regional cortical thickness as a pivotal diagnostic and prognostic indicator for iNPH. Notably, the superior frontal and precuneus regions presented significant inverse correlations with the variations in UPDRS scores post-CSF subtraction. Furthermore, the subtle trend observed in the superior parietal region, while just shy of statistical significance, aligns with these findings. The consistent pattern across these regions suggests that certain areas of the brain, when manifesting reduced cortical thickness, may indicate a less favorable response to interventions like CSF subtraction.

Our study demonstrates the effectiveness of the Super Learner stacked ensemble model in predicting iNPH by leveraging multiple ML techniques. The model's use of predictors from

the frontal and parietal regions aligns with previous research and reinforces their potential relevance in diagnosis. Moreover, the success of our predictive model becomes even more pronounced when considering the PPV of 90% achieved when considering both R_{out} and cortical thickness. This represents a significant enhancement from the approximate 80% PPV when relying on R_{out} alone, underscoring the value of incorporating multiple parameters in iNPH diagnosis. This comprehensive approach, informed by fluid dynamics and structural neuroimaging, may pave the way for more nuanced and tailored therapeutic interventions.

Limitations

Although this study provides novel insights into the diagnostic value of cortical thickness analysis in iNPH, several limitations should be acknowledged. First, the relatively small sample size, while comparable to prior single-center iNPH studies, may limit the generalizability of the findings and increase variability in model performance estimates. To mitigate this, we employed strict training–testing separation and cross-validation–based model selection; however, larger datasets would enable more stable estimation of model parameters and further refinement of complex models, such as ensemble learners.

Second, the study was conducted at a single tertiary referral center, which may introduce center-specific biases in patient selection, imaging protocols, or clinical evaluation. Although standardized diagnostic criteria and imaging procedures were applied, external validation in independent cohorts from other institutions is necessary to confirm the robustness and transportability of the proposed model.

Third, although internal cross-validation was used for model selection and hyperparameter tuning, an independent external validation cohort was unavailable. Although the final model was evaluated on a strictly held-out test set, future studies should aim to validate the proposed approach across multicenter datasets and heterogeneous imaging platforms to further assess real-world performance.

Finally, cortical thickness measurements were derived from automated segmentation pipelines. Although extensively validated, these pipelines may be sensitive to image quality and scanner-specific characteristics. Future work should explore harmonization strategies and assess the impact of acquisition variability on model performance.

Conclusions

Our analysis suggests that patients who exhibit negative CSFTT or remain unresponsive to VPS may have a distinct cortical thinning pattern. Such patterns may indicate comorbidities or alternative neurological pathologies. Consequently, a comprehensive and multi-dimensional diagnostic approach is necessary for iNPH to address overlapping conditions.

6. Radiomic Features of MRI Subcompartments Associate with Angiogenic and Inflammatory Transcriptomic Programs in Glioblastoma: An IvyGAP Exploratory Analysis

Introduction

Glioblastoma (GBM) is the most common and aggressive primary malignant brain tumor in adults, with a median overall survival of approximately 15 months despite maximal safe resection, temozolomide chemotherapy, and radiotherapy[110,111]. A defining feature of GBM is profound intratumoral spatial heterogeneity, with molecularly and histologically distinct regions coexisting within the same tumor[112,113]. This heterogeneity drives therapy resistance by harboring treatment-refractory cellular subpopulations and creating diverse microenvironmental niches[114]. Understanding the molecular programs that govern these distinct tumor regions is therefore critical for developing spatially informed therapeutic strategies.

The Ivy Glioblastoma Atlas Project (IvyGAP) represents a landmark effort to characterize this spatial heterogeneity at the transcriptomic level[115]. In IvyGAP, laser microdissection (LMD) was performed on tumor sections from 41 patients to isolate RNA from five histologically defined anatomic zones: cellular tumor (CT), microvascular proliferation (CTmvp), pseudopalisading cells adjacent to necrosis (CTpan), infiltrating tumor (IT), and leading edge (LE). The resulting 270 RNA-sequencing samples provide a spatially resolved transcriptomic atlas of GBM. Separately, the IVYGAP-RADIOMICS companion dataset[116] provides 3920 International Biomarker Standardization Initiative (IBSI)-compliant radiomic features per MRI-defined subcompartment (enhancing tumor [ET], non-enhancing tumor [NET], and peritumoral edema [ED]) for 31 of these patients with available multiparametric MRI (T1, T1-gadolinium, T2, FLAIR). Despite both datasets being publicly available since 2020, no study has linked the IVYGAP-RADIOMICS feature set to the zone-level IvyGAP RNA-seq data to test whether radiomic features reflect the transcriptomic programs of specific anatomic zones.

Prior studies have explored imaging-transcriptomic associations in GBM[117–121], including radiomic prediction of immune enrichment scores[122] and radiomic-genomic survival models[123,124], but none have connected zone-level transcriptomic data with IBSI radiomic features. Park et al.[117] used ADC/CBV clustering in only five patients ($r < 0.30$); Le et al.[118] and Zhang et al.[121] used whole-tumor approaches; Hu et al.[119] achieved spatially matched biopsies in a different dataset; Beig et al.[120] linked IvyGAP imaging subcompartments to bulk TCGA gene expression and ssGSEA for survival prediction, but did not use zone-level RNA-seq or transcriptomic pathway scores as outcomes. None has linked the IvyGAP atlas with the IVYGAP-RADIOMICS feature set.

In this study, we test whether radiomic features extracted from MRI-defined tumor subcompartments (ET, NET, ED) associate with transcriptomic pathway enrichment scores

derived from the biologically approximate IvyGAP anatomic zones. We employ a biologically motivated but spatially approximate zone-to-subcompartment mapping (e.g., CT and CTmvp to ET; CTpan to NET; IT and LE to ED) and use both linear mixed-effects models (associational analysis) and nested cross-validated Elastic Net regression (predictive analysis) as complementary analytical frameworks. We explicitly frame this as a hypothesis-generating exploratory analysis, acknowledging that the absence of voxel-level spatial co-registration between LMD sites and MRI subcompartments is a fundamental limitation that attenuates all observed associations.

Materials and Methods

Datasets and patient matching

Two publicly available datasets were used. The IvyGAP RNA-seq dataset was obtained from the Allen Institute for Brain Science portal and comprises 270 LMD samples from 41 patients across five anatomic zones (CT, CTmvp, CTpan, IT, LE)[115]. All IvyGAP patients are IDH-wildtype GBM (pre-WHO 2016 cohort). Gene expression values were provided as fragments per kilobase of transcript per million mapped reads (FPKM). The IVYGAP-RADIOMICS dataset was obtained from The Cancer Imaging Archive (TCIA) and provides 3920 IBSI-compliant radiomic features per subcompartment (ET, NET, ED) across four MRI sequences for 31 patients[116]. All MRI data originate from a single institution, eliminating the need for ComBat harmonization. Radiomic features were pre-extracted by Pati et al.[116] from BraTS-style tumor segmentations.

Patient identifiers were matched across the two datasets, yielding 28 patients with both transcriptomic and radiomic data. Zone availability varied across patients (CT available for all 28; other zones for 7–15), resulting in an unbalanced design after subcompartment aggregation (Section 3.1). Under WHO 2021 criteria, all patients classify as glioblastoma, IDH-wildtype. Detailed prior treatment history (chemotherapy, radiation, dexamethasone) was not available from the IvyGAP metadata. Because CTmvp samples were available for only 9 of 28 patients, the ET transcriptomic score represents CT alone for 19 patients and mean(CT, CTmvp) for 9. This compositional inconsistency is particularly relevant for the Angiogenesis pathway, as CTmvp is the angiogenesis-enriched zone; sensitivity analysis S1a (Section 2.9) addresses this by restricting ET to CT only. The overall study design is summarized in Figure 1.

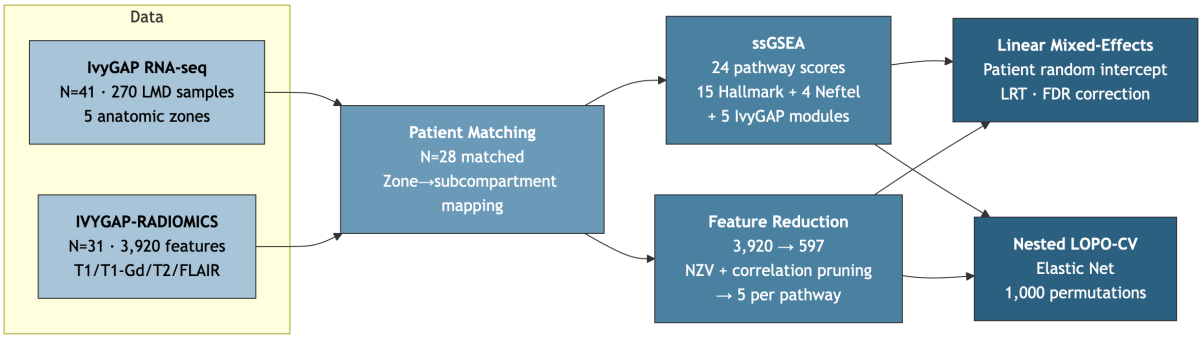


Figure 1 Study design. LMD = laser microdissected; ssGSEA = single-sample gene set enrichment analysis; NZV = near zero variance; LRT = likelihood ratio test; FDR = false discovery rate; LOPO-CV = leave-one-patient-out cross-validation.

Zone-to-Subcompartment Mapping

Because no spatial fiducials or validated registration pipeline links IvyGAP LMD sampling sites (identified post hoc on hematoxylin and eosin histology by neuropathologists[115]) to preoperative MRI voxels, we employed a biologically motivated but spatially approximate mapping between IvyGAP zones and MRI subcompartments (Table 1). Park et al.[117] reported weak but directionally consistent correlations (mean r = 0.242), implying substantial signal attenuation ($r^2 = 0.059$). All reported associations must be interpreted as zone-approximate, not spatially precise, and the degree of attenuation cannot be precisely estimated.

Transcriptomic Target Definition

Twenty-four gene sets were used to compute single-sample Gene Set Enrichment Analysis (ssGSEA) pathway enrichment scores for each LMD sample[125,126]: (1) fifteen GBM-relevant Hallmark gene sets from the Molecular Signatures Database (MSigDB): Hypoxia, Angiogenesis, Epithelial-Mesenchymal Transition (EMT), Inflammatory Response, TNF-alpha/NF-kB Signaling, IL-6/JAK/STAT3 Signaling, Interferon Gamma Response, P53 Pathway, MYC Targets V1, E2F Targets, G2M Checkpoint, mTORC1 Signaling, Glycolysis, Oxidative Phosphorylation, and Complement; (2) four Neftel et al.[114] cellular state signatures: mesenchymal-like (MES; MES1 and MES2 collapsed), astrocyte-like (AC), oligodendrocyte progenitor-like (OPC), and neural progenitor-like (NPC; NPC1 and NPC2 collapsed, as sub-states share core transcription factor programs and the sample size limits power to distinguish sub-state nuances); and (3) five IvyGAP zone-specific gene modules defined as the top 200 differentially expressed genes per zone (Wilcoxon test, FDR < 0.05, $|\log_2$ fold change$| > 1$).

ssGSEA was computed on $\log_2$(FPKM + 1)-transformed expression values using the GSVA package[126] with Gaussian kernel cumulative density function estimation. Zone-level enrichment scores were aggregated to the subcompartment level by computing the mean across zones mapped to each subcompartment (Table 1).

Table 1. Zone-to-subcompartment mapping with biological rationale and empirical support.

IvyGAP Zone	MRI Subcompartment	Biological Rationale	Park et al.[117] Correlation
CT + CTmvp[1]	Enhancing Tumor (ET)	Viable proliferating core and active angiogenesis are the principal sources of gadolinium enhancement	r = 0.238 (CT-ET), r = 0.195 (CTmvp-ET)
CTpan	Non-Enhancing Tumor (NET)	Pseudopalisading necrosis regions are predominantly located within the non-enhancing tumor core	r = 0.241
IT + LE	Peritumoral Edema (ED)	Infiltrating tumor and leading edge extend into the FLAIR-hyperintense peritumoral zone	r = 0.294 (mean IT/LE-ED)

[1] CTmvp (microvascular proliferation) samples are transcriptomically heterogeneous and may share features with CTpan. Sensitivity analysis S1a tests the robustness of all results when CTmvp is excluded from the ET subcompartment.

We note that the five IvyGAP zone modules were derived from the same expression data used to score them, introducing potential circularity. Whether this affects the associational results depends on how many zone module features pass the univariate filter (Section 3.2). Pairwise Jaccard similarity indices were computed for all 24 gene sets to quantify redundancy (Section 3.9).

Feature Reduction Pipeline

The 3920 radiomic features per subcompartment were reduced through a two-stage unsupervised filtering pipeline: (1) near-zero-variance filtering using the nearZeroVar function from the caret package[127], removing features with near-zero variance ratios (3920 to 3860 features); and (2) pairwise Spearman correlation filtering, removing one feature from each pair with $|r| > 0.90$ (3860 to 597 features). These two steps are unsupervised (outcome-agnostic) and do not introduce data leakage.

For the associational analysis (Section 2.5), a third supervised step was applied: per-pathway univariate Spearman correlation with each pathway enrichment score within each subcompartment separately, retaining features with minimum FDR < 0.10 across subcompartments (Benjamini-Hochberg correction[128] applied within each subcompartment) and selecting the top five per pathway. This step was performed on the full dataset, which is standard practice for exploratory associational analyses but introduces potential optimistic bias that must be acknowledged. After this pipeline, 12 of 24 pathways had at least one radiomic feature and were carried forward to mixed-effects modeling. For the predictive analysis (Section 2.6), supervised feature selection was instead performed inside each cross-validation fold to prevent data leakage.

Associational Analysis: Linear Mixed-Effects Models

For each of the 12 pathways with at least one radiomic feature, a linear mixed-effects model (LMM) was fitted using the lme4 package[129] with lmerTest[130] for denominator degrees of freedom (Satterthwaite approximation):

Full model: z_pathway ~ z_rad_1 + … + z_rad_k + subcompartment + (1 | patient_id)

Null model: z_pathway ~ subcompartment + (1 | patient_id)

where z_pathway denotes the globally z-scored pathway enrichment score, z_rad_1 through z_rad_k denote the within-subcompartment z-scored radiomic features (k <= 5), subcompartment is a fixed factor (ET, NET, ED), and patient_id is a random intercept to account for repeated measures within patients.

The omnibus radiomic contribution was assessed via likelihood ratio test (LRT) comparing full and null models. Marginal R^2 and conditional R^2 were computed using the Nakagawa-Schielzeth method[131,132]. FDR correction was applied across all 24 pathways[128], assigning p = 1.0 to 12 pathways with zero features. This conservative approach inflates the FDR denominator, making it harder to achieve significant results, but does not correspond to a formal BH-FDR guarantee, since the assigned values are not true p-values under the null. For FDR-significant pathways, coefficient p-values were Holm-corrected[133]. For the LMM, FDR < 0.05 was the primary threshold; for nested CV results, FDR < 0.10 was used for exploratory interpretation given the conservative 24-pathway denominator, which assigns p = 1.0 to all pathways with no signal. Feature pre-selection may inflate LMM significance; this is independently evaluated through nested CV (Section 2.6).

Primary Predictive Analysis: Nested Cross-Validated Elastic Net

Elastic Net regression[134] was selected because it combines L1 and L2 penalties, promoting grouped selection of correlated features while maintaining sparsity[135], addressing the instability of LASSO[136] in settings with highly correlated radiomic features. The conservative lambda.1se rule was used to guard against overfitting[137].

To provide an unbiased assessment free from data leakage, we implemented nested leave-one-patient-out cross-validation (LOPO-CV) for all 24 pathways. Radiomic features and pathway scores were averaged across subcompartments within each patient, yielding N = 28 independent observations. Patient-level averaging sacrifices subcompartment-level resolution; the LMM (Section 2.5) directly models this variation and is complementary.

For each LOPO fold (28 folds per pathway): (1) one patient was held out for testing; (2) univariate Spearman correlations were computed between each of the 597 candidate features and the pathway score using only the 27 training patients; (3) features with FDR < 0.10 were retained, and the top five by raw p-value were selected; and (4) Elastic Net regression[134] was fitted on the selected features with alpha grid search (0.1 to 1.0, step 0.1) and lambda selected at the conservative one-standard-error rule (lambda.1se), using inner 5-fold CV within the training set.

Performance was evaluated via R^2_{cv} ($1 - SS_{res}/SS_{tot}$), MAE, and Spearman correlation. Uncertainty was quantified via nonparametric bootstrap (B = 1000) for R^2_{cv} confidence

intervals, which condition on fitted predictions without refitting the full pipeline. These CIs capture only metric sampling variability, not the variability from feature selection or hyperparameter tuning across folds. As such, they should not be interpreted as reflecting the full uncertainty of the predictive signal; the nested permutation p-values (Section 2.7), which re-execute the complete pipeline under the null, are the primary inferential tool.

Feature stability was assessed by recording which features were selected in each LOPO fold. Features appearing in more than 50% of folds were designated as "stability-selected"[138]. Inner 5-fold CV on 27 training observations produces noisy hyperparameter selection; alpha and lambda distributions are reported in Supplementary Table S11.

Permutation Testing

Two permutation frameworks were employed. For the LMM associational analysis, patient-level permutation (N = 1000) shuffled the mapping between patients' pathway scores and radiomic features, preserving the within-patient correlation structure. The LRT chi-squared statistic served as the primary test statistic.

For the nested CV predictive analysis, a fully nested permutation test (N = 1000) was performed for pathways with $R^2_{cv} > 0$. In each permutation, pathway scores were shuffled across patients, and the entire nested CV pipeline, including feature selection inside each fold, was re-executed. This provides a truly unbiased permutation p-value that accounts for the adaptive feature selection process. Permutation p-values were computed as $n_{extreme}/N_{perm}$.

Legacy Analysis

A legacy analysis with pre-screened features (selected on the full dataset before cross-validation) is reported in Supplementary Table S5 for methodological comparison with the bias-free nested CV.

Sensitivity Analyses

Ten pre-specified sensitivity analyses tested robustness to alternative zone mappings (S1a, S1b), patient exclusion (S2), aggregation method (S3), FDR subset correction (S4a-c), random effects structure (S6), standardization approach (S7), denominator degrees of freedom (S8), and clinical covariate adjustment for Inflammatory Response (age, MGMT status, Verhaak molecular subtype; S9). Full details are in Supplementary Table S3.

Sample Size Justification

The minimum sample size for P = 5 predictors, $R^2 = 0.20$, and shrinkage S >= 0.90 was $N_{min} = 240$ (pmsampsize[139], Riley et al.[140], Criterion 4). With N = 28, the minimum detectable R^2 at S >= 0.90 was 0.641.

Software and Reproducibility

All analyses were conducted in R version 4.5.0 using the following key packages: lme4 v1.1-37[129], lmerTest v3.1-3[130], performance v0.13.0[132], glmnet v4.1-8[134], pmsampsize v1.1.3[139], GSVA[126], caret v7.0-1[127], and pbkrtest v0.5.5 (for Kenward-Roger correction). A random seed of 42 was used for all stochastic procedures. A CLEAR (CheckList for EvaluAtion of Radiomics research)[141] compliance table is provided as Supplementary Table S8.

Results

Throughout, R^2_{cv} denotes cross-validated R^2, R^2_m marginal R^2 (fixed effects), R^2_c conditional R^2 (fixed + random), and ΔR^2_m the radiomic increment beyond subcompartment effects.

Data Availability and Feature Reduction

Of 41 IvyGAP patients and 31 IVYGAP-RADIOMICS patients, 28 were present in both datasets and constituted the analysis cohort. Of these, 27 (96%) underwent primary surgery and 1 (4%) had recurrent tumor (mean age 58.5 years, SD 7.8; median KPS 90; MGMT methylated 13/28). These 28 patients contributed a total of 50 observations across three MRI subcompartments (ET: n = 28; NET: n = 15; ED: n = 7), with only 6 of 28 patients having data for all three subcompartments. The imbalance reflects differential availability of zone-level RNA-seq data across patients (Figure S8).

The unsupervised radiomic feature reduction pipeline progressively reduced dimensionality: 3920 initial features were filtered to 3860 after near-zero-variance removal, and to 597 after Spearman correlation pruning ($|r| > 0.90$) (Figure S9).

Nested Cross-Validation: Predictive Performance

Of the 24 pathways evaluated, three showed positive predictive signal ($R^2_{cv} > 0$) in the nested LOPO-CV analysis with internal feature selection (Table 2, Figure 2).

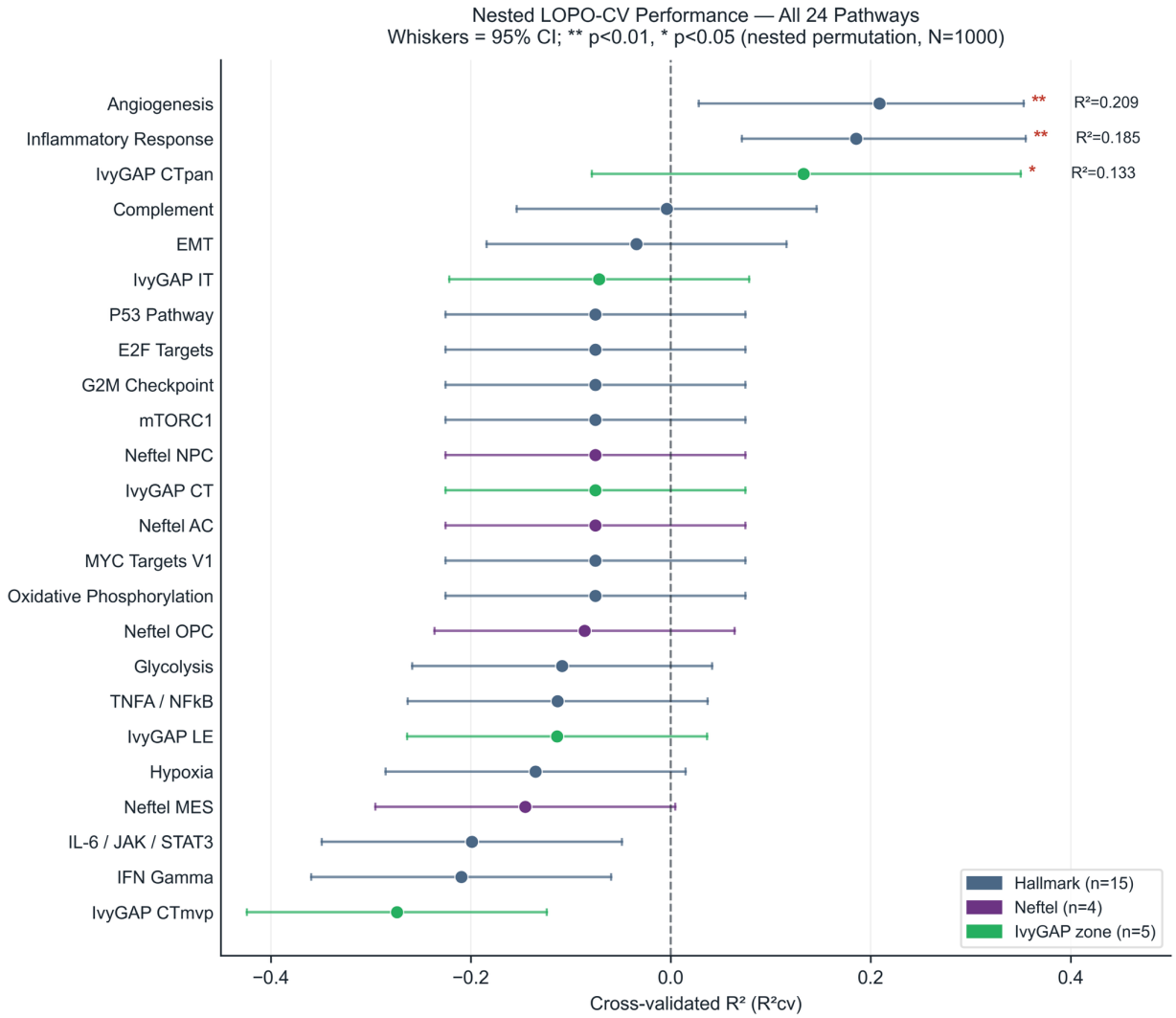


Figure 2 Nested cross-validated predictive performance of radiomic features for 24 transcriptomic pathway enrichment scores. Each point shows R^2_{cv} from leave-one-patient-out cross-validation (LOPO-CV; N = 28 patients) with Elastic Net regression and feature selection performed independently inside each fold. Whiskers indicate bootstrap 95% confidence intervals. Points are colored by gene set category: dark blue = Hallmark pathways (n = 15), purple = Neftel cellular states (n = 4), green = IvyGAP zone modules (n = 5). Significance annotations denote nested permutation p-values (** $p < 0.01$, * $p < 0.05$; N = 1000). Pathways are ordered by descending R^2_{cv}. The dashed vertical line marks $R^2_{cv} = 0$. R^2_{cv} = cross-validated coefficient of determination.

Table 2. Nested cross-validation results for pathways with $R^2_{cv} > 0$.

Pathway	R^2_{cv}	95% CI[e]	MAE	Spearman rho	Stable features (>50% folds)	Nested perm p	FDR (24)[1]
Angiogenesis	0.209	[0.028, 0.353]	0.702	0.581	5	0.006	0.096
Inflammatory Response[c]	0.185	[0.071, 0.355]	0.674	0.524	5	0.008	0.096
IvyGAP CTpan module[d]	0.133	[-0.079, 0.350]	0.740	0.348	4	0.013	0.104

[1] BH-FDR correction applied across all 24 pathways, assigning p = 1.0 to 21 pathways with $R^2_{cv} <= 0$ (Table S10). This is conservative; FDR across the 3 tested pathways = 0.012. [c] Designated as primary based on convergent LMM evidence (Section 3.4). Both analyses use the same outcome data and are not independent; this represents consistency across different statistical models on the same dataset, not independent replication. [d] CI crosses zero; gene module scored on the same expression data from which it was derived (see Limitation 7); composition baseline test significant (p = 0.001). Interpret with caution. [e] Bootstrap CIs (B = 1000) condition on fitted predictions and capture metric sampling variability only; see Section 2.6 for interpretation.

After BH-FDR correction across all 24 pathways, Angiogenesis and Inflammatory Response both reached FDR < 0.10 (FDR = 0.096), while the CTpan module did not (FDR = 0.104). Based on convergent evidence from the complementary LMM analysis (Section 3.4), which identified Inflammatory Response as the sole FDR-significant pathway, Inflammatory Response was designated as the primary pathway for nested CV evaluation.

The remaining 21 pathways all had $R^2_{cv} <= 0$ (Table S1), indicating no patient-level predictive signal. Twelve pathways had zero features passing the univariate filter in any fold (median features per fold = 0), indicating complete absence of radiomic-transcriptomic associations at the patient level.

The CTpan module had a confidence interval spanning zero, is circular by construction (Section 2.3), and showed a significant composition effect (Spearman rho = 0.579, p = 0.001); it should not be considered robust. Composition baseline tests for Angiogenesis (rho = -0.105, p = 0.594) and Inflammatory Response (rho = -0.003, p = 0.988) were non-significant.

Feature Stability and Identification

All features contributing to the Inflammatory Response prediction were T2-derived (Table S12, Figure 3), consistent with the known sensitivity of T2-weighted imaging to inflammatory edema and tissue water content.

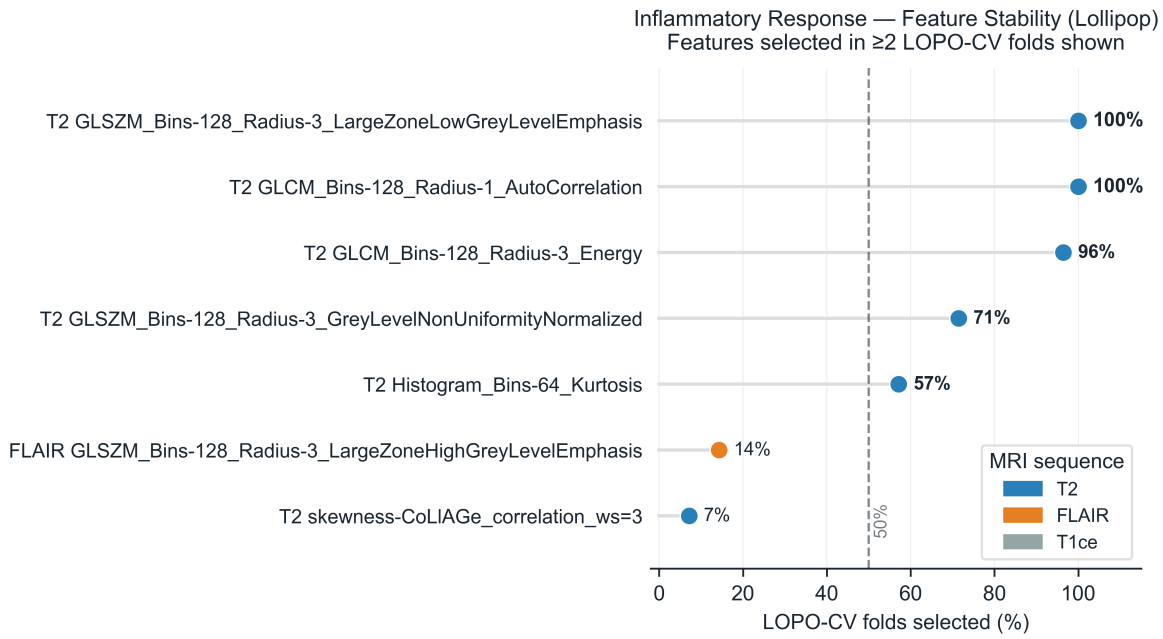


Figure 3 Feature stability for the Inflammatory Response pathway across 28 LOPO-CV folds. Each row represents a radiomic feature selected in at least two folds; the x-axis shows selection frequency (percentage of folds in which the feature was selected by univariate Spearman screening, FDR < 0.10, top 5). Points are colored by MRI sequence: blue = T2, orange = FLAIR, gray = T1ce. The dashed vertical line marks the 50% stability threshold. LOPO-CV = leave-one-patient-out cross-validation; FDR = false discovery rate; GLCM = Gray-Level Co-occurrence Matrix; GLSZM = Gray-Level Size Zone Matrix.

Two features were selected in all 28 folds (100% stability): T2 GLCM AutoCorrelation (rad_1707), reflecting tissue homogeneity patterns, and T2 GLSZM Large Zone Low Grey Level Emphasis (Bins-128, Radius-3; rad_1950), quantifying large contiguous low-signal regions. A third GLCM feature (Energy, rad_1930) was selected in 96% of folds, suggesting that T2 texture features consistently capture inflammatory microenvironment signatures. Note that rad_1950 (Bins-128, Radius-3) and the LMM feature rad_1732 (Bins-128, Radius-1) are distinct radius configurations of the same GLSZM texture class, explaining their independent selection across the two analytical frameworks.

For the Angiogenesis pathway, stability selection identified five features in more than 50% of LOPO folds (Table S9), all T2-derived: three GLSZM texture features (100% stability) and two first-order/GLSZM features (96% stability).

A complete feature lookup table mapping all radiomic feature indices to their full IBSI names is provided in Supplementary Table S6.

Associational Analysis: Mixed-Effects Models

Of the 12 pathways tested in the LMM analysis (after supervised univariate pre-screening), only Inflammatory Response reached significance after FDR correction across all 24 pathways: FDR = 0.024, marginal R^2 = 0.384, of which ΔR^2 = 0.214 is attributable to radiomic features beyond the subcompartment effect (null model R^2_m = 0.170), conditional R^2 = 0.687, LRT χ^2 = 20.53 (df = 5, p = 0.001) (Table 3, Figure S1). Note that features were

pre-selected on the full dataset; marginal R^2 values are likely optimistically biased (see Section 2.4). The intraclass correlation for Inflammatory Response was ICC = 0.492, indicating moderate within-patient correlation.

Table 3. Linear mixed-effects model results for all 12 tested pathways, ordered by FDR. R^2_m (null) = subcompartment + random intercept only; ΔR^2_m = radiomic increment.

Pathway	Category	k	R^2_m (null)	R^2_m (full)	ΔR^2_m	R^2_c	LRT p	FDR (24)
Inflammatory Response	Hallmark	5	0.170	0.384	0.214	0.687	0.001	0.024 *
Angiogenesis	Hallmark	1	0.425	0.459	0.034	0.582	0.053	0.445
Hypoxia	Hallmark	5	0.837	0.847	0.010	0.918	0.085	0.445
P53 Pathway	Hallmark	5	0.470	0.533	0.063	0.716	0.101	0.445
Glycolysis	Hallmark	5	0.789	0.804	0.015	0.849	0.112	0.445
mTORC1 Signaling	Hallmark	3	0.671	0.694	0.024	0.709	0.133	0.445
Neftel MES	Neftel	2	0.546	0.562	0.016	0.699	0.133	0.445
Complement	Hallmark	5	0.059	0.208	0.149	0.462	0.148	0.445
EMT	Hallmark	5	0.339	0.409	0.070	0.589	0.176	0.469
TNFA/NF-kB	Hallmark	5	0.580	0.610	0.030	0.703	0.352	0.845
IvyGAP CTpan Module	IvyGAP	5	0.906	0.902	-0.004	0.922	0.773	1.000
Oxidative Phosphorylation	Hallmark	3	0.352	0.344	-0.008	0.557	0.800	1.000

* FDR < 0.05 across all 24 pathways (BH correction). Because the Nakagawa-Schielzeth R^2 is not strictly additive in mixed models, ΔR^2_m values are approximate; negative values (CTpan, OxPhos) reflect variance repartitioning.

Pathways with high R^2_m (full) but minimal ΔR^2_m, such as Hypoxia (ΔR^2_m = 0.010) and Glycolysis (ΔR^2_m = 0.015), showed negligible radiomic increments. Inflammatory Response had the largest ΔR^2_m (0.214).

Residual diagnostics (Figure S3) showed approximate normality (Shapiro-Wilk p = 0.384). One observation exceeded Cook's D = 1.0 (D = 1.097); sensitivity analyses (S1a, S2) confirm robust results (Table S3).

Angiogenesis had only one feature surviving the univariate pre-screen and did not reach LMM FDR significance (FDR = 0.445), but the nested CV analysis identified a more informative feature set, underscoring the importance of the nested CV approach.

The Inflammatory Response model included five radiomic features (Table S13, Figure S2). The strongest individual contributor was T2 GLSZM Large Zone Low Grey Level Emphasis (Bins-128, Radius-1; rad_1732; beta = -0.471, 95% CI [-0.758, -0.184], Holm-adjusted p = 0.010), indicating a negative association between this texture feature and inflammatory pathway enrichment. One additional feature, T2 CoLIAGe skewness of difference variance (ws=5; rad_1080), showed nominal significance (beta = 0.308, 95% CI [0.007, 0.610], uncorrected p = 0.045) but did not survive Holm correction (adjusted p = 0.181).

The subcompartment fixed effect was also significant ($F(2, 17.5) = 9.48$, $p = 0.002$), indicating that inflammatory pathway enrichment scores differed across ET, NET, and ED subcompartments independently of radiomic features.

Permutation Testing

Patient-level permutation testing for the LMM (N = 1000 permutations) assessed whether the Inflammatory Response LRT χ^2 statistic exceeded chance expectation. The observed χ^2 of 20.53 yielded a permutation p-value of 0.055 for the χ^2 statistic and p = 0.050 for marginal R^2 (Figure S4). These values indicate that the Inflammatory Response association exceeds approximately 95% of the null distribution but does not reach conventional significance at $\alpha = 0.05$. This permutation test used pre-selected features (i.e., the same features in every permutation), which does not fully account for the adaptive feature selection process.

The nested permutation test (Section 2.7), which re-selects features inside each fold of each permutation, provides a more rigorous assessment. All three pathways with $R^2_{cv} > 0$ reached significance at the uncorrected level: Angiogenesis (p = 0.006), Inflammatory Response (p = 0.008), and IvyGAP CTpan module (p = 0.013) (Table 2). After BH-FDR correction across all 24 pathways (Table S10), Angiogenesis and Inflammatory Response reached FDR < 0.10 (FDR = 0.096 each), while CTpan did not (FDR = 0.104). The decision to test only pathways with $R^2_{cv} > 0$ in the nested permutation is data-dependent; this is accounted for by reporting FDR-corrected p-values across all 24 pathways.

Clinical Covariate Adjustment

To assess whether the radiomic-transcriptomic association for Inflammatory Response was confounded by clinical variables, we progressively added covariates to the LMM (Table S7).

The radiomic-transcriptomic association for Inflammatory Response remained significant across all covariate models (LRT $p < 0.005$ in all cases). Adding age and MGMT methylation had minimal impact on the marginal R^2 (Models B-C), while adding Verhaak molecular subtype (Model D) increased the marginal R^2 to 0.479, suggesting that molecular subtype explains additional variance in inflammatory pathway activity beyond radiomic features alone. Critically, the radiomic contribution remained significant even in the fully adjusted model (LRT p = 0.004), indicating that the association is not driven by confounding from age, MGMT status, or molecular subtype. Model D results should be interpreted with caution, given the high parameter-to-observation ratio (12 parameters from 50 observations across 28 clusters).

Sensitivity Analyses

The Inflammatory Response association was robust across the majority of sensitivity analyses (Table S3, Figure S5). The LRT p-value remained below 0.01 under alternative zone mappings (S1a: p = 0.003; S1b: p = 0.003), exclusion of patients with single subcompartments (S2: p = 0.007), median aggregation (S3: p = 0.001), Hallmark-only FDR re-correction (S4a: FDR = 0.010), and Kenward-Roger denominator degrees of freedom (S8: p = 0.006).

Robustness under S1a (ET = CT only, excluding CTmvp) addresses the concern that CTmvp samples could confound the ET signal (S1a: p = 0.003, R^2_m = 0.365).

Two analyses qualified the primary result. Global standardization (S7) weakened the association to p = 0.076, indicating that the radiomic signal captures within-subcompartment variation in tissue properties rather than between-subcompartment differences, which are already captured by the subcompartment fixed effect. The random slopes model (S6) could not be fitted due to insufficient sample size.

Legacy Pre-Screened Analysis

The legacy pre-screened Elastic Net yielded R^2_{cv} = -0.104 for Inflammatory Response (Table S5), compared with R^2_{cv} = 0.185 from nested CV, confirming that pre-screening introduced overfitting.

Gene Set Overlap

Jaccard similarity analysis confirmed that Angiogenesis and Inflammatory Response represent independent signals (maximum J < 0.10). Only the IvyGAP IT and LE modules showed high overlap (J = 0.653).This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation, as well as the experimental conclusions that can be drawn.

Discussion

This exploratory analysis tested whether MRI-derived radiomic features from tumor subcompartments associate with regional transcriptomic programs in GBM, leveraging the IvyGAP and IVYGAP-RADIOMICS public datasets. Using a nested cross-validation framework, not employed in prior IvyGAP radiomic studies[117–121], we identified two transcriptomic programs with positive predictive signal at the patient level: Angiogenesis (R^2_{cv} = 0.209) and Inflammatory Response (R^2_{cv} = 0.185), while 21 of 24 pathways showed no signal. In the complementary associational analysis, Inflammatory Response was the sole pathway reaching FDR significance (FDR = 0.024 across all 24 pathways, ΔR^2 = 0.214 beyond subcompartment effects), driven primarily by T2-derived texture features.

Biological Interpretation

The identification of Angiogenesis and Inflammatory Response as the two pathways with predictive radiomic signal is biologically plausible. Both processes directly modulate MRI signal through mechanisms that alter tissue contrast.

Angiogenesis drives gadolinium enhancement through leaky neovasculature. However, the dominant features were T2-derived texture features. This likely reflects a methodological

constraint: within the ET subcompartment (defined by enhancement), T1-gadolinium features have a restricted dynamic range because the region is already selected for high enhancement, whereas T2 features retain full dynamic range and capture internal heterogeneity of the enhancing core. Biologically, T2 texture within the enhancing tumor reflects downstream consequences of angiogenesis (irregular vascular architecture, patchy edema, and microhemorrhage) rather than the vascular process directly. This finding aligns with Dextraze et al.[123], who identified associations between angiogenesis-related pathways and MRI-defined imaging habitats in 85 GBM patients. Angiogenesis had the highest R^2_{cv} (0.209) despite not reaching LMM FDR significance, underscoring the advantage of nested CV feature selection.

Inflammatory Response involves tumor-associated macrophages, which constitute 30-50% of the GBM mass[142,143], modulating vascular permeability and blood-brain barrier integrity, processes that directly affect T2/FLAIR signal. The dominant feature (T2 GLSZM Large Zone Low Grey Level Emphasis, beta = -0.471) showed a negative association, suggesting that inflammatory activity corresponds to more heterogeneous tissue architecture.

Why These Pathways Survive Spatial Mismatch

The selective survival of Inflammatory Response is consistent with the spatial scale of this process. Tumor-associated macrophages constitute 30-50% of GBM mass[142,143], permeating the entire subcompartment and creating a spatially homogeneous alteration of tissue properties. Because inflammation modulates the full subcompartment volume, any LMD sample captures the inflammatory signal, and radiomic features integrate the same signal across the volume. The effective signal attenuation from spatial mismatch may therefore be lower than the 94% predicted by Park et al.[117] correlations for spatially localized processes.

Angiogenesis operates differently: neovascularization creates focal structures (glomeruloid microvascular proliferation bodies), but its downstream tissue-level consequences, such as edema, vascular permeability, and microhemorrhage, are diffuse and captured by T2 texture across the enhancing core. The mechanism is structural-consequential rather than diffuse per se.

Scope and Negative Results

The dominant finding is negative: 21 of 24 pathways showed no predictive signal ($R^2_{cv} <= 0$). Several patterns are informative.

All five IvyGAP zone-specific gene modules showed no radiomic signal (zero features passed the univariate filter in four of five). This zone module paradox is striking: genes that define the zones themselves fail to map to MRI subcompartments, suggesting either a prohibitive spatial scale mismatch or that the zone-to-subcompartment mapping is too imprecise for these signatures.

All four Neftel cellular state signatures (MES, AC, OPC, NPC) showed no signal. These signatures represent cell-intrinsic transcription factor programs (e.g., CEBP/D for MES, OLIG1/2 for OPC) that do not directly alter tissue-level MRI contrast, unlike angiogenesis or inflammation, which physically modulate vascular permeability and tissue water content. Furthermore, single-cell data show that all four states coexist within the same histological

zone, and bulk RNA-seq from that zone averages across these states, washing out spatial specificity. The spatial scale mismatch is also maximal: Neftel states vary at the 10-50 micron single-cell level, two to three orders of magnitude below the millimeter-scale radiomic features. The selective failure of cell-intrinsic signatures, combined with the selective success of microenvironmental pathways, is consistent with the interpretation that radiomics captures tissue-level rather than cell-level biology, a selectivity that itself argues against the positive results being artifacts of an overfitting pipeline.

Several Hallmark pathways with a biological rationale for MRI detectability, including Hypoxia, EMT, and Glycolysis, also showed no radiomic association. For Hypoxia, the high null-model R^2_m (0.837) indicates that subcompartment membership alone captures the hypoxia gradient, leaving only $\Delta R^2_m = 0.010$ for radiomic features. These negative results constrain the assumption that radiomic features can serve as proxies for arbitrary molecular programs.

Methodological Considerations

Data leakage through feature pre-selection is a critical concern in radiomic studies[144]. Our legacy Elastic Net with pre-screened features produced $R^2_{cv} = -0.104$ for Inflammatory Response, while the bias-free nested CV yielded $R^2_{cv} = 0.185$, demonstrating that pre-selection introduced optimistic bias that paradoxically worsened predictions. We addressed this by implementing nested CV as the primary analysis, with feature selection inside each LOPO fold. For the LMM, features were pre-selected on the full dataset; the independent confirmation in nested CV provides convergent evidence that the Inflammatory Response association is not an artifact of data leakage. The discrepancy between LMM permutation $p = 0.055$ and nested CV permutation $p = 0.008$ for Inflammatory Response reflects different null hypotheses: the LMM permutation uses pre-selected features in every permutation, while the nested CV permutation re-executes the full pipeline, including feature selection under the null, generating a tighter null distribution.

The designation of Inflammatory Response as the primary pathway was based on consistent results across the nested CV (FDR = 0.096) and LMM (FDR = 0.024) analyses. Because both analyses use the same outcome data (ssGSEA scores from the same 28 patients), this consistency should not be interpreted as independent replication. Rather, the two approaches, one predictive with internal feature selection and one associational with pre-screened features, converge on the same pathway despite different statistical frameworks, model structures, and units of analysis (patient-level vs observation-level). The dependence of results on the standardization approach (S7: $p = 0.076$ under global standardization) suggests that the association is partly driven by within-subcompartment relative feature values rather than absolute magnitudes.

Limitations

This study has several important limitations. Firstly, the zone-to-subcompartment mapping is biologically approximate rather than spatially precise, with weak correlations (mean $r = 0.242$), as reported by Park et al.[117]. All results should be interpreted within this context. Although a classical measurement error attenuation model suggests true effect sizes could be larger, its applicability to categorical zone-to-subcompartment mapping is uncertain and cannot be verified with the current data. The sample size of 28 patients limits statistical

power; the Riley criterion requires 240 subjects for P = 5, R2 = 0.20, S >= 0.90. Consequently, this study is hypothesis-generating rather than definitive, and the possibility of false-positive associations cannot be excluded. Additionally, there is no external validation cohort because IvyGAP is the only dataset that combines zone-level RNA-seq with matched MRI, preventing independent replication. The MRI data were obtained from a single institution, which may limit generalizability across scanners, though this also reduces batch effects. BraTS-style segmentations are radiological rather than biological boundaries. In the associational analysis, feature pre-selection on the full dataset may inflate significance, though nested cross-validation results help mitigate this concern. The R2 decomposition now separates radiomic increments from subcompartment effects, offering a clearer interpretation. Since IvyGAP zone modules are scored on the same expression data used for their derivation, circularity affects certain modules like CTpan, which showed positive R2cv but should be interpreted cautiously. The LMD samples represent microscopic tissue volumes—roughly 500-2000 cells—covering less than 0.01% of the subcompartment, across centimeters, which introduces a cross-scale gap between the transcriptomic measurements and radiomic features that integrate signals over full subcompartment volumes. FPKM normalization was used instead of TPM or raw counts, with rank-based ssGSEA partially mitigating this. Permutation p-values for the LMM were around 0.05, warranting cautious interpretation, while the nested CV permutation p-value of 0.008 offers a more rigorous bound; the true significance likely lies between these values (see Section 4). Treatment confounding could influence results, especially given the effects of prior therapies such as bevacizumab and dexamethasone, which can alter enhancement patterns and inflammation. Although most patients had primary tumors, dexamethasone use at the time of MRI is unknown, and its suppression of inflammatory gene expression and edema may attenuate observed associations. Finally, WHO 2021 molecular markers and the top feature caps were not available or were sensitivity-tested.

Future Directions

If confirmed with spatially precise co-registration, these findings could motivate non-invasive molecular profiling of tumor subregions. Future studies should prioritize spatially co-registered datasets extending the Hu et al.[119] approach, multi-institutional harmonized radiomic extraction, and deep learning-derived features that may capture relationships beyond hand-crafted IBSI features. Specifically, future studies should test whether radiomic identification of inflammatory subregions could predict regional response to immunotherapy or anti-angiogenic agents.

Conclusions

This exploratory analysis of 28 patients with matched MRI radiomic features and zone-level RNA-seq data from the IvyGAP atlas identified two transcriptomic programs with positive radiomic associations in bias-free nested cross-validation: Angiogenesis (R^2_{cv} = 0.209, 95% CI [0.028, 0.353]) and Inflammatory Response (R^2_{cv} = 0.185, 95% CI [0.071, 0.355]), while 21 of 24 pathways showed no signal. In the associational analysis, Inflammatory Response was the sole FDR-significant pathway (FDR = 0.024 across all 24 pathways, ΔR^2 = 0.214 beyond subcompartment effects), driven by T2-derived GLSZM and CoLIAGe texture features. These findings provide hypothesis-generating evidence that angiogenic and inflammatory microenvironment programs are selectively reflected in MRI-derived radiomic

features, while constraining claims about the broader utility of radiomics as a proxy for arbitrary molecular programs. Critically, the absence of radiomic signal for 21 of 24 pathways, including all cell-intrinsic signatures, constrains expectations for MRI-based molecular profiling in glioblastoma. Validation in cohorts with spatially precise co-registration is needed.

Supplementary materials

Figure S1. Exploratory linear mixed-effects model (LMM) results for all 12 pathways that retained radiomic features after univariate screening. Each point represents the marginal R^2 (variance explained by fixed effects only) for a given pathway. Horizontal lines extend from zero to the point estimate. Red points indicate pathways reaching FDR < 0.05 after Benjamini-Hochberg correction across 12 tests; gray points indicate non-significant pathways (NS). The dashed vertical line marks $R^2_{marginal} = 0$. Only the Inflammatory Response pathway reached significance (FDR = 0.012). Features were pre-selected on the full dataset; these results carry optimistic bias and should be interpreted alongside the nested cross-validation analysis (Figure 2). LMM = linear mixed-effects model; FDR = false discovery rate; $R^2_{marginal}$ = Nakagawa marginal R^2.

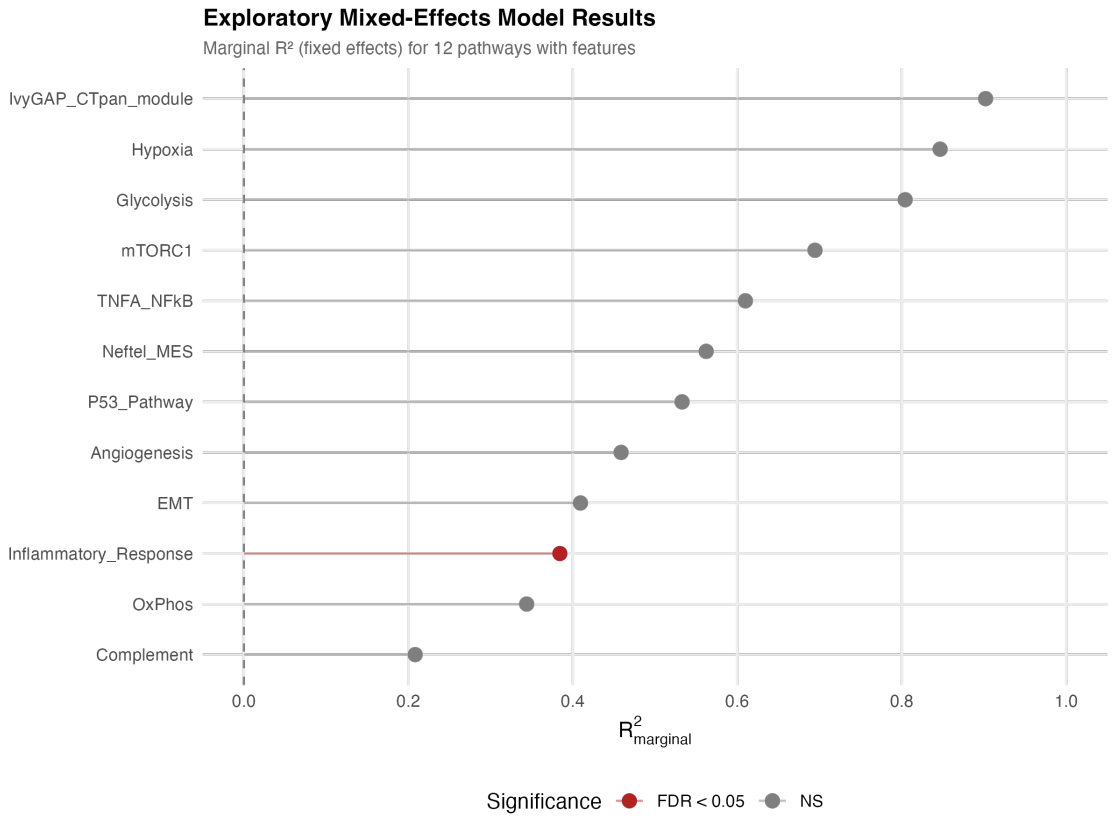


Figure S2. Standardized coefficients for the five radiomic features in the Inflammatory Response LMM. Points represent standardized beta coefficients; horizontal error bars indicate 95% confidence intervals. The dashed vertical line marks zero. Red points denote features significant after Holm correction ($p < 0.05$); gray points denote non-significant features (NS). All five features are T2-derived. $R^2_m = 0.384$, FDR = 0.012. LMM = linear mixed-effects model; GLSZM = Gray-Level Size Zone Matrix; CoLIAGe = Co-occurrence of Local Anisotropic Gradient Orientations.

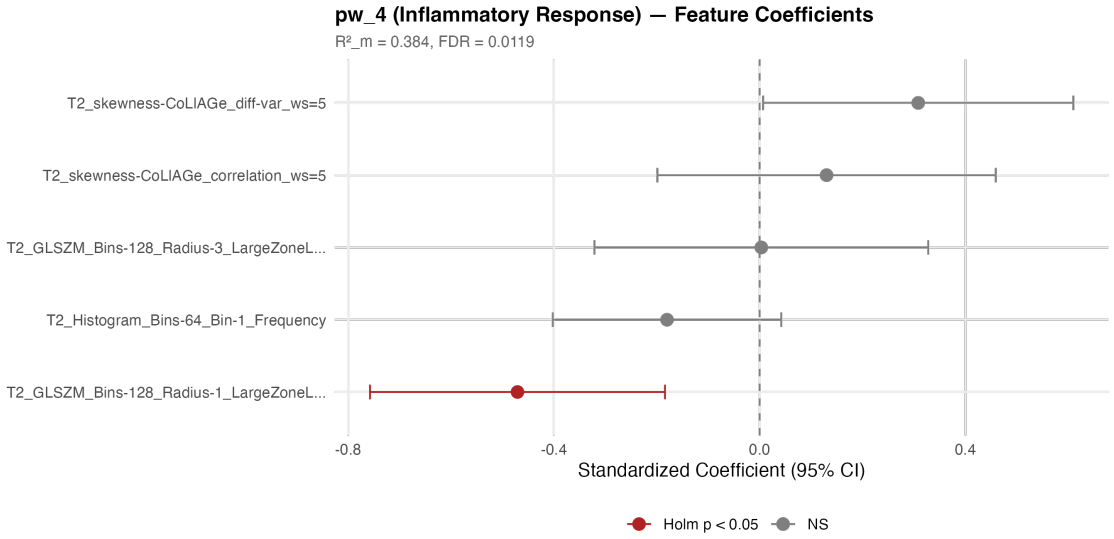


Figure S3. Residual diagnostics for the Inflammatory Response LMM. **(A)** Normal Q-Q plot of standardized residuals. Blue points represent individual observations; the red line indicates the theoretical normal distribution. **(B)** Cook's distance for each observation. Red bars indicate observations exceeding the conventional influence threshold of 4/n = 0.080 (dashed red horizontal line); blue bars indicate non-influential observations. One observation exceeded Cook's D = 1.0. **(C)** Residuals versus fitted values, colored by MRI subcompartment: blue = enhancing tumor (ET), orange = non-enhancing tumor (NET), green = peritumoral edema (ED). The dashed horizontal line marks zero. No systematic pattern is evident. LMM = linear mixed-effects model.

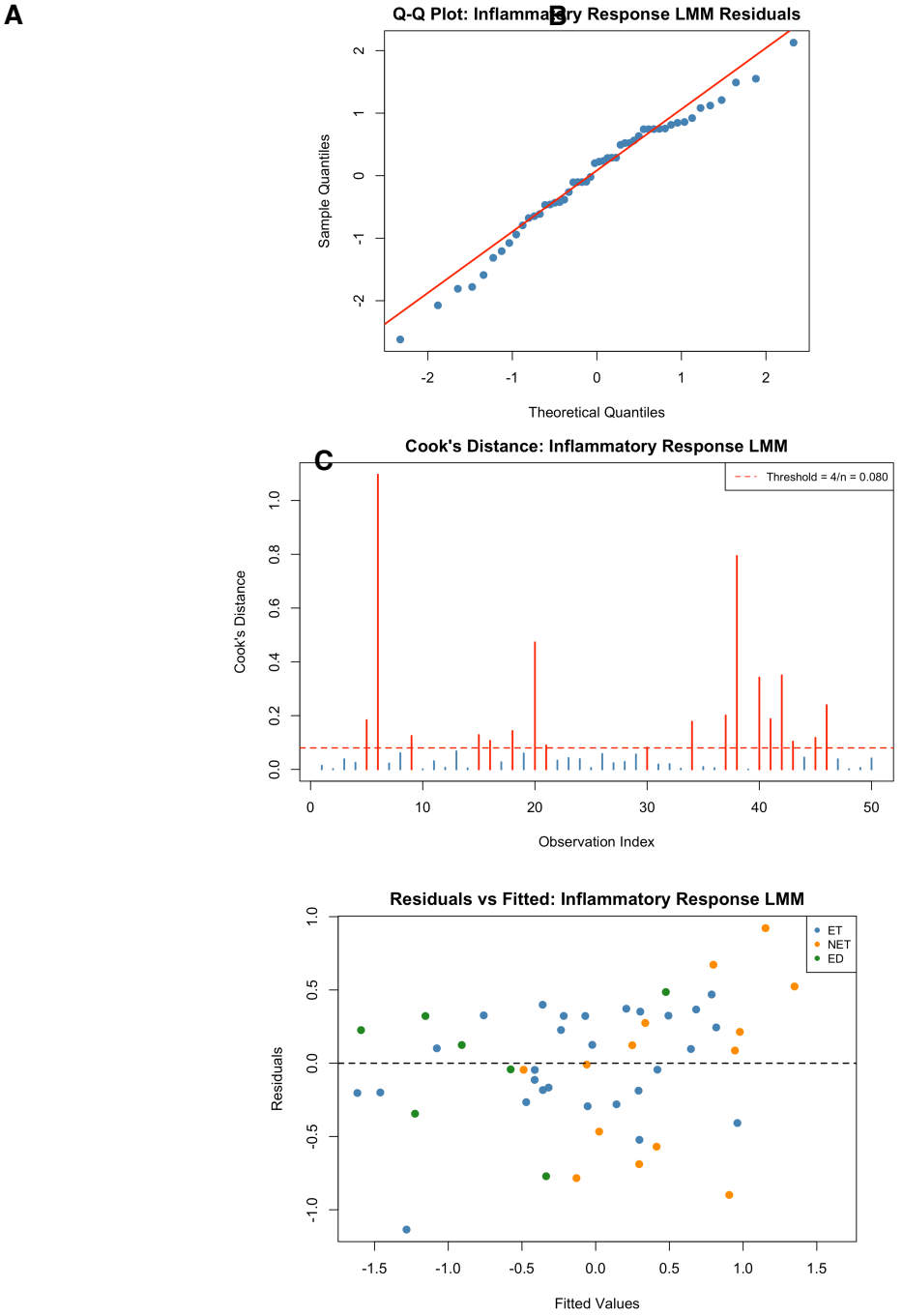


Figure S4. Permutation null distribution for the Inflammatory Response LMM. The histogram shows the distribution of likelihood ratio test (LRT) chi-squared statistics obtained from 1000 permutations of the pathway enrichment scores. The dashed red vertical line indicates the observed LRT chi-squared = 20.53 (permutation p = 0.055). LMM = linear mixed-effects model; LRT = likelihood ratio test.

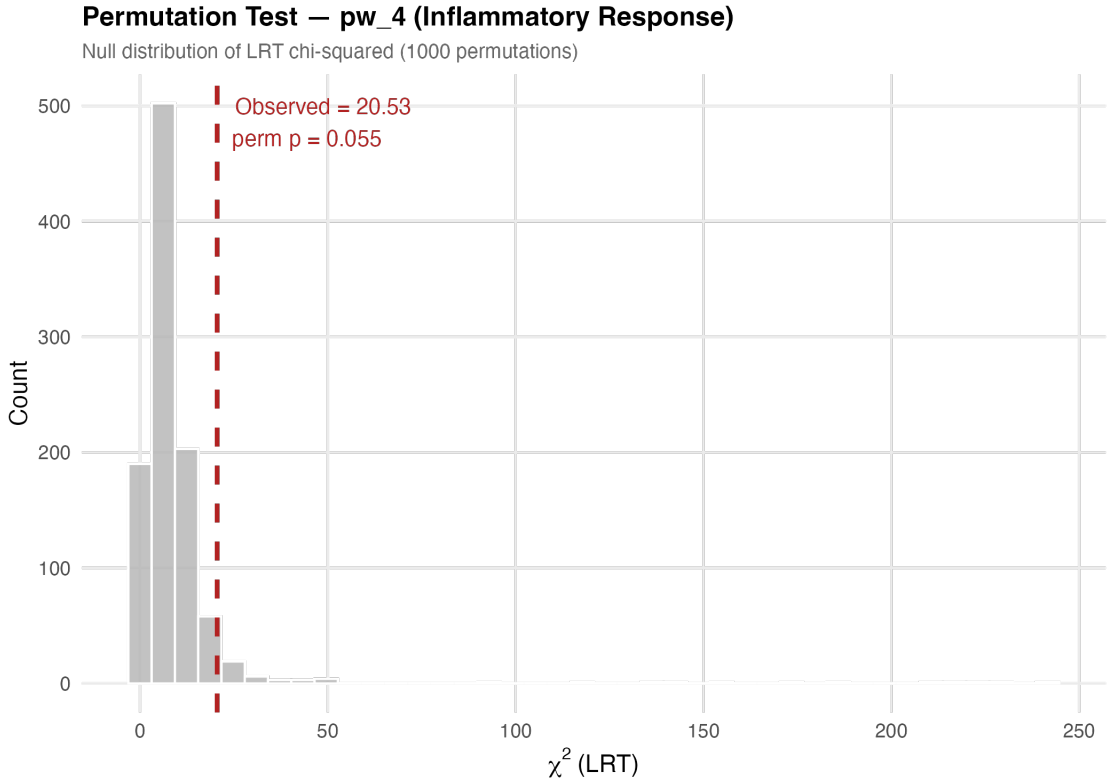


Figure S5. Sensitivity analysis for the Inflammatory Response LMM across seven analysis variants. Each point represents the LRT p-value for a given variant, plotted as -$\log_{10}$(p). Red points indicate p < 0.05; the gray point indicates p >= 0.05. The dashed vertical line marks p = 0.05. Variants: Primary = default analysis; S1a = ET mapped to Cellular Tumor only (excluding MVP); S1b = conservative zone mapping; S2 = patients with >= 2 subcompartments only; S3 = median aggregation (instead of mean); S7 = global standardization; S8 = Kenward-Roger denominator degrees of freedom. Six of seven variants retained significance. LMM = linear mixed-effects model; LRT = likelihood ratio test; ET = enhancing tumor; MVP = microvascular proliferation.

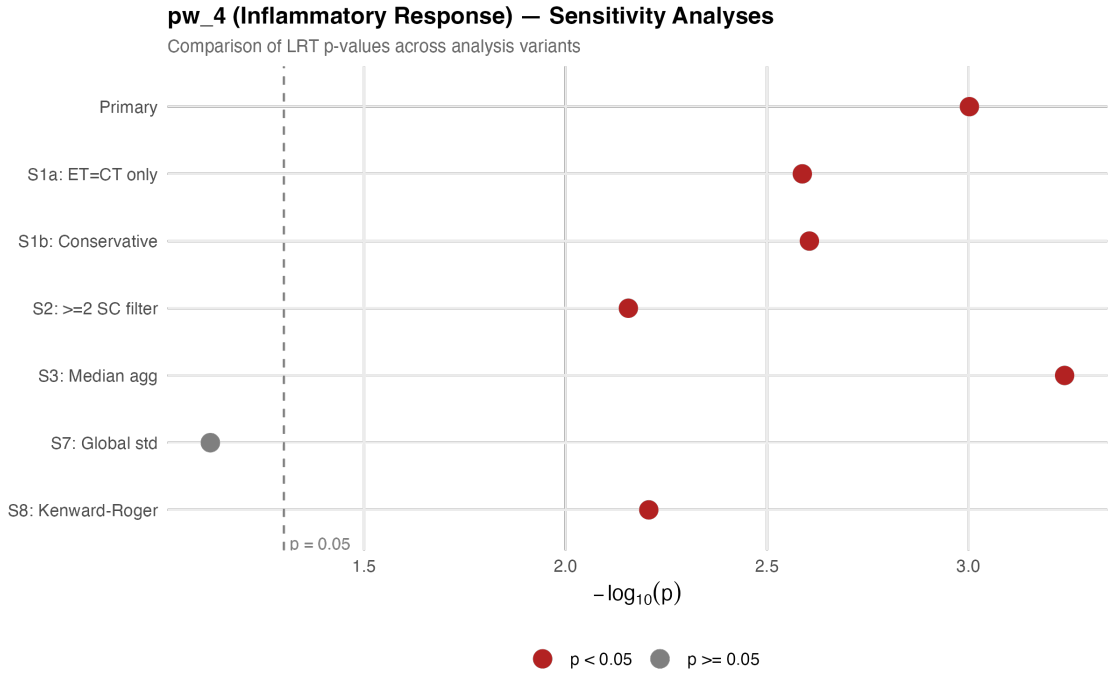


Figure S6. Heatmap of Spearman correlations between the top 30 radiomic features and 12 pathway enrichment scores that retained features after univariate screening. Color scale represents Spearman rho (red = positive correlation, blue = negative correlation). Hierarchical clustering dendrograms (Ward's method) are applied to both rows (features) and columns (pathways). FDR significance is annotated in the color bar (< 0.05 vs. >= 0.05). Feature names follow the convention: MRI sequence, texture class, extraction parameters. FDR = false discovery rate.

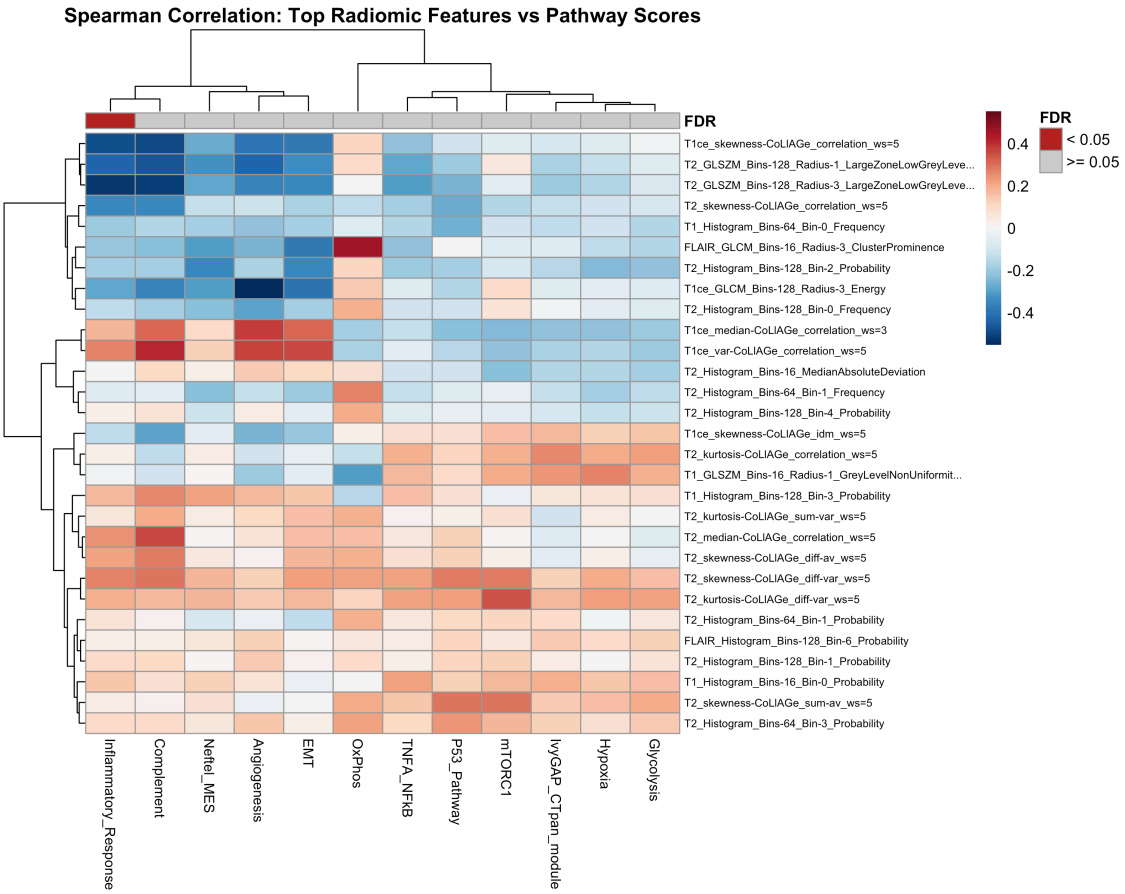


Figure S7. Volcano plot of univariate radiomic-transcriptomic associations for the Inflammatory Response pathway. Each point represents one of 597 candidate radiomic features after unsupervised filtering. The x-axis shows the maximum absolute Spearman rho across the three MRI subcompartments; the y-axis shows -$\log_{10}$(FDR). Red points indicate FDR < 0.05 (1 feature); orange points indicate FDR < 0.10 (24 features); gray points indicate non-significant features. Dashed horizontal lines mark the FDR < 0.05 and FDR < 0.10 thresholds. FDR = false discovery rate.

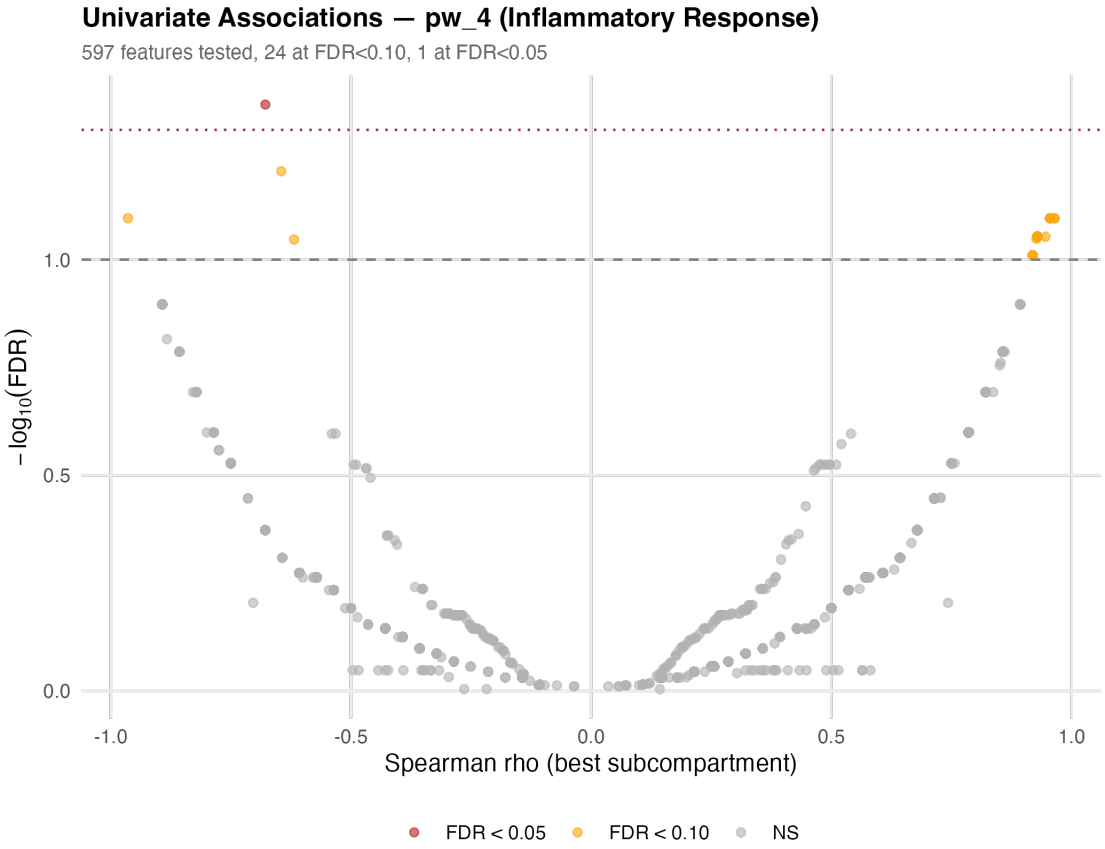


Figure S8. Cohort structure and data availability. Heatmap showing the number of zone-aggregated RNA-seq samples per patient (rows) per MRI subcompartment (columns). Color intensity is proportional to sample count (white = 0, dark blue = 8); numbers within cells indicate exact counts. All 28 patients with matched IvyGAP transcriptomic and IVYGAP-RADIOMICS data are shown. ET = enhancing tumor (mapped from Cellular Tumor and Microvascular Proliferation IvyGAP zones); NET = non-enhancing tumor (mapped from Pseudopalisading Necrosis); ED = peritumoral edema (mapped from Infiltrating Tumor and Leading Edge).

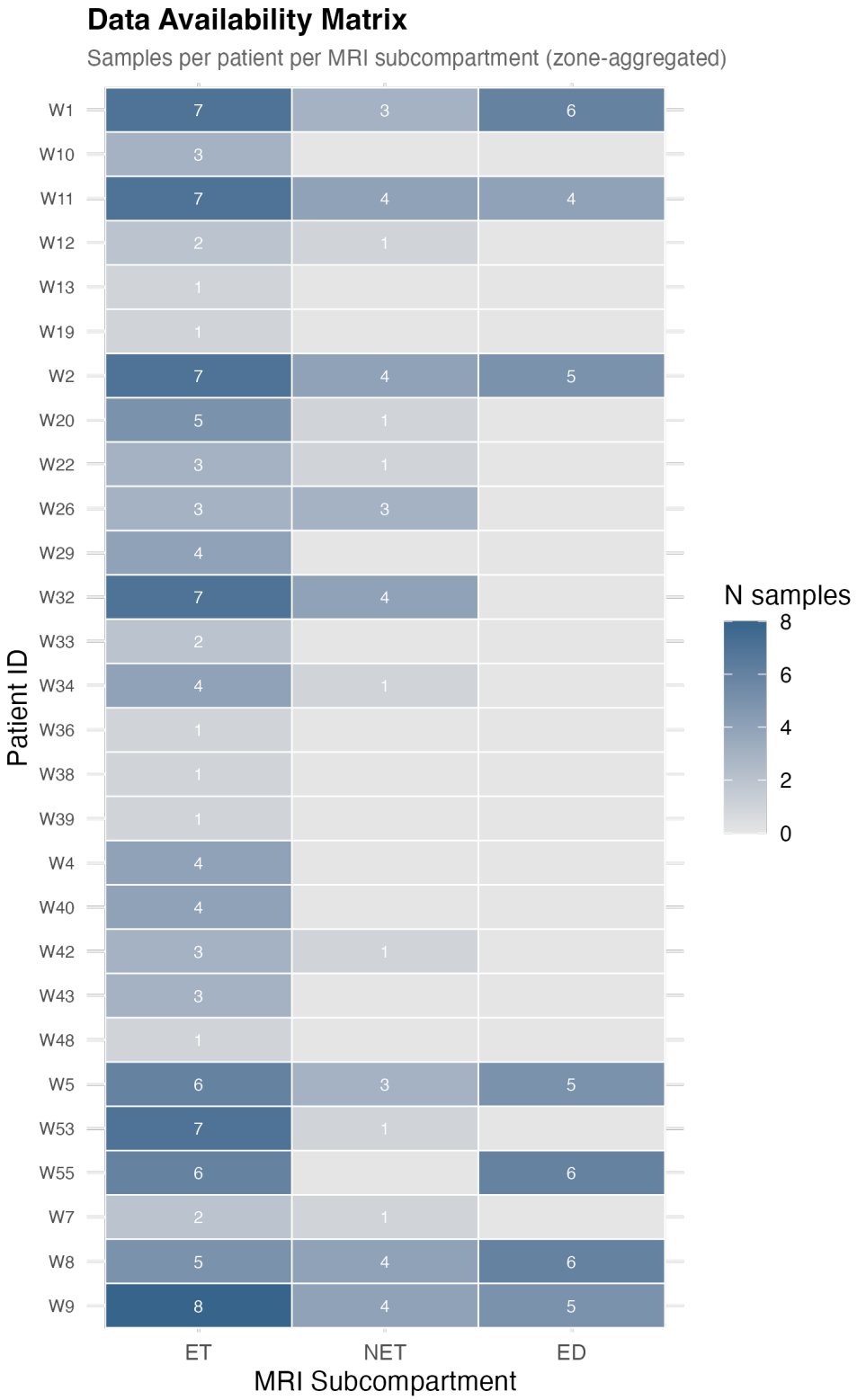


Figure S9. Unsupervised radiomic feature reduction pipeline. Bar chart showing progressive dimensionality reduction across four stages: 3920 raw IBSI-compliant features per subcompartment, 3860 after near-zero-variance (NZV) filtering, 597 after Spearman correlation pruning (|r| > 0.90, retaining one feature per correlated cluster), and 29 unique features entering final models after pathway-specific univariate screening. NZV = near-zero-variance; IBSI = Image Biomarker Standardisation Initiative.

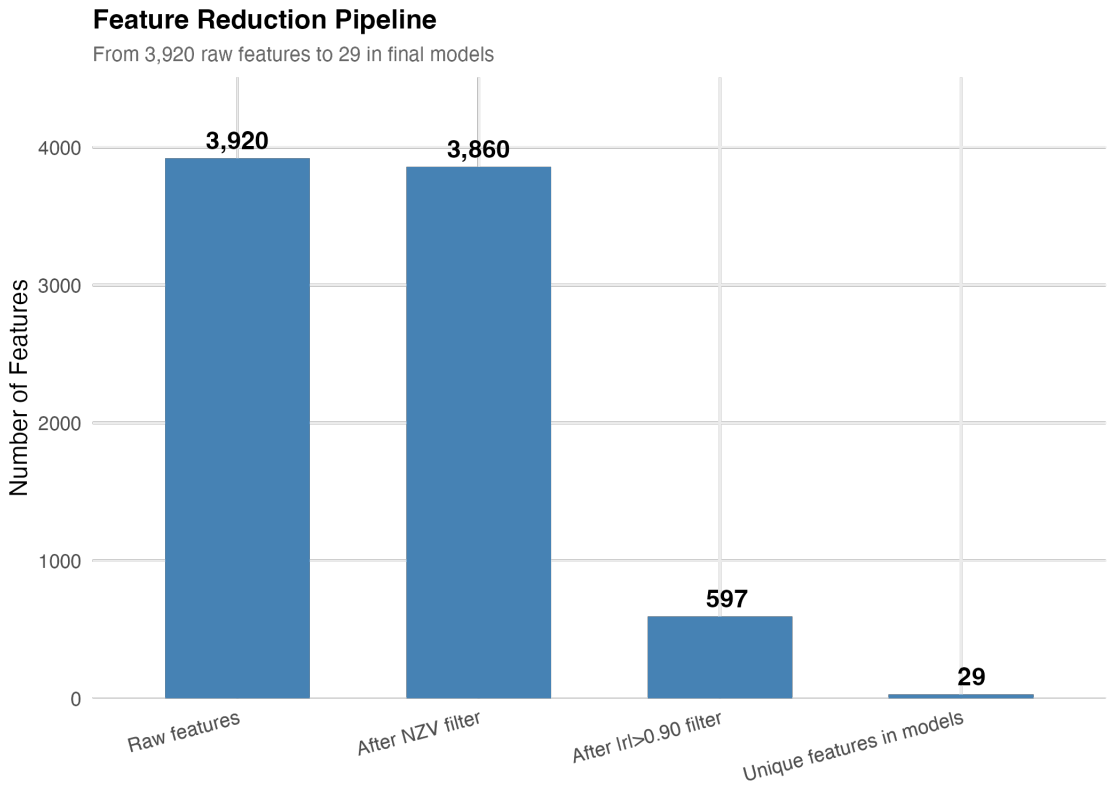


Table S1. Nested cross-validation results for all 24 pathways (R^2_{cv}, MAE, stable features). Pathways are ordered by descending R^2_{cv}. Feature selection was performed independently inside each LOPO fold (no data leakage). Stable features = features selected in >50% of folds.

Pathway	R^2_{cv}	MAE	Spearman rho	Stable features	Median features/fold
Angiogenesis	0.209	0.702	0.581	5	5
Inflammatory Response	0.185	0.674	0.524	5	5
IvyGAP CTpan module	0.133	0.740	0.348	4	5
Complement	-0.004	0.680	0.194	5	5
EMT	-0.034	0.713	-0.607	1	1
IvyGAP IT module	-0.072	0.833	-0.784	1	1
P53 Pathway	-0.075	0.762	-1.000	0	0
E2F Targets	-0.075	0.823	-1.000	0	0
G2M Checkpoint	-0.075	0.831	-1.000	0	0
mTORC1 Signaling	-0.075	0.899	-1.000	0	0
Neftel NPC	-0.075	0.881	-1.000	0	0
IvyGAP CT module	-0.075	0.796	-1.000	0	0
MYC Targets V1	-0.075	0.753	-1.000	0	0
Oxidative Phosphorylation	-0.075	0.833	-1.000	0	0
Neftel AC	-0.075	0.789	-1.000	0	0
Neftel OPC	-0.086	0.820	-0.968	0	0
Glycolysis	-0.109	0.822	-1.000	0	0
TNFA/NF-kB	-0.113	0.795	-0.990	0	0
IvyGAP LE module	-0.114	0.828	-0.773	1	1
Hypoxia	-0.135	0.813	-0.443	0	0.5
Neftel MES	-0.146	0.790	-0.778	0	0
IL6/JAK/STAT3	-0.199	0.894	-0.748	3	3
IFN Gamma Response	-0.210	0.868	-0.806	0	0
IvyGAP CTmvp module	-0.274	0.893	-0.582	2	4

Table S2. Status of all 24 pathways in the associational analysis (12 tested in LMM, 12 with zero features passing univariate filter, 1 FDR-significant). Pathways with zero univariate features were not tested in the LMM.

Pathway	Category	Univariate features	Status	R^2_m	LRT p	FDR
Hypoxia	Hallmark	46	Tested in LMM	0.847	0.085	0.223
Angiogenesis	Hallmark	1	Tested in LMM	0.459	0.053	0.223
EMT	Hallmark	21	Tested in LMM	0.409	0.176	0.234
Inflammatory Response	Hallmark	24	FDR < 0.05	0.384	0.001	0.012
TNFA/NF-kB	Hallmark	22	Tested in LMM	0.610	0.352	0.423
IL6/JAK/STAT3	Hallmark	0	No features at FDR < 0.10	—	—	—
IFN Gamma Response	Hallmark	0	No features at FDR < 0.10	—	—	—
P53 Pathway	Hallmark	23	Tested in LMM	0.533	0.101	0.223
MYC Targets V1	Hallmark	0	No features at FDR < 0.10	—	—	—
E2F Targets	Hallmark	0	No features at FDR < 0.10	—	—	—
G2M Checkpoint	Hallmark	0	No features at FDR < 0.10	—	—	—
mTORC1 Signaling	Hallmark	3	Tested in LMM	0.694	0.133	0.223
Glycolysis	Hallmark	21	Tested in LMM	0.804	0.112	0.223
Oxidative Phosphorylation	Hallmark	3	Tested in LMM	0.344	0.800	0.800
Complement	Hallmark	21	Tested in LMM	0.208	0.148	0.223
Neftel MES	Neftel	2	Tested in LMM	0.562	0.133	0.223
Neftel AC	Neftel	0	No features at FDR < 0.10	—	—	—
Neftel OPC	Neftel	0	No features at FDR < 0.10	—	—	—
Neftel NPC	Neftel	0	No features at FDR < 0.10	—	—	—
IvyGAP CT module	IvyGAP Zone	0	No features at FDR < 0.10	—	—	—
IvyGAP CTmvp module	IvyGAP Zone	0	No features at FDR < 0.10	—	—	—
IvyGAP CTpan module	IvyGAP Zone	5	Tested in LMM	0.902	0.773	0.800
IvyGAP IT module	IvyGAP Zone	0	No features at FDR < 0.10	—	—	—
IvyGAP LE module	IvyGAP Zone	0	No features at FDR < 0.10	—	—	—

Table S3. Full sensitivity analysis results for the Inflammatory Response pathway across all analysis variants. R^2_m = marginal R^2 (fixed effects only); LRT p = likelihood ratio test p-value comparing full model (radiomic features + subcompartment) to null model (subcompartment only). All variants use the same five radiomic features except where noted.

Variant	Description	R^2_m	LRT p	N patients	Significant (p < 0.05)
Primary	Default analysis	0.384	0.0010	28	Yes
S1a	ET = CT only (excluding MVP)	0.365	0.0026	28	Yes
S1b	Conservative zone mapping	0.421	0.0025	28	Yes
S2	Patients with >= 2 subcompartments only	0.412	0.0070	16	Yes
S3	Median aggregation (instead of mean)	0.366	0.0006	28	Yes
S7	Global standardization (instead of within-subcompartment)	0.317	0.0761	28	No
S8	Kenward-Roger denominator degrees of freedom	0.384	0.0062	28	Yes

Six of seven variants retained significance (LRT $p < 0.05$). The sole exception (S7, global standardization) approached significance ($p = 0.076$), consistent with attenuated signal when between-subcompartment variance is not removed prior to modeling.

Table S4. Gene set pairwise Jaccard similarity matrix (24 x 24). Values represent the Jaccard index (intersection/union of gene members) between gene sets. Higher values indicate greater overlap. Most pairs show minimal overlap (J < 0.10), confirming that the 24 gene sets capture largely distinct biological programs. Notable exceptions: IvyGAP IT and LE modules (J = 0.653); E2F Targets and G2M Checkpoint (J = 0.223); Glycolysis and Hypoxia (J = 0.194); IvyGAP CTpan and Hypoxia (J = 0.146); Inflammatory Response and TNFA/NF-kB (J = 0.146).

	Hyp	Ang	EMT	IR	TNF	IL6	IFN	P53	MYC	E2F	G2M	mTO	Gly	OxP	Com	MES	AC	OPC	NPC	CT	mvp	pan	IT	LE
Hyp	1.000	.013	.047	.028	.078	.014	.020	.036	.005	.005	.008	.084	.194	.005	.028	.037	.025	.000	.000	.011	.018	.146	.003	.000
Ang	.013	1.000	.054	.013	.017	.017	.000	.022	.000	.000	.000	.004	.013	.000	.009	.047	.000	.000	.000	.000	.017	.009	.000	.000
EMT	.047	.054	1.000	.034	.058	.029	.013	.018	.003	.003	.005	.005	.036	.000	.031	.139	.004	.004	.008	.004	.075	.047	.005	.003
IR	.028	.013	.034	1.000	.146	.087	.102	.031	.003	.005	.008	.013	.013	.000	.055	.037	.012	.000	.000	.000	.010	.026	.008	.008
TNF	.078	.017	.058	.146	1.000	.067	.072	.070	.003	.005	.013	.026	.013	.000	.042	.033	.008	.000	.000	.000	.003	.081	.000	.000
IL6	.014	.017	.029	.087	.067	1.000	.083	.025	.000	.000	.004	.004	.011	.000	.025	.022	.000	.000	.000	.000	.011	.025	.000	.000
IFN	.020	.000	.013	.102	.072	.083	1.000	.028	.005	.005	.008	.020	.008	.003	.061	.012	.004	.000	.000	.000	.008	.020	.000	.000
P53	.036	.022	.018	.031	.070	.025	.028	1.000	.008	.018	.003	.028	.013	.010	.010	.012	.004	.000	.000	.004	.005	.036	.010	.008
MYC	.005	.000	.003	.003	.003	.000	.005	.008	1.000	.096	.078	.050	.013	.028	.000	.000	.000	.000	.000	.000	.000	.005	.000	.000
E2F	.005	.000	.003	.005	.005	.000	.005	.018	.096	1.000	.223	.036	.018	.000	.000	.000	.000	.000	.004	.004	.000	.003	.000	.000
G2M	.008	.000	.005	.008	.013	.004	.008	.003	.078	.223	1.000	.023	.018	.000	.000	.008	.004	.000	.008	.015	.000	.003	.000	.000
mTO	.084	.004	.005	.013	.026	.004	.020	.028	.050	.036	.023	1.000	.072	.018	.013	.012	.004	.000	.004	.007	.005	.072	.003	.003
Gly	.194	.013	.036	.013	.013	.011	.008	.013	.013	.018	.018	.072	1.000	.020	.013	.020	.008	.000	.004	.004	.005	.070	.000	.000
OxP	.005	.000	.000	.000	.000	.000	.003	.010	.028	.000	.000	.018	.020	1.000	.008	.000	.000	.000	.000	.000	.000	.003	.003	.003
Com	.028	.009	.031	.055	.042	.025	.061	.010	.000	.000	.000	.013	.013	.008	1.000	.028	.012	.000	.000	.000	.015	.018	.008	.005
MES	.037	.047	.139	.037	.033	.022	.012	.012	.000	.000	.008	.012	.020	.000	.028	1.000	.000	.000	.000	.000	.037	.050	.000	.000
AC	.025	.000	.004	.012	.008	.000	.004	.004	.000	.000	.004	.004	.008	.000	.012	.000	1.000	.000	.000	.075	.000	.012	.012	.004
OPC	.000	.000	.004	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	1.000	.000	.084	.000	.000	.012	.004
NPC	.000	.000	.008	.000	.000	.000	.000	.000	.000	.004	.008	.004	.004	.000	.000	.000	.000	.000	1.000	.008	.000	.004	.045	.045
CT	.011	.000	.004	.000	.000	.000	.000	.004	.000	.004	.015	.007	.004	.000	.000	.000	.075	.084	.008	1.000	.000	.000	.007	.000
mvp	.018	.017	.075	.010	.003	.011	.008	.005	.000	.000	.000	.005	.005	.000	.015	.037	.000	.000	.000	.000	1.000	.000	.005	.000
pan	.146	.009	.047	.026	.081	.025	.020	.036	.005	.003	.003	.072	.070	.003	.018	.050	.012	.000	.004	.000	.000	1.000	.000	.000
IT	.003	.000	.005	.008	.000	.000	.000	.010	.000	.000	.000	.003	.000	.003	.008	.000	.012	.012	.045	.007	.005	.000	1.000	.653
LE	.000	.000	.003	.008	.000	.000	.000	.008	.000	.000	.000	.003	.000	.003	.005	.000	.004	.004	.045	.000	.000	.000	.653	1.000

Column abbreviations: Hyp = Hypoxia, Ang = Angiogenesis, EMT = Epithelial-Mesenchymal Transition, IR = Inflammatory Response, TNF = TNFA/NF-kB, IL6 = IL6/JAK/STAT3, IFN = IFN Gamma Response, P53 = P53 Pathway, MYC = MYC Targets V1, E2F = E2F Targets, G2M = G2M Checkpoint, mTO = mTORC1 Signaling, Gly = Glycolysis, OxP = Oxidative Phosphorylation, Com = Complement, MES = Neftel MES, AC = Neftel AC, OPC = Neftel OPC, NPC = Neftel NPC, CT = IvyGAP CT module, mvp = IvyGAP CTmvp module, pan = IvyGAP CTpan module, IT = IvyGAP IT module, LE = IvyGAP LE module.

Table S5. Legacy pre-screened Elastic Net results compared with nested CV results. The legacy analysis pre-selected features on the full dataset before LOPO-CV, introducing data leakage. Only pathways with $R^2_{cv} > 0$ in either analysis are shown.

Pathway	Legacy pre-screened R^2	Legacy S5 R^2	Nested R^2_{cv}
Angiogenesis	—	—	0.209
Inflammatory Response	-0.104	0.180	0.185
IvyGAP CTpan module	—	—	0.133

Table S6. Feature lookup table mapping radiomic feature indices to full IBSI names, MRI sequence, and feature type. All 89 unique radiomic features that entered any model across all 24 pathways are listed.

Feature ID	IBSI Feature Name	Sequence	Feature Type
rad_71	T1 kurtosis-CoLIAGe correlation ws=5	T1	Higher-order (CoLIAGe)
rad_126	T1 Histogram Bins-16 Bin-0 Probability	T1	First-order (Histogram)
rad_163	T1 Histogram Bins-16 Kurtosis	T1	First-order (Histogram)
rad_166	T1 Histogram Bins-16 MeanAbsoluteDeviation	T1	First-order (Histogram)
rad_227	T1 GLSZM Bins-16 Radius-1 GreyLevelNonUniformityNormalized	T1	Texture (GLSZM)
rad_240	T1 GLSZM Bins-16 Radius-1 ZoneSizeNonUniformity	T1	Texture (GLSZM)
rad_249	T1 Histogram Bins-64 Bin-0 Frequency	T1	First-order (Histogram)
rad_418	T1 GLRLM Bins-64 Radius-1 LongRunLowGreyLevelEmphasis	T1	Texture (GLRLM)
rad_439	T1 GLSZM Bins-64 Radius-1 ZoneSizeNonUniformity	T1	Texture (GLSZM)
rad_565	T1 Histogram Bins-128 Bin-3 Probability	T1	First-order (Histogram)
rad_601	T1 Histogram Bins-128 Bin-56 Probability	T1	First-order (Histogram)
rad_611	T1 Histogram Bins-128 Bin-60 Probability	T1	First-order (Histogram)
rad_733	T1 GLCM Bins-128 Radius-1 Entropy	T1	Texture (GLCM)
rad_1000	T2 skewness-CoLIAGe correlation ws=3	T2	Higher-order (CoLIAGe)
rad_1049	T2 median-CoLIAGe correlation ws=5	T2	Higher-order (CoLIAGe)
rad_1051	T2 kurtosis-CoLIAGe correlation ws=5	T2	Higher-order (CoLIAGe)
rad_1052	T2 skewness-CoLIAGe correlation ws=5	T2	Higher-order (CoLIAGe)
rad_1062	T2 var-CoLIAGe sum-av ws=5	T2	Higher-order (CoLIAGe)
rad_1064	T2 skewness-CoLIAGe sum-av ws=5	T2	Higher-order (CoLIAGe)
rad_1067	T2 kurtosis-CoLIAGe sum-var ws=5	T2	Higher-order (CoLIAGe)
rad_1074	T2 var-CoLIAGe diff-av ws=5	T2	Higher-order (CoLIAGe)
rad_1076	T2 skewness-CoLIAGe diff-av ws=5	T2	Higher-order (CoLIAGe)
rad_1079	T2 kurtosis-CoLIAGe diff-var ws=5	T2	Higher-order (CoLIAGe)
rad_1080	T2 skewness-CoLIAGe diff-var ws=5	T2	Higher-order (CoLIAGe)
rad_1148	T2 Histogram Bins-16 MedianAbsoluteDeviation	T2	First-order (Histogram)
rad_1220	T2 GLSZM Bins-16 Radius-1 ZoneSizeNonUniformity	T2	Texture (GLSZM)
rad_1226	T2 NGTDM Contrast	T2	Texture (NGTDM)

Feature ID	IBSI Feature Name	Sequence	Feature Type
rad_1251	T2 Histogram Bins-64 Bin-1 Frequency	T2	First-order (Histogram)
rad_1252	T2 Histogram Bins-64 Bin-1 Probability	T2	First-order (Histogram)
rad_1296	T2 Histogram Bins-64 Bin-3 Probability	T2	First-order (Histogram)
rad_1363	T2 Histogram Bins-64 Kurtosis	T2	First-order (Histogram)
rad_1422	T2 Histogram Bins-128 Bin-0 Frequency	T2	First-order (Histogram)
rad_1501	T2 Histogram Bins-128 Bin-1 Probability	T2	First-order (Histogram)
rad_1523	T2 Histogram Bins-128 Bin-2 Probability	T2	First-order (Histogram)
rad_1567	T2 Histogram Bins-128 Bin-4 Probability	T2	First-order (Histogram)
rad_1707	T2 GLCM Bins-128 Radius-1 AutoCorrelation	T2	Texture (GLCM)
rad_1732	T2 GLSZM Bins-128 Radius-1 LargeZoneLowGreyLevelEmphasis	T2	Texture (GLSZM)
rad_1763	T2 GLSZM Bins-16 Radius-2 GreyLevelNonUniformityNormalized	T2	Texture (GLSZM)
rad_1872	T2 GLSZM Bins-16 Radius-3 GreyLevelNonUniformityNormalized	T2	Texture (GLSZM)
rad_1920	T2 GLSZM Bins-64 Radius-3 ZoneSizeEntropy	T2	Texture (GLSZM)
rad_1930	T2 GLCM Bins-128 Radius-3 Energy	T2	Texture (GLCM)
rad_1931	T2 GLCM Bins-128 Radius-3 Entropy	T2	Texture (GLCM)
rad_1945	T2 GLSZM Bins-128 Radius-3 GreyLevelNonUniformityNormalized	T2	Texture (GLSZM)
rad_1950	T2 GLSZM Bins-128 Radius-3 LargeZoneLowGreyLevelEmphasis	T2	Texture (GLSZM)
rad_1956	T2 GLSZM Bins-128 Radius-3 ZoneSizeEntropy	T2	Texture (GLSZM)
rad_1977	T1ce median-CoLIAGe correlation ws=3	T1ce	Higher-order (CoLIAGe)
rad_1978	T1ce var-CoLIAGe correlation ws=3	T1ce	Higher-order (CoLIAGe)
rad_1979	T1ce kurtosis-CoLIAGe correlation ws=3	T1ce	Higher-order (CoLIAGe)
rad_2002	T1ce var-CoLIAGe diff-av ws=3	T1ce	Higher-order (CoLIAGe)
rad_2003	T1ce kurtosis-CoLIAGe diff-av ws=3	T1ce	Higher-order (CoLIAGe)
rad_2007	T1ce kurtosis-CoLIAGe diff-var ws=3	T1ce	Higher-order (CoLIAGe)
rad_2023	T1ce kurtosis-CoLIAGe inertia ws=5	T1ce	Higher-order (CoLIAGe)
rad_2028	T1ce skewness-CoLIAGe idm ws=5	T1ce	Higher-order (CoLIAGe)
rad_2030	T1ce var-CoLIAGe correlation ws=5	T1ce	Higher-order (CoLIAGe)
rad_2031	T1ce kurtosis-CoLIAGe correlation ws=5	T1ce	Higher-order (CoLIAGe)

Feature ID	IBSI Feature Name	Sequence	Feature Type
rad_2032	T1ce skewness-CoLIAGe correlation ws=5	T1ce	Higher-order (CoLIAGe)
rad_2044	T1ce skewness-CoLIAGe sum-av ws=5	T1ce	Higher-order (CoLIAGe)
rad_2048	T1ce skewness-CoLIAGe sum-var ws=5	T1ce	Higher-order (CoLIAGe)
rad_2055	T1ce kurtosis-CoLIAGe diff-av ws=5	T1ce	Higher-order (CoLIAGe)
rad_2056	T1ce skewness-CoLIAGe diff-av ws=5	T1ce	Higher-order (CoLIAGe)
rad_2073	T1ce Intensity MedianAbsoluteDeviation	T1ce	Other
rad_2074	T1ce Intensity Minimum	T1ce	Other
rad_2077	T1ce Intensity QuartileCoefficientOfVariation	T1ce	Other
rad_2134	T1ce Histogram Bins-16 QuartileCoefficientOfVariation	T1ce	First-order (Histogram)
rad_2198	T1ce GLSZM Bins-16 Radius-1 ZoneSizeEntropy	T1ce	Texture (GLSZM)
rad_2209	T1ce Histogram Bins-64 Bin-0 Frequency	T1ce	First-order (Histogram)
rad_2397	T1ce GLSZM Bins-64 Radius-1 ZoneSizeEntropy	T1ce	Texture (GLSZM)
rad_2481	T1ce Histogram Bins-128 Bin-1 Probability	T1ce	First-order (Histogram)
rad_2495	T1ce Histogram Bins-128 Bin-26 Probability	T1ce	First-order (Histogram)
rad_2511	T1ce Histogram Bins-128 Bin-33 Probability	T1ce	First-order (Histogram)
rad_2635	T1ce Histogram Bins-128 Bin-8 Probability	T1ce	First-order (Histogram)
rad_2662	T1ce Histogram Bins-128 FifthPercentileMean	T1ce	First-order (Histogram)
rad_2711	T1ce GLSZM Bins-128 Radius-1 LargeZoneHighGreyLevelEmphasis	T1ce	Texture (GLSZM)
rad_2718	T1ce GLSZM Bins-128 Radius-1 ZoneSizeEntropy	T1ce	Texture (GLSZM)
rad_2794	T1ce GLSZM Bins-64 Radius-2 ZoneSizeNoneUniformityNormalized	T1ce	Texture (GLSZM)
rad_2910	T1ce GLCM Bins-128 Radius-3 Energy	T1ce	Texture (GLCM)
rad_3064	FLAIR Intensity Variance	FLAIR	Other
rad_3066	FLAIR Histogram Bins-16 Bin-0 Probability	FLAIR	First-order (Histogram)
rad_3131	FLAIR Morphologic EllipseDiameter Axis-2	FLAIR	Other
rad_3180	FLAIR GLSZM Bins-16 Radius-1 ZoneSizeNonUniformity	FLAIR	Texture (GLSZM)
rad_3234	FLAIR Histogram Bins-64 Bin-2 Probability	FLAIR	First-order (Histogram)
rad_3375	FLAIR GLSZM Bins-64 Radius-1 SmallZoneLowGreyLevelEmphasis	FLAIR	Texture (GLSZM)
rad_3549	FLAIR Histogram Bins-128 Bin-5 Probability	FLAIR	First-order (Histogram)
rad_3571	FLAIR Histogram Bins-128 Bin-6 Probability	FLAIR	First-order (Histogram)
rad_3647	FLAIR Histogram Bins-128 MeanAbsoluteDeviation	FLAIR	First-order (Histogram)
rad_3813	FLAIR GLCM Bins-16 Radius-3 ClusterProminence	FLAIR	Texture (GLCM)

Feature ID	IBSI Feature Name	Sequence	Feature Type
rad_3848	FLAIR LBP Radius-3 LBP	FLAIR	Other
rad_3880	FLAIR GLSZM Bins-64 Radius-3 ZoneSizeEntropy	FLAIR	Texture (GLSZM)
rad_3909	FLAIR GLSZM Bins-128 Radius-3 LargeZoneHighGreyLevelEmphasis	FLAIR	Texture (GLSZM)

Table S7. Clinical covariate adjustment results for the Inflammatory Response pathway (sensitivity analysis S9; Section 2.9). Progressive covariate adjustment demonstrating that radiomic features remain significant after accounting for age, MGMT methylation status, and molecular subtype. LRT tests the contribution of the five radiomic features above the covariates included in each model.

Model	Covariates	R^2_m	R^2_c	LRT chi²	LRT p	N obs	N patients	Converged
A: Primary	Subcompartment only	0.384	0.687	20.53	0.0010	50	28	Yes
B: + age	Subcompartment + age	0.378	0.702	20.03	0.0012	50	28	Yes
C: + age + MGMT	Subcompartment + age + MGMT	0.369	0.719	19.26	0.0017	50	28	Yes
D: + age + MGMT + subtype	Subcompartment + age + MGMT + molecular subtype	0.479	0.785	17.51	0.0036	50	28	Yes

Radiomic features remain significant (LRT $p < 0.005$) in all four models. R^2_m = marginal R^2 (fixed effects only); R^2_c = conditional R^2 (fixed + random effects); LRT = likelihood ratio test comparing full model (with radiomic features) to reduced model (covariates only).

Individual radiomic feature coefficients across adjustment models:

Model A: Primary (no clinical covariates)

Feature	IBSI Name	Beta	SE	95% CI	p (Holm)
rad_1732	T2 GLSZM Bins-128 Radius-1 LargeZoneLowGreyLevelEmphasis	-0.471	0.141	[-0.758, -0.184]	0.010 *
rad_1080	T2 skewness-CoLIAGe diff-var ws=5	0.308	0.149	[0.007, 0.610]	0.181
rad_1251	T2 Histogram Bins-64 Bin-1 Frequency	-0.180	0.110	[-0.402, 0.042]	0.326
rad_1052	T2 skewness-CoLIAGe correlation ws=5	0.130	0.163	[-0.199, 0.459]	0.858
rad_1950	T2 GLSZM Bins-128 Radius-3 LargeZoneLowGreyLevelEmphasis	0.003	0.159	[-0.321, 0.328]	0.984

Model D: Fully adjusted (+ age + MGMT + molecular subtype)

Feature	IBSI Name	Beta	SE	95% CI	p (Holm)
rad_1732	T2 GLSZM Bins-128 Radius-1 LargeZoneLowGreyLevelEmphasis	-0.471	0.138	[-0.753, -0.188]	0.010 *
rad_1080	T2 skewness-CoLIAGe diff-var ws=5	0.295	0.150	[-0.011, 0.601]	0.232
rad_1052	T2 skewness-CoLIAGe correlation ws=5	0.239	0.170	[-0.108, 0.585]	0.510
rad_1251	T2 Histogram Bins-64 Bin-1 Frequency	-0.119	0.108	[-0.339, 0.101]	0.558
rad_1950	T2 GLSZM Bins-128 Radius-3 LargeZoneLowGreyLevelEmphasis	0.068	0.154	[-0.248, 0.383]	0.665

* Holm-adjusted $p < 0.05$. rad_1732 (T2 GLSZM LargeZoneLowGreyLevelEmphasis) is the only individually significant feature, remaining significant across all four adjustment models.

Table S8. CLEAR (CheckList for EvaluAtion of Radiomics research) compliance table. Self-assessment following CLEAR v1.0 (Kwan et al., Insights Imaging 2023;14:75).

Item	Domain	Description	Compliance	Section	Notes
1	Title	Indicate use of radiomics in title	Yes	Title	"Radiomic Features" in title
2	Abstract	Structured summary with methods, results, uncertainty	Yes	Abstract	R^2_{cv} with 95% CIs and permutation p-values
3	Keywords	Keywords indicating radiomics study	Yes	Keywords	"radiomics" and related terms
4	Introduction	Scientific/clinical problem with literature review	Yes	Section 1	GBM heterogeneity, prior work reviewed
5	Introduction	Rationale for radiomic approach	Yes	Section 1	Zone-to-subcompartment mapping rationale
6	Introduction	Study objectives	Yes	Section 1	Hypothesis-generating analysis stated
7	Study Design	Indicate CLEAR checklist use	Yes	Section 2.13	CLEAR compliance declared
8	Study Design	Ethical approval	Yes	IRB Statement	Public datasets, no IRB required
9	Study Design	Sample size calculation	Yes	Section 2.12	Riley criterion: N_{min} = 240, acknowledged underpowered
10	Study Design	Study nature	Yes	Section 2.1	Retrospective analysis of public datasets
11	Study Design	Inclusion/exclusion criteria	Partial	Section 2.1	Patient matching described
12	Study Design	Flowchart of methodology	Yes	Figures 1-2	Data flow and feature reduction shown
13	Data	Data source with links	Yes	Section 2.1	IvyGAP and TCIA cited
14	Data	Prior dataset usage declared	Yes	Section 1	Le et al., Park et al. discussed
15	Data	Data split with leakage prevention	Yes	Sections 2.6-2.7	Nested LOPO-CV with internal feature selection
16	Data	Imaging protocol and scanner	Partial	Section 2.1	Deferred to Pati et al.
17	Data	Non-radiomic predictor variables	Yes	Section 2.8	Age, MGMT, molecular subtype (S9)
18	Data	Reference standard/outcome measure	Yes	Section 2.3	ssGSEA pathway enrichment scores
19	Segmentation	Segmentation software and method	Partial	Section 2.1	BraTS-style from Pati et al.
20	Segmentation	Number of readers and experience	Partial	Section 2.1	Multi-expert per Pati et al.
21	Pre-processing	Pre-processing software and parameters	N/A	Section 2.1	Pre-extracted features
22	Pre-processing	Resampling technique	N/A	—	Details in Pati et al.
23	Pre-processing	Discretization method	N/A	—	Multiple bin configurations
24	Pre-processing	Image types and filter parameters	N/A	—	4 MRI sequences
25	Feature Extraction	Feature extraction software, IBSI	Yes	Section 2.1	IBSI-compliant, CaPTk

Item	Domain	Description	Compliance	Section	Notes
26	Feature Extraction	Feature classes using IBSI terminology	Yes	Table 3, S6	GLCM, GLSZM, Histogram, CoLIAGe
27	Feature Extraction	Total features per instance	Yes	Section 2.1	3920 per subcompartment
28	Feature Extraction	Default parameters stated	N/A	—	Pre-extracted features
29	Data Preparation	Missing data handling	Partial	Section 2.4	Near-zero-variance removal
30	Data Preparation	Class balance	N/A	—	Regression task
31	Data Preparation	Segmentation reliability	Partial	Section 2.1	Reproducibility in Pati et al.
32	Data Preparation	Feature normalization	Yes	Sections 2.4-2.5	Within-subcompartment z-scoring
33	Data Preparation	Dimension reduction	Yes	Section 2.4	Three-stage pipeline
34	Modeling	Software and algorithm details	Yes	Sections 2.5-2.6	lme4, glmnet with versions
35	Modeling	Training process and hyperparameters	Yes	Section 2.6	Alpha grid, inner 5-fold CV, lambda.1se
36	Modeling	Confounder detection	Yes	Section 2.8	Progressive covariate adjustment (S9)
37	Modeling	Final model selection	Yes	Section 2.6	Conservative lambda.1se rule
38	Evaluation	Internal or external testing	Yes	Sections 2.6, 2.12	Internal LOPO-CV only
39	Evaluation	Performance metrics	Yes	Section 2.6	R^2_{cv}, MAE, Spearman rho
40	Evaluation	Uncertainty quantification	Yes	Section 2.6	Bootstrap CIs, permutation p-values
41	Evaluation	Statistical software and comparison	Yes	Sections 2.5, 2.7	LRT, FDR, Holm corrections
42	Evaluation	Comparison with non-radiomic approaches	Yes	Section 2.8	Clinical covariate-only models
43	Evaluation	Interpretability/explainability	Partial	Section 3.3	Feature stability selection
44	Results	Baseline characteristics	Partial	Section 3.1	Sample counts reported
45	Results	Flowchart with inclusion/exclusion	Yes	Section 3.1	41 -> 31 -> 28 patients
46	Results	Feature statistics and selection	Yes	Tables 3, S6, S12	Stability, IBSI names, coefficients
47	Results	Performance metrics for train/val/test	Yes	Tables 2, 3	Nested CV R^2_{cv} with CIs
48	Results	Comparison with non-radiomic approaches	Yes	S7	Radiomic contribution above covariates
49	Discussion	Summary and study categorization	Yes	Discussion	Exploratory/hypothesis-generating
50	Discussion	Comparison with previous works	Yes	Discussion	Park, Le, Hu, Zhang compared
51	Discussion	Practical implications and future	Yes	Discussion	Speculative nature acknowledged
52	Discussion	Strengths and limitations	Yes	Limitations	10 limitations enumerated

Item	Domain	Description	Compliance	Section	Notes
53	Open Science	Raw/processed image data shared	N/A	Data Availability	Public datasets
54	Open Science	Radiomic feature data shared	Yes	Data Availability	Pre-extracted from TCIA
55	Open Science	Pre-processing/extraction scripts	N/A	—	Pre-extracted features
56	Open Science	Modeling scripts shared	Yes	Section 2.13	Code on GitHub
57	Open Science	Final model files shared	Partial	Section 2.13	Code shared, not serialized models
58	Open Science	Ready-to-use tool	No	—	Exploratory study

CLEAR summary: 40 Yes, 8 Partial, 8 N/A, 1 No (out of 57 items; the single “No” reflects the exploratory nature of the study).

Table S9. Top radiomic features for the Angiogenesis pathway, identified by stability selection across 28 LOPO folds. Features selected in >50% of folds are considered stable.

Feature ID	IBSI Feature Name	Folds selected	Stability (%)
rad_1763	T2 GLSZM Bins-16 Radius-2 GreyLevelNonUniformityNormalized	28	100.0
rad_1872	T2 GLSZM Bins-16 Radius-3 GreyLevelNonUniformityNormalized	28	100.0
rad_1920	T2 GLSZM Bins-64 Radius-3 ZoneSizeEntropy	28	100.0
rad_1363	T2 Histogram Bins-64 Kurtosis	27	96.4
rad_1956	T2 GLSZM Bins-128 Radius-3 ZoneSizeEntropy	27	96.4
rad_163	T1 Histogram Bins-16 Kurtosis	1	3.6
rad_2077	T1ce Intensity QuartileCoefficientOfVariation	1	3.6

Table S10. BH-FDR-corrected nested CV permutation p-values for all 24 pathways. Pathways with R^2_{cv} <= 0 were assigned p = 1.0 before FDR correction (conservative approach).

Pathway	R^2_{cv}	Permutation p (raw)	Permutation p (FDR)
Angiogenesis	0.209	0.006	0.096
Inflammatory Response	0.185	0.008	0.096
IvyGAP CTpan module	0.133	0.013	0.104
Hypoxia	-0.135	1.000	1.000
EMT	-0.034	1.000	1.000
TNFA/NF-kB	-0.113	1.000	1.000
IL6/JAK/STAT3	-0.199	1.000	1.000
IFN Gamma Response	-0.210	1.000	1.000
P53 Pathway	-0.075	1.000	1.000
MYC Targets V1	-0.075	1.000	1.000
E2F Targets	-0.075	1.000	1.000
G2M Checkpoint	-0.075	1.000	1.000
mTORC1 Signaling	-0.075	1.000	1.000
Glycolysis	-0.109	1.000	1.000
Oxidative Phosphorylation	-0.075	1.000	1.000
Complement	-0.004	1.000	1.000
Neftel MES	-0.146	1.000	1.000
Neftel AC	-0.075	1.000	1.000
Neftel OPC	-0.086	1.000	1.000
Neftel NPC	-0.075	1.000	1.000
IvyGAP CT module	-0.075	1.000	1.000
IvyGAP CTmvp module	-0.274	1.000	1.000
IvyGAP IT module	-0.072	1.000	1.000
IvyGAP LE module	-0.114	1.000	1.000

Table S11. Hyperparameter distributions (alpha, lambda) across LOPO folds for the three pathways with positive predictive signal. Alpha was selected from a grid of 0.1 to 1.0 (step 0.1); lambda was selected using the 1-SE rule (lambda.1se) from inner 5-fold CV.

Pathway	Folds with features	Alpha (min)	Alpha (max)	Alpha (median)	Lambda (min)	Lambda (max)	Lambda (median)
Angiogenesis	28	0.1	1.0	0.5	0.264	2.506	0.656
Inflammatory Response	28	0.1	1.0	0.4	0.270	1.637	0.663
IvyGAP CTpan module	28	0.1	1.0	0.6	0.175	1.698	0.455

Table S12. Top radiomic features for the Inflammatory Response pathway, identified by stability selection across 28 LOPO folds. Full IBSI feature names, fold selection counts, and stability percentages are provided. Features selected in >50% of folds are considered stable.

Feature ID	IBSI Feature Name	Folds selected	Stability (%)
rad_1707	T2 GLCM Bins-128 Radius-1 AutoCorrelation	28	100.0
rad_1950	T2 GLSZM Bins-128 Radius-3 LargeZoneLowGreyLevelEmphasis	28	100.0
rad_1930	T2 GLCM Bins-128 Radius-3 Energy	27	96.4
rad_1945	T2 GLSZM Bins-128 Radius-3 GreyLevelNonUniformityNormalized	20	71.4
rad_1363	T2 Histogram Bins-64 Kurtosis	16	57.1
rad_3909	FLAIR GLSZM Bins-128 Radius-3 LargeZoneHighGreyLevelEmphasis	4	14.3
rad_1000	T2 skewness-CoLIAGe correlation ws=3	2	7.1
rad_2134	T1ce Histogram Bins-16 QuartileCoefficientOfVariation	1	3.6
rad_2635	T1ce Histogram Bins-128 Bin-8 Probability	1	3.6

Table S13. Coefficient-level results for the Inflammatory Response linear mixed-effects model. Standardized beta coefficients, standard errors, 95% confidence intervals, Type II Satterthwaite ANOVA F-values, and Holm-adjusted p-values for all five radiomic features and the subcompartment fixed effect. Model: $R^2_m = 0.384$, $R^2_c = 0.687$, LRT $chi^2 = 20.53$, df = 5, p = 0.001.

Term	Beta	SE	95% CI	F	df	p	p (Holm)
rad_1732 (T2 GLSZM Bins-128 R1 LargeZoneLowGreyLevelEmphasis)	-0.471	0.141	[-0.758, -0.184]	11.16	33.0	0.0021	0.010 *
rad_1080 (T2 skewness-CoLIAGe diff-var ws=5)	0.308	0.149	[0.007, 0.610]	4.26	40.8	0.0454	0.181
rad_1251 (T2 Histogram Bins-64 Bin-1 Frequency)	-0.180	0.110	[-0.402, 0.042]	2.70	37.9	0.1088	0.326
rad_1052 (T2 skewness-CoLIAGe correlation ws=5)	0.130	0.163	[-0.199, 0.459]	0.64	40.3	0.4291	0.858
rad_1950 (T2 GLSZM Bins-128 R3 LargeZoneLowGreyLevelEmphasis)	0.003	0.159	[-0.321, 0.328]	0.00	32.9	0.9842	0.984
Subcompartment (NET vs ET)	1.945	0.093	[1.755, 2.135]	9.48[a]	17.5	0.0016	—
Subcompartment (ED vs ET)	-0.577	0.274	[-1.152, -0.001]	—	17.9	0.0496	—

* Holm-adjusted $p < 0.05$ among the five radiomic features.

[a] Joint F-test for the subcompartment factor (2 df). Individual subcompartment contrasts are reported as t-tests. The subcompartment effect accounts for R^2_m (null) = 0.170; the radiomic increment is ΔR^2_m = 0.214. All features are T2-derived. Coefficients are standardized (z-scored predictors).

7. OTHER CONTRIBUTIONS

Efficacy of VNS Stimulation in Drug-Resistant Epilepsy: An Analysis with Quantum and Artificial Intelligence Algorithms

Introduction

Vagal nerve stimulation (VNS) has for several years represented a safe and effective therapeutic option for seizure control in patients with drug-resistant epilepsy. However, the optimal criteria for patient selection toward the surgical pathway are not yet well defined. The aim of this preliminary study is to evaluate the efficacy of VNS, while attempting to identify predictive features of clinical response through the use of quantum and artificial intelligence tools.

Materials and Methods

A retrospective study was conducted on patients diagnosed with drug-resistant epilepsy according to ILAE3, who underwent VNS implantation between January 2008 and January 2022, with regular clinical follow-up longer than 24 months. Data were collected on 23 patients, including: age, sex, time since diagnosis, possible syndromic condition, seizure frequency and type, number of antiseizure medications used, genetic or structural etiology. Feature selection techniques based on quantum algorithms (QAOA, SB, SA) were employed to identify the most relevant characteristics. Predictive models were then built using machine learning methods such as Distributed Random Forest (DRF), Gradient Boosting Machines (GBM), and Deep Learning Machines, and their performance was evaluated through 5-fold cross-validation.

Results

Eight patients (34.8%) met the responder criterion, defined as a reduction in seizure frequency greater than 50%. Quantum feature selection techniques consistently highlighted the importance of age, time since diagnosis, and seizure frequency, with particular emphasis on focal seizures. Among predictive models, DRF showed the best performance with an accuracy of 78.26% and an AUC of 0.81.

Conclusions

This preliminary study suggests that age, time since diagnosis, and seizure frequency are key factors in predicting response to VNS in patients with drug-resistant epilepsy. The use of quantum techniques for feature selection combined with machine learning models represents a significant innovation in the field, offering new perspectives for the selection of suitable candidates for surgical therapy. These preliminary results warrant further investigation to refine therapeutic strategies and optimize surgical indications.

Characteristic	Overall, N = 23[1]	Non Responders, N = 15[1]	Responders, N = 8[1]	p-value[2]
Sex				0.7
Female	12 (52%)	7 (47%)	5 (63%)	
Male	11 (48%)	8 (53%)	3 (38%)	
Age	41.84 (11.35)	40.32 (10.07)	44.67 (13.73)	0.2
Lenght of disease	20.30 (16.41)	18.95 (14.67)	22.84 (20.12)	0.7
Syndrome				0.2
Lennox Gastaut Syndrome	3 (13%)	2 (13%)	1 (13%)	
Non syndromic epilepsy	16 (70%)	12 (80%)	4 (50%)	
Temporal Lobe Epilepsy	4 (17%)	1 (6.7%)	3 (38%)	
Etiology				>0.9
Genetic	3 (13%)	2 (13%)	1 (13%)	
Immune	1 (4.3%)	1 (6.7%)	0 (0%)	
Structural	14 (61%)	9 (60%)	5 (63%)	
Unknown	5 (22%)	3 (20%)	2 (25%)	
Total Focal Seizures	44.17 (43.68)	43.73 (45.48)	45.00 (43.12)	0.8
Total Generalized Seizures	23.70 (83.36)	35.53 (102.39)	1.50 (2.78)	>0.9
Total Seizures	67.87 (89.06)	79.27 (105.01)	46.50 (45.53)	0.3
Number of ASM	3.39 (0.89)	3.40 (0.83)	3.38 (1.06)	0.8

[1] n (%); Mean (SD)

[2] Fisher's exact test; Wilcoxon rank sum exact test; Wilcoxon rank sum test

Tumor Grade Unlocks Cortical Clues: Predicting Memory After Glioma Surgery

Objectives

To determine if pre-operative brain (cortical) thickness predicts memory function before and after glioma surgery, and memory changes. To investigate whether tumor grade (Low-Grade vs. High-Grade Glioma) alters how pre-operative cortical thickness relates to memory changes after surgery.

Background

Preserving memory is crucial in glioma surgery. While brain structure before surgery is thought to influence cognitive outcomes, its direct ability to predict memory, especially how this is affected by the tumor's biological aggressiveness (grade), is unclear. This study uses advanced imaging analysis to clarify these relationships.

Methods

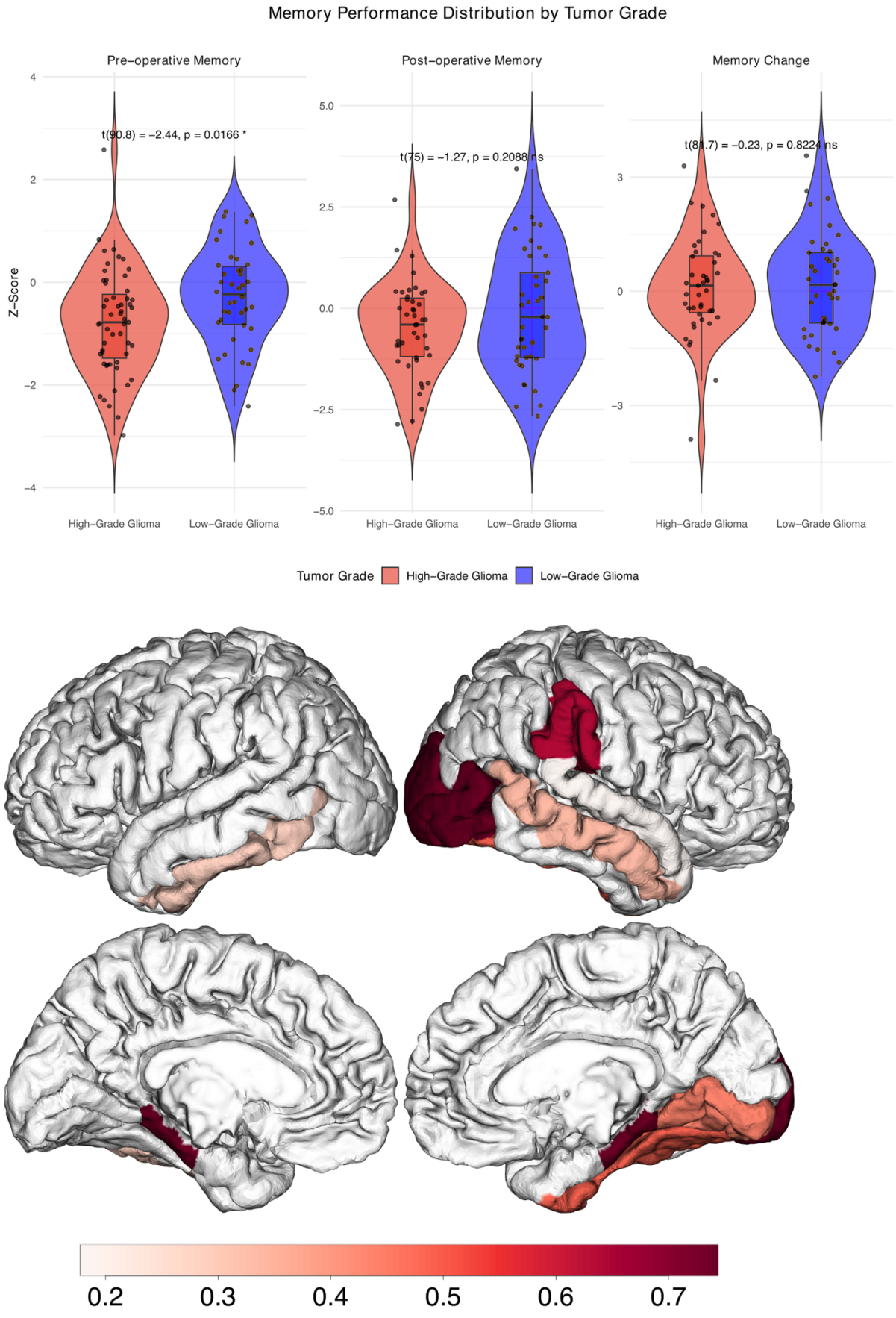
Pre-operative MRI-derived cortical thickness (regional/network-based) and memory scores (ZMPRE, ZMPOST, ZMDELTA) were analyzed in glioma patients. A rigorous iterative analytical approach, including univariate regressions, multivariate GLMs, and robust Elastic Net regularized regression, controlled for clinical covariates. Interaction models specifically tested the tumor grade's moderation of the thickness-memory change relationship.

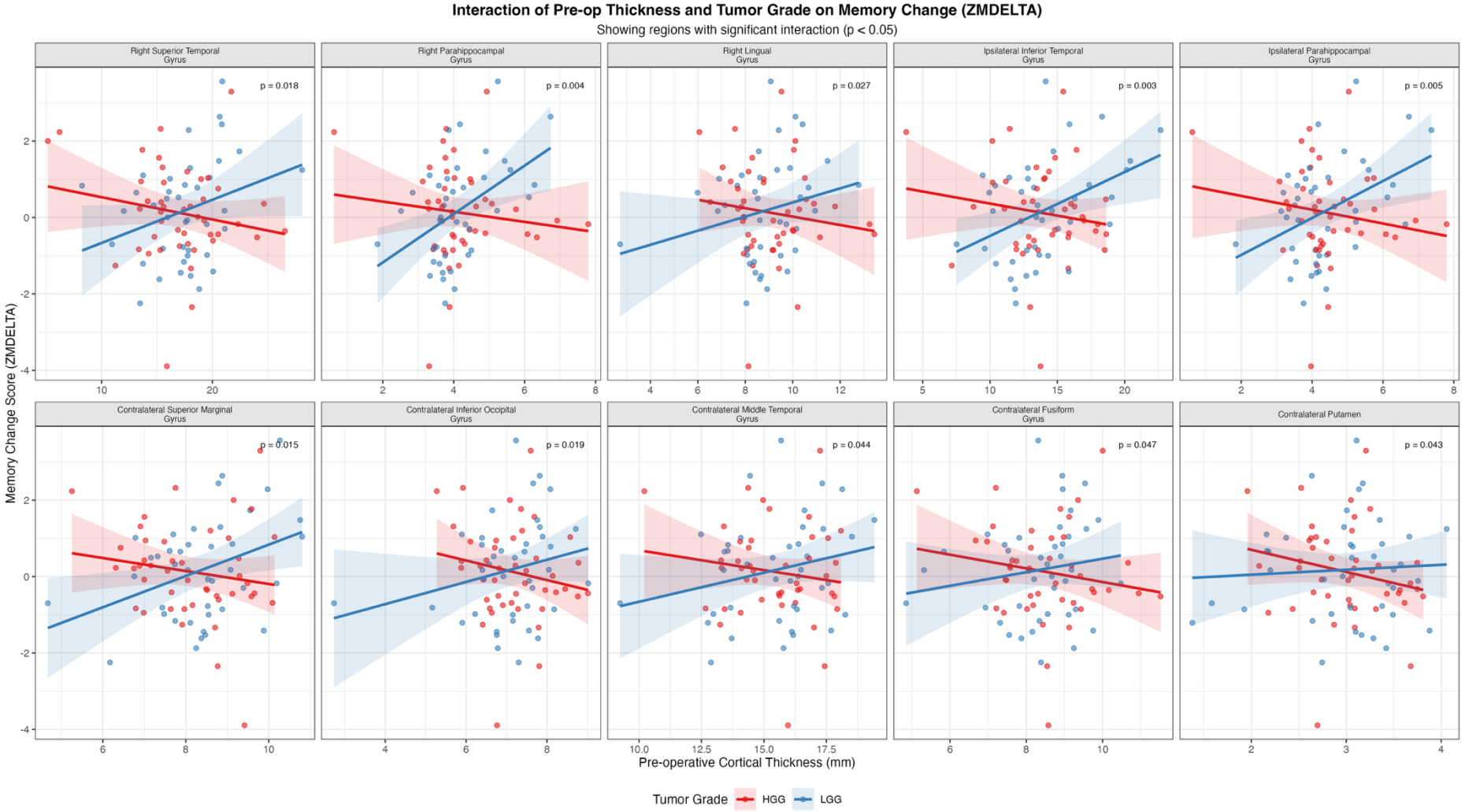
Results

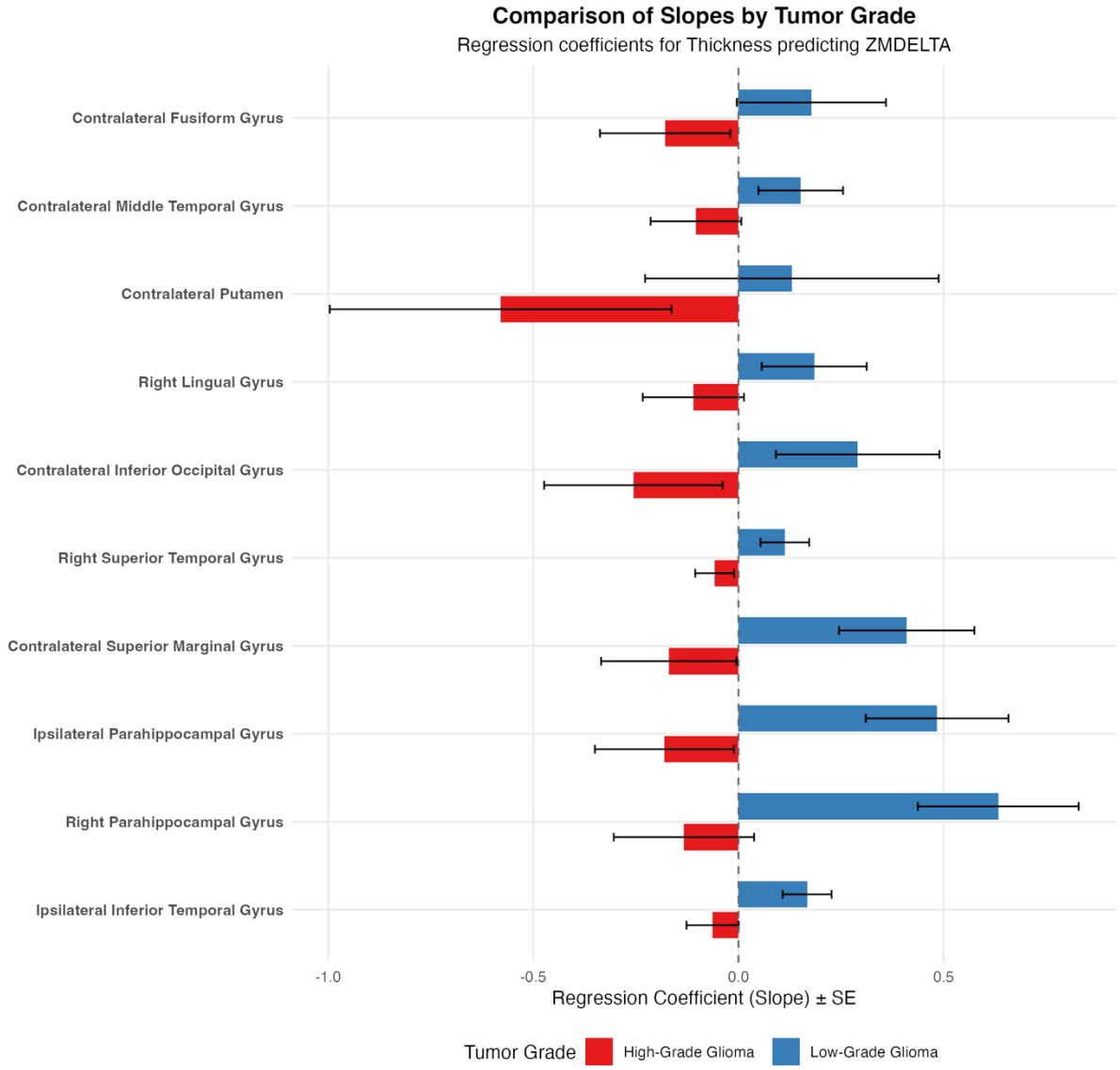
While pre-operative cortical thickness alone was not a strong, independent predictor of memory outcomes, we found a critical interaction with tumor grade. For memory change after surgery, the relationship with pre-operative thickness in ten specific brain regions – notably the right parahippocampal gyrus, ipsilateral inferior temporal gyrus, and contralateral middle temporal gyrus – was significantly different for LGG compared to HGG patients (p-interaction <0.05 to <0.004). This means that in these regions, whether a thicker cortex predicted better or worse memory change depended fundamentally on whether the tumor was low or high grade.

Conclusions

The predictive value of pre-operative cortical thickness for memory changes after glioma surgery is not straightforward; it is significantly shaped by tumor grade. This suggests that the brain's response and compensatory mechanisms differ based on tumor aggressiveness. Understanding this interplay between brain structure and tumor biology is vital for better predicting patient outcomes and could help tailor neurosurgical approaches.


Memory Performance Distribution by Tumor Grade
Pre−operative Memory
Post−operative Memory
Memory Change
Z−Score
t(90.8) = −2.44, p = 0.0166 *
t(75) = −1.27, p = 0.2088 ns
t(84.7) = −0.23, p = 0.8224 ns
High−Grade Glioma
Low−Grade Glioma
Tumor Grade
0.2
0.3
0.4
0.5
0.6
0.7


Interaction of Pre-op Thickness and Tumor Grade on Memory Change (ZMDELTA)
Showing regions with significant interaction (p < 0.05)
Right Superior Temporal Gyrus
p = 0.018
Right Parahippocampal Gyrus
p = 0.004
Right Lingual Gyrus
p = 0.027
Ipsilateral Inferior Temporal Gyrus
p = 0.003
Ipsilateral Parahippocampal Gyrus
p = 0.005
Contralateral Superior Marginal Gyrus
p = 0.015
Contralateral Inferior Occipital Gyrus
p = 0.019
Contralateral Middle Temporal Gyrus
p = 0.044
Contralateral Fusiform Gyrus
p = 0.047
Contralateral Putamen
p = 0.043
Memory Change Score (ZMDELTA)
Pre-operative Cortical Thickness (mm)
Tumor Grade
HGG
LGG


Comparison of Slopes by Tumor Grade
Regression coefficients for Thickness predicting ZMDELTA
Contralateral Fusiform Gyrus
Contralateral Middle Temporal Gyrus
Contralateral Putamen
Right Lingual Gyrus
Contralateral Inferior Occipital Gyrus
Right Superior Temporal Gyrus
Contralateral Superior Marginal Gyrus
Ipsilateral Parahippocampal Gyrus
Right Parahippocampal Gyrus
Ipsilateral Inferior Temporal Gyrus
-1.0
-0.5
0.0
0.5
Regression Coefficient (Slope) ± SE
Tumor Grade
High-Grade Glioma
Low-Grade Glioma

Prediction of Cerebellar Mutism Syndrome in Posterior fossa tumors: the role of the Rotterdam Score and multivariate analysis of associated factors

Objectives

Posterior fossa syndrome (PFS) is a rare but debilitating complication following neurosurgical treatment of infratentorial tumors in children. The Rotterdam model (RM) is a numerical index based on MRI features of the tumor and has been reported to predict a 66% risk of PFS in medulloblastoma (MB) patients with scores ≥100. However, its applicability to broader tumor types remains unverified.

Materials and Methods

We conducted a retrospective, single-center study from January 2011 to May 2025, enrolling 50 patients aged 0–16 years (mean 7.7 years) diagnosed with posterior fossa tumors and treated surgically. A multivariate analysis assessed correlations between sex, anthropometric characteristics, radiological features and surgical variables. Key predictors were analyzed in relation to PFS onset.

Results

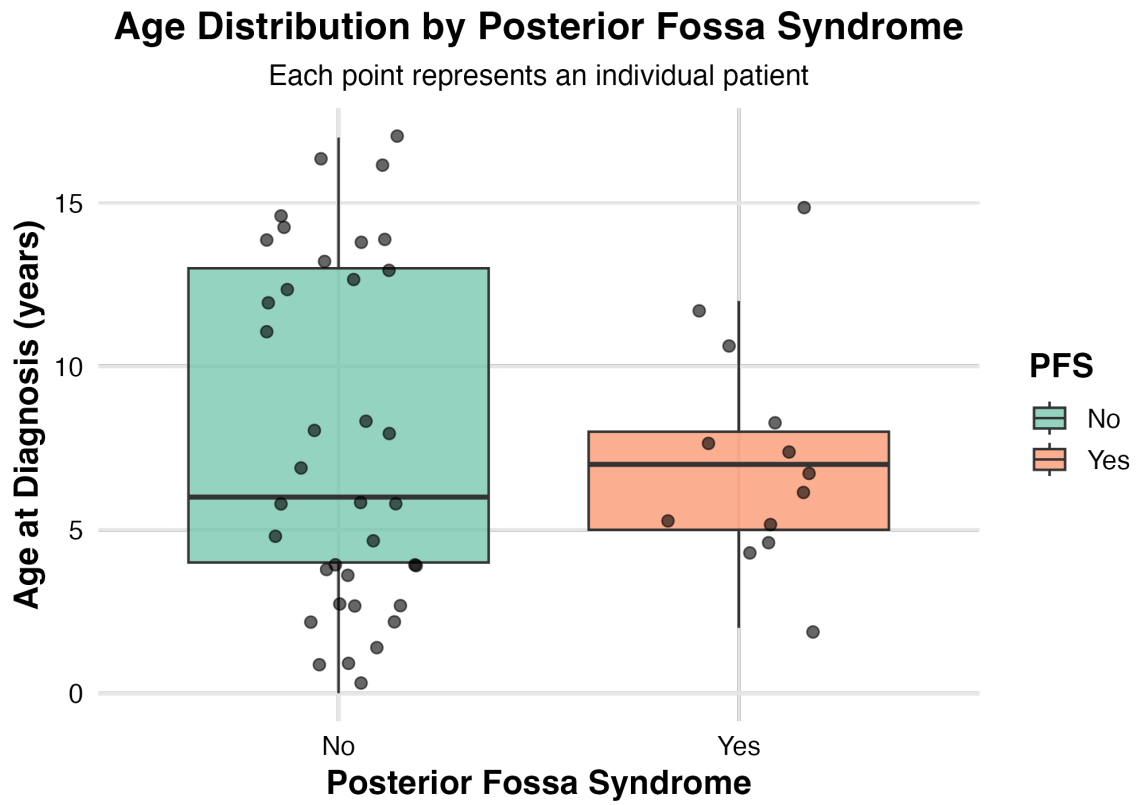
Children who developed PFS had significantly higher Rotterdam Index scores (mean 93.5 ± 34.2) compared to those without PFS (mean 51.2 ± 28.7), with a bimodal peak. However, applying the previously proposed cutoff of Rotterdam ≥100, only 28.6% of our cohort developed PFS, suggesting limited generalizability outside MB.

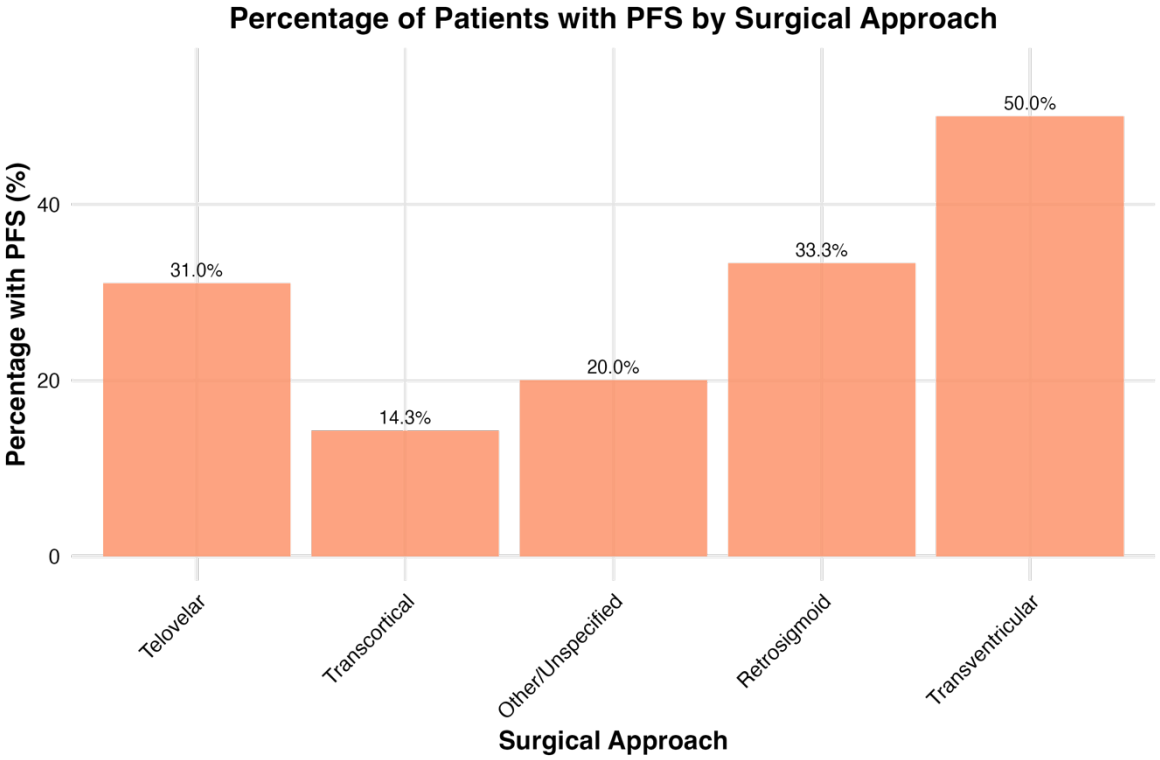
Similarly, body weight was lower in PFS patients (mean 21.6 ± 7.3 kg) than in unaffected patients (mean 34.8 ± 11.2 kg).

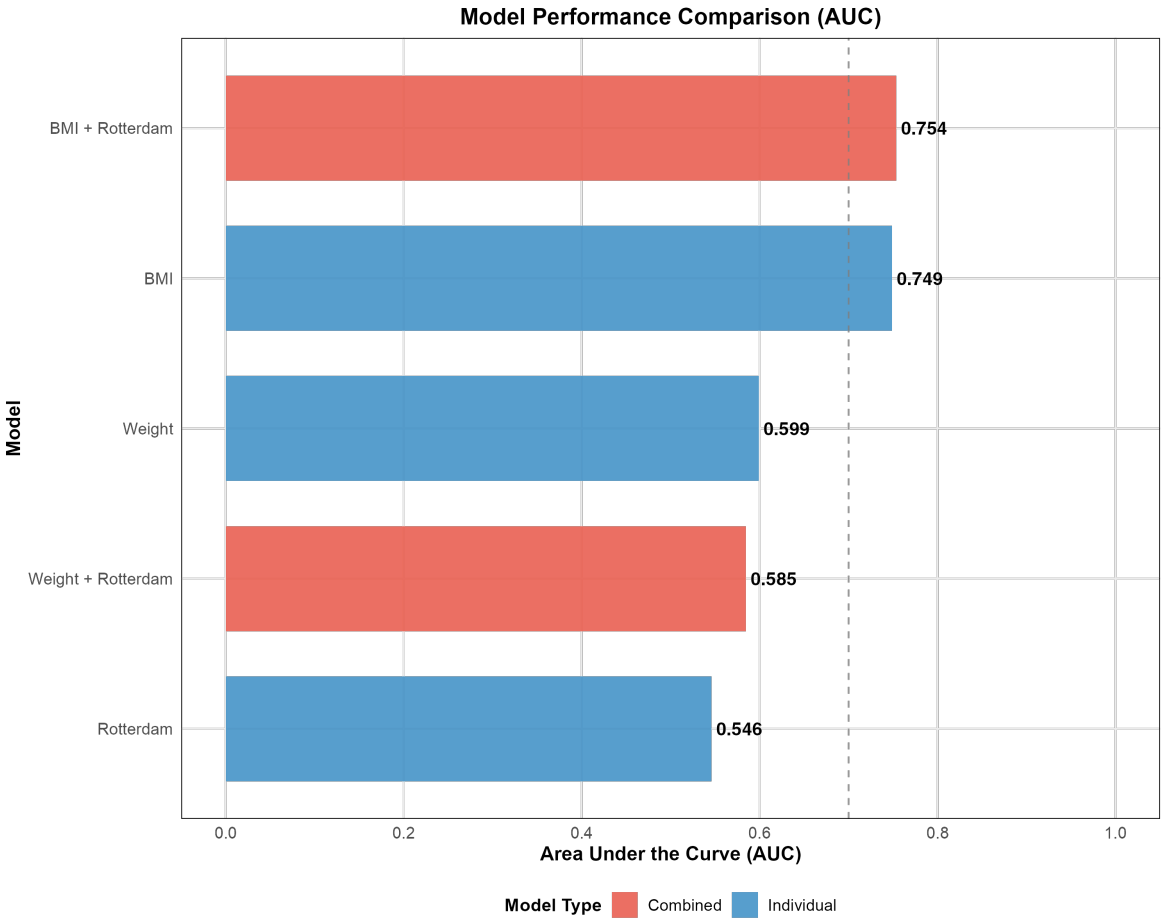
A significant interaction was identified between weight and Rotterdam Index: children with both a high Rotterdam score (>52.5) and low body weight (<17.9 kg) had the highest incidence of PFS (up to 100%), whereas no cases were observed in patients with weight >38.5 kg and Rotterdam Index <52.5. This suggests a strong protective effect of higher body mass.

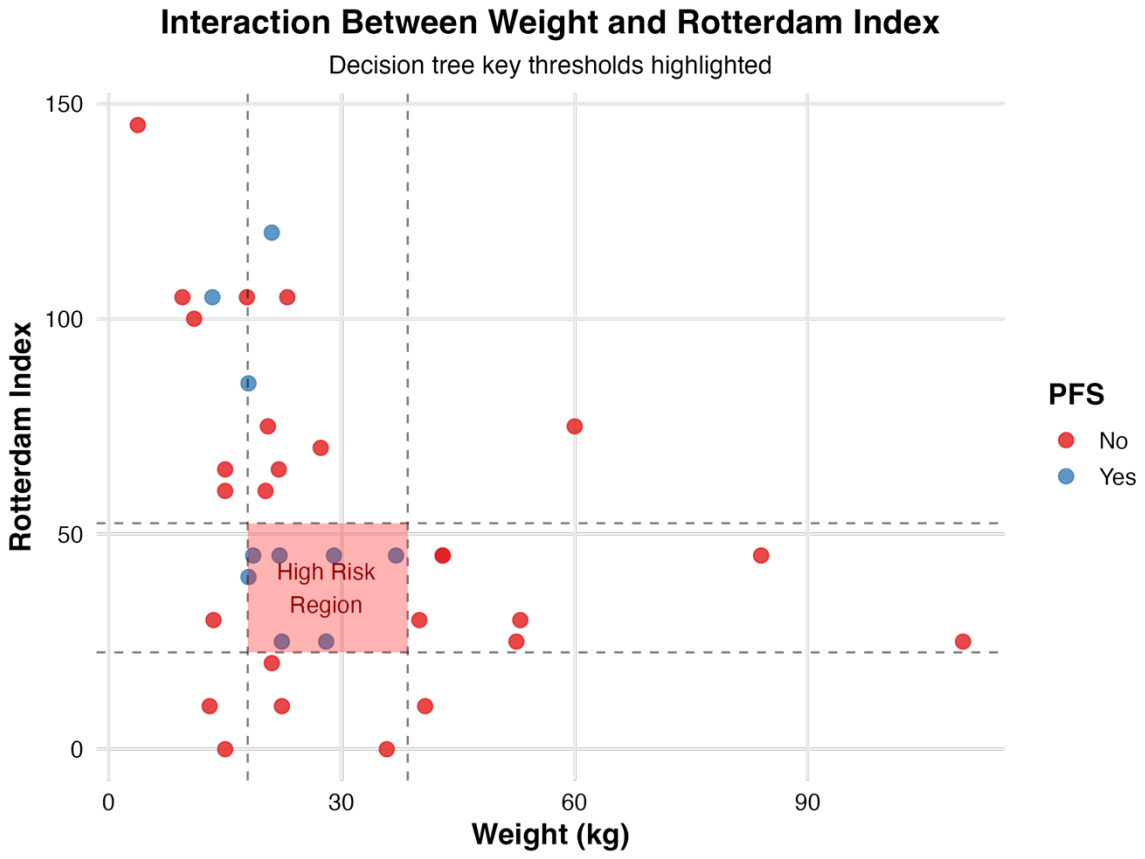
Conclusion

While the Rotterdam Index remains associated with PFS, its predictive power is limited when applied uniformly across tumor types. Body weight appears to be an independent and interacting protective factor, and the combined use of both parameters may allow more accurate preoperative risk stratification


Age Distribution by Posterior Fossa Syndrome
Each point represents an individual patient
Age at Diagnosis (years)
15
10
5
0
No
Yes
Posterior Fossa Syndrome
PFS
No
Yes


Percentage of Patients with PFS by Surgical Approach
Percentage with PFS (%)
40
20
0
31.0%
14.3%
20.0%
33.3%
50.0%
Telovelar
Transcortical
Other/Unspecified
Retrosigmoid
Transventricular
Surgical Approach


Model Performance Comparison (AUC)
BMI + Rotterdam
BMI
Weight
Weight + Rotterdam
Rotterdam
0.754
0.749
0.599
0.585
0.546
Model
0.0
0.2
0.4
0.6
0.8
1.0
Area Under the Curve (AUC)
Model Type
Combined
Individual


Interaction Between Weight and Rotterdam Index
Decision tree key thresholds highlighted
150
100
50
0
Rotterdam Index
High Risk
Region
0
30
60
90
Weight (kg)
PFS
No
Yes

8. General Discussion

This thesis set out to examine the integration of artificial intelligence into surgical practice across two complementary axes: the ethical and professional framework governing AI adoption in emergency surgery, and the development and evaluation of AI-based diagnostic and prognostic tools in neurosurgery. The preceding chapters have moved from normative principles (Chapter 2), through empirical assessment of surgeon readiness (Chapter 3), to a series of clinical applications spanning hydrocephalus diagnosis (Chapter 4), glioblastoma molecular profiling (Chapter 5), epilepsy neuromodulation (Chapter 6a), neuro-oncological cognition (Chapter 6b), and posterior fossa surgery (Chapter 6c). The present chapter synthesizes these contributions, identifies the connections and tensions among them, and situates the collective findings within the broader landscape of translational surgical AI.

The Ethical and Professional Landscape

The Delphi consensus study (Chapter 2) distilled the seven requirements for trustworthy AI articulated by the European Commission's High-Level Expert Group into the specific context of surgical practice. Through structured deliberation among twelve experts drawn from surgery, computer science, ethics, and healthcare management, the study identified the principal ethical tensions that arise when algorithmic systems enter the operating theatre and the clinical ward: the allocation of liability between surgeon and algorithm, the opacity of predictive models that cannot explain their reasoning in clinically interpretable terms, the risk that biased training data will perpetuate or amplify existing healthcare disparities, and the challenge of preserving meaningful human oversight when AI systems operate at speeds and scales that exceed human cognitive capacity. These findings established a normative framework, grounded in human agency, technical robustness, privacy, transparency, fairness, societal well-being, and accountability, against which any proposed surgical AI application should be evaluated.

The international survey of 650 emergency surgeons across 71 countries (Chapter 3) then tested the empirical preconditions for that framework's implementation. The results exposed a fundamental disconnect. While 69% of respondents reported familiarity with AI and machine learning terminology, only 17% provided a definition that met the concordance criterion upon qualitative analysis; 55% of definitions were rated as discordant, reflecting either substantive misunderstanding or absence of knowledge. This pattern carries direct consequences for the ethical principles established in Chapter 2. Transparency and explainability presuppose a recipient capable of evaluating algorithmic outputs; informed human oversight requires a human who is, in fact, informed. When the majority of surveyed surgeons cannot accurately characterize the technology they would be asked to oversee, the governance architecture envisioned by the Delphi consensus rests on an unstable foundation.

The survey data further revealed that emergency surgeons do not regard AI as either a priority resource or a pressing professional threat. Machine learning and artificial intelligence ranked 10th of 11 clinical decision-making facilitators (mean 3.56, SD 1.07) and last among 13 perceived challenges to clinical judgment (mean 3.10, SD 1.14). Both items recorded the

highest standard deviations in their respective scales, indicating a polarized community rather than uniform indifference. This polarization is consistent with early-phase technology diffusion, in which attitudes distribute bimodally between enthusiasts and skeptics before convergence toward adoption or rejection.

The temporal dimension of the survey data introduces a more nuanced interpretation. Surgeons rated the current importance of AI at 3.06 but projected its five-year importance at 3.88, a statistically significant increase that signals anticipatory awareness without present engagement. This gap between acknowledged trajectory and current preparation constitutes a window of opportunity: the profession expects AI to arrive but has not yet invested in the knowledge infrastructure necessary to receive it. Geographic variation reinforced this point. Respondents from Argentina (mean 4.50), Brazil (4.38), and Saudi Arabia (4.43) expressed substantially greater enthusiasm than those from Switzerland (2.60) and Canada (2.60), suggesting that institutional context, resource availability, and local technological culture modulate readiness in ways that preclude a uniform global implementation strategy.

Clinical AI Applications: From Diagnosis to Prognosis

The clinical studies presented in Chapters 4, 5, and 6, instantiate the translational challenges identified by the ethical and survey analyses, demonstrating both the promise and the constraints of AI in neurosurgical practice.

The Super Learner ensemble for idiopathic normal pressure hydrocephalus diagnosis (Chapter 4) represents the most methodologically mature application in this thesis. The study enrolled 100 patients, employed a standardized deep-learning neuroimaging pipeline (CAT12) for cortical thickness extraction across the Desikan-Killiany-Tourville atlas, and combined multiple machine learning algorithms (Distributed Random Forest, Extremely Randomized Trees, Generalized Linear Model with Regularization, Gradient Boosting Machine, Extreme Gradient Boosting, and a fully connected neural network) into a stacked ensemble. The resulting Super Learner achieved an AUC of 0.843 on the held-out test set, with a positive predictive value of 90% when integrating cortical thickness with CSF dynamic parameters. SHAP analysis identified R_{out} as the dominant predictor and highlighted contributions from frontal and parietal cortical regions, particularly the caudal middle frontal (Cohen's D = 0.86, corrected p = 0.002), superior frontal (D = 0.82, p = 0.002), rostral middle frontal (D = 0.62, p = 0.038), and superior parietal (D = 0.80, p = 0.006) areas. These cortical thickness patterns provide explainable biomarkers, a direct response to the transparency requirement articulated in Chapter 2, and align with prior research linking frontal and parietal thinning to iNPH pathophysiology and its differential diagnosis from neurodegenerative conditions.

The RADIOMAP-IvyGAP analysis (Chapter 5) tested whether MRI-derived radiomic features from tumor subcompartments associate with transcriptomic pathway enrichment in glioblastoma. Of 24 pathways evaluated, only Angiogenesis (R^2_{cv} = 0.209, permutation p = 0.006) and Inflammatory Response (R^2_{cv} = 0.185, p = 0.008) showed genuine radiomic-transcriptomic associations, both reaching FDR = 0.096. The remaining 21 of 24 pathways exhibited no predictive signal. This dominant negative finding is itself a substantive scientific contribution: it constrains the assumption that radiomic features can serve as non-invasive proxies for arbitrary molecular programmes and argues for pathway-selective rather than

pan-molecular claims in the habitat imaging literature. The consistency of T2-derived texture features across both positive pathways is biologically plausible, given T2 sensitivity to tissue water content, inflammatory edema, and angiogenic vascular changes.

The VNS response prediction study (Chapter 6a) explored a different methodological frontier. In a cohort of 31 patients with drug-resistant epilepsy, a quantum-enhanced hybrid approach to feature selection combined quantum-inspired algorithms (QAOA, Simulated Bifurcation, Simulated Annealing) via the Falcondale SDK with clinical domain constraints. The approach identified total seizure frequency (Random Forest importance 33.2%), age at implantation (30.5%), and time since diagnosis (28.9%) as primary predictors of VNS response, with age and time since diagnosis together accounting for 59% of predictive importance. The Gradient Boosting Machine achieved the best cross-validated performance at 77.1% +/- 8.6% accuracy (AUC = 0.70). The gap between this cross-validated accuracy and the model's perfect training-set classification (100%) exemplifies the overfitting risk inherent in gradient boosting on small neurosurgical datasets and underscores the need for external validation. The quantum contribution should be understood as a proof-of-concept for emerging computational paradigms in surgical prediction rather than a mature clinical tool: direct quantum optimization was constrained by the small sample size, and clinical domain knowledge guided the inclusion of key demographic predictors that were not independently discovered by all quantum algorithms.

The tumor grade-memory interaction analysis (Chapter 6b) introduced a layer of biological complexity that challenges simple biomarker models. The finding that tumor grade moderates the relationship between pre-operative cortical thickness and post-operative memory change, with the direction of this relationship reversing between low-grade and high-grade glioma patients in regions including the right parahippocampal gyrus ($p = 0.004$), ipsilateral inferior temporal gyrus ($p = 0.003$), and contralateral middle temporal gyrus ($p = 0.044$), suggests that cognitive reserve mechanisms interact with tumor biology in non-linear ways. A thicker cortex that predicts better memory outcomes in low-grade glioma may carry different prognostic significance in high-grade disease, where infiltrative growth patterns and peritumoral edema alter the structure-function relationship. This complexity is directly relevant to AI model design: predictive algorithms that treat cortical thickness as a uniform biomarker without accounting for tumor grade will systematically misestimate cognitive outcomes.

The cerebellar mutism prediction study (Chapter 6c) yielded a clinically actionable finding. Among 50 pediatric patients with posterior fossa tumors, body mass index emerged as the sole significant independent predictor of posterior fossa syndrome ($p = 0.031$, AUC = 0.749), while the Rotterdam Index, previously validated in medulloblastoma, showed limited predictive power when applied uniformly across tumor types (AUC = 0.546). The interaction between weight and Rotterdam score was notable: children with both high Rotterdam scores (>52.5) and low body weight (<17.9 kg) exhibited the highest incidence of PFS, whereas no cases occurred in patients with weight above 38.5 kg and Rotterdam Index below 52.5. That BMI is a modifiable pre-operative parameter distinguishes this finding from purely prognostic biomarkers and opens a potential pathway to nutritional optimization before posterior fossa surgery.

Bridging the Translational Gap

The synthesis of ethical framework and clinical applications reveals a set of interdependencies that no single chapter could articulate in isolation.

The explainability requirement established through Delphi consensus (Chapter 2) finds variable compliance across the clinical studies. The Super Learner for iNPH (Chapter 4) offers the strongest example: SHAP analysis renders individual predictions interpretable, cortical thickness measurements map onto anatomically defined brain regions, and the ensemble's reliance on CSF dynamics connects to established pathophysiological reasoning. By contrast, the VNS quantum-enhanced feature selection (Chapter 6a) operates through algorithmic processes that are less transparent to the clinician, and the radiomic texture features identified in the RADIOMAP-IvyGAP analysis (Chapter 5) lack intuitive clinical interpretation despite their statistical validity. This gradient of interpretability across the thesis studies illustrates that explainability is not a binary property but a continuum, and that different clinical applications will occupy different positions along it.

The survey finding that 55% of surgeons provided discordant definitions of AI (Chapter 3) has direct implications for the deployment of every clinical model presented in this thesis. A surgeon who cannot define artificial intelligence cannot meaningfully evaluate a Super Learner ensemble's recommendation for shunt surgery, assess the credibility of quantum-enhanced feature importance rankings, or interpret a radiomic-transcriptomic association in the context of treatment planning. The governance framework of Chapter 2 requires technically literate human oversight; the readiness data of Chapter 3 demonstrate that this literacy is largely absent. The thesis therefore reveals a recursive gap: ethical AI governance requires informed clinicians, but clinicians lack foundational AI knowledge, and without that knowledge they cannot meaningfully participate in the ethical oversight that responsible deployment demands.

Data governance considerations (Chapter 2) are equally relevant to the clinical studies. The sample sizes involved — iNPH (N = 100), VNS (N = 31), cerebellar mutism (N = 50), RADIOMAP (N = 28) — reflect the reality of neurosurgical research, where rare conditions, specialized procedures, and single-center recruitment constrain cohort assembly. Each study has demonstrated internal validity through appropriate cross-validation strategies (10-fold, 5-fold stratified, leave-one-patient-out), but none has undergone external validation on an independent cohort. Multicenter collaboration represents the critical next step for every predictive model in this thesis, and such collaboration will itself require the data-sharing governance structures that Chapter 2 identified as ethically necessary.

The bias detection principle (Chapter 2) relates to a more subtle observation. The geographic variation in AI perception documented in Chapter 3, from Argentina's enthusiasm (4.50) to Switzerland's skepticism (2.60), mirrors a potential concern in clinical studies: all neurosurgical cohorts were gathered from a single Italian center. Algorithms trained on uniform populations risk inheriting demographic and institutional biases that limit their applicability. The Delphi panel's focus on diversity and non-discrimination in AI development therefore goes beyond abstract principles to a concrete methodological requirement for the clinical research program.

Limitations

Several limitations qualify the conclusions drawn from this body of work.

The Delphi consensus (Chapter 2) engaged twelve experts, a panel sufficient for the modified EFTE methodology employed but potentially unrepresentative of the broader surgical and technological community. The geographic and disciplinary composition of the panel may have shaped the consensus toward perspectives prevalent in European academic surgery.

The WSES survey (Chapter 3) achieved a response rate of approximately 70%, but respondents were WSES members, a population likely more research-oriented and internationally connected than the global surgical workforce. Self-reported familiarity introduces social desirability bias, possibly inflating the 69% familiarity rate. The survey assessed knowledge at a single time point and cannot capture the trajectory of AI literacy within the profession.

The Super Learner study (Chapter 4) was retrospective and single-centre, with cortical thickness derived from CAT12, a computational pipeline whose measurements may not be directly comparable across different software platforms. No external validation cohort was available.

The RADIOMAP-IvyGAP analysis (Chapter 5) was exploratory in design, with no spatial co-registration between MRI voxels and laser microdissection sampling sites. The zone-to-subcompartment mapping is biologically approximate (mean correlation $r = 0.242$), and the sample size ($N = 28$) falls substantially below the minimum required for reliable prediction.

The VNS study (Chapter 6a) was limited by its small sample size ($N = 31$). The gap between 77.1% cross-validated accuracy and 100% training-set accuracy indicates overfitting. Clinical constraints guided feature selection, and per-fold cross-validated sensitivity and specificity were not available from existing outputs.

The tumor grade-memory analysis (Chapter 6b) was exploratory, with multiple interaction tests across brain regions raising concerns about Type I error despite the biological plausibility of the findings.

The cerebellar mutism study (Chapter 6c) enrolled 50 patients from a single paediatric centre, with complete data available for 32. BMI was assessed as a static pre-operative variable; dynamic nutritional status may carry different predictive value.

No predictive model in this thesis has undergone prospective validation. All neurosurgical studies were conducted at a single centre.

Future Directions

The findings of this thesis point toward several concrete research priorities. Prospective multicenter validation studies are needed for the Super Learner iNPH diagnostic model and for the BMI-based prediction of cerebellar mutism — the two applications with the most direct clinical utility. Integration of AI decision support tools into emergency surgery workflows, building on the readiness baseline established in Chapter 3, would permit longitudinal assessment of whether validated tools shift surgeon attitudes and knowledge over time. Spatial co-registration studies linking MRI voxels to tissue-level transcriptomic data would address the principal limitation of the RADIOMAP-IvyGAP analysis and test whether the Angiogenesis and Inflammatory Response associations identified in this thesis survive spatially precise validation. Educational programmes targeting foundational AI literacy among surgeons — the prerequisite identified by the convergence of Chapters 2 and 3 — should be developed and evaluated through professional societies such as WSES, with pre- and post-intervention assessment of both knowledge and attitudes. Larger, multicentre epilepsy cohorts would enable true quantum optimization without the clinical constraints that the small VNS sample necessitated. Longitudinal studies tracking how surgeon perceptions of AI evolve with exposure to validated, well-explained clinical tools would close the feedback loop between ethical governance, professional readiness, and technical deployment that this thesis has mapped.

9. CONCLUSIONS

This thesis has demonstrated that the responsible integration of artificial intelligence into surgical practice demands simultaneous advancement on ethical, professional, and technical fronts. The Delphi consensus study (Chapter 2) established a governance framework grounded in the seven requirements for trustworthy AI — human agency, technical robustness, privacy, transparency, fairness, societal well-being, and accountability — and identified the specific ethical tensions that arise when algorithmic systems enter surgical decision-making, including liability allocation, model opacity, and the preservation of meaningful human oversight. The international survey of 650 emergency surgeons across 71 countries (Chapter 3) revealed that while 69% of respondents reported familiarity with AI terminology, only 17% demonstrated concordant understanding, exposing a pattern of familiarity without comprehension that undermines the informed oversight upon which ethical governance depends. The Super Learner ensemble for iNPH diagnosis (Chapter 4) achieved an AUC of 0.843 and a positive predictive value of 90%, providing an explainable, clinically actionable diagnostic tool grounded in cortical thickness biomarkers. The RADIOMAP-IvyGAP analysis (Chapter 5) provided the first evidence linking radiomic features to transcriptomic pathway enrichment in glioblastoma, while constraining this association to two of 24 pathways — Angiogenesis and Inflammatory Response — and reporting that 21 pathways showed no signal. The VNS response prediction study (Chapter 6a) applied a quantum-enhanced hybrid feature selection approach to a 31-patient epilepsy cohort, identifying age, seizure frequency, and time since diagnosis as primary predictors (GBM cross-validated accuracy 77.1%), while honestly acknowledging the overfitting constraints of small-sample gradient boosting. The tumor grade-memory interaction analysis (Chapter 6b) demonstrated that the relationship between cortical thickness and post-operative memory is non-linear and grade-dependent. The cerebellar mutism study (Chapter 6c) identified body mass index as a modifiable pre-operative predictor of posterior fossa syndrome (AUC = 0.749).

The dual-axis structure of this thesis — ethical and professional framework in emergency surgery, clinical AI applications in neurosurgery — reflects an interdependence rather than a division. The ethical principles of transparency and explainability articulated through expert consensus acquire operational meaning only when tested against concrete clinical models: the SHAP-interpretable Super Learner satisfies these principles more fully than the opaque radiomic texture features, illustrating that explainability varies across applications rather than adhering to a single standard. Conversely, even high-performing predictive models cannot achieve clinical impact without a workforce prepared to evaluate, adopt, and oversee them, a precondition that the survey data demonstrate is not yet met. The recursive relationship between ethical governance, professional readiness, and technical development constitutes the central finding of this thesis: none of these three dimensions can advance in isolation, and progress on any one axis is constrained by deficits on the others.

Three priorities define the path forward. First, prospective multicenter validation of the Super Learner for iNPH diagnosis and the BMI-based cerebellar mutism predictor would establish whether the performance metrics observed in single-centre retrospective studies translate to broader clinical populations. Second, structured educational programmes on AI fundamentals, delivered through professional societies and integrated into surgical training curricula, must address the knowledge deficit that currently prevents the surgical community from exercising the informed oversight that ethical governance requires. Third, prospective clinical trials embedding AI decision support tools into emergency and neurosurgical workflows would generate the longitudinal evidence needed to assess whether validated algorithms improve patient outcomes, alter clinician behaviour, and justify the institutional investment that deployment demands.

The path from algorithmic proof-of-concept to responsible clinical integration is neither short nor guaranteed, but this thesis has mapped its essential coordinates: the ethical principles that must guide the journey, the professional readiness that must sustain it, and the technical evidence that must justify each step.

10. BIBLIOGRAPHY

1. Noh SH, Cho PG, Kim KN, Kim SH, Shin DA. Artificial Intelligence for Neurosurgery : Current State and Future Directions. *J Korean Neurosurg Soc*. 2023;66(2):113-120.

2. Levy AS, Bhatia S, Merenzon MA, et al. Exploring the Landscape of Machine Learning Applications in Neurosurgery: A Bibliometric Analysis and Narrative Review of Trends and Future Directions. *World Neurosurg*. 2024;181:108-115.

3. Tangsrivimol JA, Schonfeld E, Zhang M, et al. Artificial Intelligence in Neurosurgery: A State-of-the-Art Review from Past to Future. *Diagnostics*. 2023;13(14):2429.

4. Khanna R, Raison N, Granados A, et al. Quantum computing in surgery and urology — taking a quantum leap. *Nat Rev Urol*. Published online June 16, 2025:1-2.

5. Lee B, Liu CY, Apuzzo MLJ. Quantum computing: a prime modality in neurosurgery's future. *World Neurosurg*. 2012;78(5):404-408.

6. Schilling AT, Shah PP, Feghali J, Jimenez AE, Azad TD. A Brief History of Machine Learning in Neurosurgery. *Acta Neurochir Suppl*. 2022;134:245-250.

7. Kwoh YS, Hou J, Jonckheere EA, Hayati S. A robot with improved absolute positioning accuracy for CT guided stereotactic brain surgery. *IEEE Trans Biomed Eng*. 1988;35(2):153-160.

8. Benabid AL, Cinquin P, Lavalle S, Le Bas JF, Demongeot J, de Rougemont J. Computer-driven robot for stereotactic surgery connected to CT scan and magnetic resonance imaging. Technological design and preliminary results. *Appl Neurophysiol*. 1987;50(1-6):153-154.

9. Nathoo N, Cavuşoğlu MC, Vogelbaum MA, Barnett GH. In touch with robotics: neurosurgery for the future. *Neurosurgery*. 2005;56(3):421-433; discussion 421-433.

10. Grigsby J, Kramer RE, Schneiders JL, Gates JR, Brewster Smith W. Predicting outcome of anterior temporal lobectomy using simulated neural networks. *Epilepsia*. 1998;39(1):61-66.

11. Songsaeng D, Nava-Apisak P, Wongsripuemtet J, et al. The Diagnostic Accuracy of Artificial Intelligence in Radiological Markers of Normal-Pressure Hydrocephalus (NPH) on Non-Contrast CT Scans of the Brain. *Diagn Basel Switz*. 2023;13(17):2840.

12. Lee J, Kim D, Suh CH, et al. Automated Idiopathic Normal-Pressure Hydrocephalus Diagnosis via Artificial Intelligence-Based 3D T1 MRI Volumetric Analysis. *Am J Neuroradiol*. Published online September 9, 2024.

13. Pahwa B, Tayal A, Shukla A, et al. Utility of Machine Learning in the Management of Normal Pressure Hydrocephalus: A Systematic Review. *World Neurosurg*. 2023;177:e480-e492.

14. Twinanda AP, Shehata S, Mutter D, Marescaux J, de Mathelin M, Padoy N. EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos. *IEEE Trans Med Imaging*. 2017;36(1):86-97.

15. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv*. Preprint posted online May 18, 2015:arXiv:1505.04597.

16. Menze BH, Jakab A, Bauer S, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans Med Imaging*. 2015;34(10):1993-2024.

17. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol*. 2017;14(12):749-762.

18. Oh S, Choi J, Kim J. A Tutorial on Quantum Convolutional Neural Networks (QCNN). *arXiv*. Preprint posted online September 20, 2020:arXiv:2009.09423.

19. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. *Lancet Digit Health*. 2021;3(3):e195-e203.

20. Vokinger KN, Feuerriegel S, Kesselheim AS. Continual learning in medical devices: FDA's action plan and beyond. *Lancet Digit Health*. 2021;3(6):e337-e338.

21. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion*. 2020;58:82-115.

22. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447-453.

23. Borgesius FZ. Discrimination, artificial intelligence, and algorithmic decision-making. *arXiv*. Preprint posted online October 15, 2025:arXiv:2510.13465.

24. Stiegler MP, Neelankavil JP, Canales C, Dhillon A. Cognitive errors detected in anaesthesiology: a literature review and pilot study. *Br J Anaesth*. 2012;108(2):229-235.

25. Cobianchi L, Piccolo D, Dal Mas F, et al. Surgeons' perspectives on artificial intelligence to support clinical decision-making in trauma and emergency contexts: results from an international survey. *World J Emerg Surg*. 2023;18(1):1.

26. Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial Intelligence in Surgery: Promises and Perils. *Ann Surg*. 2018;268(1):70-76.

27. Liu NT, Salinas J. Machine Learning for Predicting Outcomes in Trauma. *Shock*. 2017;48(5):504-510.

28. Kuo PJ, Wu SC, Chien PC, et al. Derivation and validation of different machine-learning models in mortality prediction of trauma in motorcycle riders: a cross-sectional retrospective study in southern Taiwan. *BMJ Open*. 2018;8(1):e018252.

29. Byerly S, Maurer LR, Mantero A, Naar L, An G, Kaafarani HMA. Machine Learning and Artificial Intelligence for Surgical Decision Making. *Surg Infect*. 2021;22(6):626-634.

30. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44-56.

31. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. *Lancet Digit Health*. 2021;3(3):e195-e203.

32. High-Level Expert Group on Artificial Intelligence. *Ethics Guidelines For Trustworthy AI*. 2019.

33. Nelms KR, Porter AL. EFTE: An interactive Delphi method. *Technol Forecast Soc Change*. 1985;28(1):43-61.

34. Bagnoli C, Dal Mas F, Biancuzzi H, Massaro M. Business Models Beyond Covid-19. A Paradoxes Approach. *J Bus Models*. 2021;(Online first).

35. Odland ML, Nepogodiev D, Morton D, et al. Identifying a Basket of Surgical Procedures to Standardize Global Surgical Metrics. *Ann Surg*. 2020;Publish Ah.

36. Asbun HJ, Abu Hilal M, Kunzler F, et al. International Delphi Expert Consensus on Safe Return to Surgical and Endoscopic Practice: From the Coronavirus Global Surgical Collaborative. *Ann Surg*. 2021;274(1):50-56.

37. D'Souza N, de Neree tot Babberich MPM, d'Hoore A, et al. Definition of the rectum: An International, expert-based Delphi consensus. *Ann Surg*. 2019;270(6):955-959.

38. Cobianchi L, Dal Mas F, Angelos P. One size does not fit all – Translating knowledge to bridge the gaps to diversity and inclusion of surgical teams. *Ann Surg*. 2021;273(2):e34-e36.

39. Loftus TJ, Tighe PJ, Filiberto AC, et al. Artificial Intelligence and Surgical Decision-making. *JAMA Surg*. 2020;155(2):148-158.

40. O'Sullivan S, Nevejans N, Allen C, et al. Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery. *Int J Med Robot*. 2019;15(1):1-12.

41. Yang GZ, Cambias J, Cleary K, et al. Medical robotics. Regulatory, ethical, and legal considerations for increasing levels of autonomy. *Sci Robot*. 2017;2(4).

42. Briggs A, Raja AS, Joyce MF, et al. The role of nontechnical skills in simulated trauma resuscitation. *J Surg Educ*. 2015;72(4):732-739.

43. Georgiou A, Lockey DJ. The performance and assessment of hospital trauma teams. *Scand J Trauma Resusc Emerg Med*. 2010;18(1):1-7.

44. Yule S, Smink DS. Non-Technical Skill Countermeasures for Pandemic Response. *Ann Surg*. 2020;272(3):e213-e215.

45. Cobianchi L, Dal Mas F, Peloso A, et al. Planning the Full Recovery Phase: An Antifragile Perspective on Surgery after COVID-19. *Ann Surg*. 2020;272(6):e296-e299.

46. Thistlethwaite PA. The sparkle of creativity. *J Thorac Cardiovasc Surg*. 2020;160(3):740-752.

47. Gauderer MWL. Creativity and the surgeon. *J Pediatr Surg*. 2009;44(1):13-20.

48. Steil J, Finas D, Beck S, Manzeschke A, Haux R. Robotic Systems in Operating Theaters: New Forms of Team-Machine Interaction in Health Care. *Methods Inf Med*. 2019;58(1):E14-E25.

49. Vokinger KN, Feuerriegel S, Kesselheim AS. Continual learning in medical devices: FDA's action plan and beyond. *Lancet Digit Health*. 2021;3(6):e337-e338.

50. Arambula AM, Bur AM. Ethical Considerations in the Advent of Artificial Intelligence in Otolaryngology. *Otolaryngol - Head Neck Surg U S*. 2020;162(1):38-39.

51. Eastwood C. *Sully: Miracle on the Hudson*. 2016.

52. Falk M. Artificial stupidity. *Interdiscip Sci Rev*. 2021;46(1-2):36-52.

53. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. *Science*. 2019;363(6433):1287-1289.

54. Roberts DC. The elephant in the room. *Nursing (Lond)*. 2020;50(12):42-46.

55. Lashbrook A. AI-Driven Dermatology Could Leave Dark-Skinned Patients Behind. The Atlantic. 2018. Accessed July 2, 2021. https://www.theatlantic.com/health/archive/2018/08/machine-learning-dermatology-skin-color/567619/

56. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion*. 2020;58:82-115.

57. Zuiderveen Borgesius F. *Discrimination, Artificial Intelligence, and Algorithmic Decision-Making*. 2018.

58. El-Bahnasawi M, Tekkis P, Kontovounisios C. Is it the surgeon or the technology performing the operation? *Tech Coloproctology*. 2019;23(9):933-934.

59. Woltz S, Krijnen P, Pieterse AH, Schipper IB. Surgeons' perspective on shared decision making in trauma surgery. A national survey. *Patient Educ Couns*. 2018;101(10):1748-1752.

60. Elwyn G, Nelson E, Hager A, Price A. Coproduction: When users define quality. *BMJ Qual Saf*. 2020;29(9):711-716.

61. UN. The Sustainable Development Goals: Our Framework for COVID-19 Recovery. 2020. Accessed October 17, 2020. https://www.un.org/sustainabledevelopment/sdgs-framework-for-covid-19-recovery/

62. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med*. 2022;28(1):31-38.

63. Healey MA, Shackford SR, Osler TM, Rogers FB, Burns E. Complications in Surgical Patients. *ARCH SURG*. 2002;137.

64. Loftus TJ, Filiberto AC, Balch J, et al. Intelligent, Autonomous Machines in Surgery. *J Surg Res*. 2020;253:92-99.

65. Bertsimas D, Dunn J, Velmahos GC, Kaafarani HMA. Surgical Risk Is Not Linear: Derivation and Validation of a Novel, User-friendly, and Machine-learning-based Predictive OpTimal Trees in Emergency Surgery Risk (POTTER) Calculator. *Ann Surg*. 2018;268(4):574-583.

66. De Simone B, Abu-Zidan FM, Gumbs AA, et al. Knowledge, attitude, and practice of artificial intelligence in emergency and trauma surgery, the ARIES project: an international web-based survey. *World J Emerg Surg*. 2022;17(1):10.

67. Andersson J, Rosell M, Kockum K, Lilja-Lund O, Söderström L, Laurell K. Prevalence of idiopathic normal pressure hydrocephalus: A prospective, population-based study. *PloS One*. 2019;14(5):e0217705.

68. Nakajima M, Yamada S, Miyajima M, et al. Guidelines for Management of Idiopathic Normal Pressure Hydrocephalus (Third Edition): Endorsed by the Japanese Society of Normal Pressure Hydrocephalus. *Neurol Med Chir (Tokyo)*. 2021;61(2):63-97.

69. Relkin N, Marmarou A, Klinge P, Bergsneider M, Black PMcL. Diagnosing Idiopathic Normal-pressure Hydrocephalus. *Neurosurgery*. 2005;57(suppl_3):S2-4-S2-16.

70. Kitagaki H, Mori E, Ishii K, Yamaji S, Hirono N, Imamura T. CSF spaces in idiopathic normal pressure hydrocephalus: morphology and volumetry. *AJNR Am J Neuroradiol*. 1998;19(7):1277-1284.

71. Vassilouthis J. The syndrome of normal-pressure hydrocephalus. *J Neurosurg*. 1984;61(3):501-509.

72. Kurihara Y, Simonson TM, Nguyen HD, Fisher DJ, Sato CSLY, Yuh WTC. Mr imaging of ventriculomegaly—a qualitative and quantitative comparison of communicating hydrocephalus, central atrophy, and normal studies. *J Magn Reson Imaging*. 1995;5(4):451-456.

73. The study of INPH on neurological improvement (SINPHONI), Hashimoto M, Ishikawa M, Mori E, Kuwana N. Diagnosis of idiopathic normal pressure hydrocephalus is supported by MRI-based scheme: a prospective cohort study. *Cerebrospinal Fluid Res*. 2010;7(1):18.

74. Iseki C, Kawanami T, Nagasawa H, et al. Asymptomatic ventriculomegaly with features of idiopathic normal pressure hydrocephalus on MRI (AVIM) in the elderly: A prospective study in a Japanese population. *J Neurol Sci*. 2009;277(1-2):54-57.

75. Siasios I, Kapsalaki EZ, Fountas KN, et al. The role of diffusion tensor imaging and fractional anisotropy in the evaluation of patients with idiopathic normal pressure hydrocephalus: a literature review. *Neurosurg Focus*. 2016;41(3):E12.

76. Tuniz F, Vescovi MC, Bagatto D, et al. The role of perfusion and diffusion MRI in the assessment of patients affected by probable idiopathic normal pressure hydrocephalus. A cohort-prospective preliminary study. *Fluids Barriers CNS*. 2017;14(1):24.

77. Bagatto D, Piccolo D, Fabbro S, et al. Intravoxel incoherent motion magnetic resonance imaging in the assessment of brain microstructure and perfusion in idiopathic normal-pressure hydrocephalus. *Neuroradiology*. Published online January 26, 2024.

78. Tawfik AM, Elsorogy L, Abdelghaffar R, Naby AA, Elmenshawi I. Phase-Contrast MRI CSF Flow Measurements for the Diagnosis of Normal-Pressure Hydrocephalus: Observer Agreement of Velocity Versus Volume Parameters. *Am J Roentgenol*. 2017;208(4):838-843.

79. Fabbro S, Piccolo D, Vescovi MC, et al. Resting-state functional-MRI in iNPH: can default mode and motor networks changes improve patient selection and outcome? Preliminary report. *Fluids Barriers CNS*. 2023;20(1):7.

80. Lenfeldt N, Hauksson J, Birgander R, Eklund A, Malm J. Improvement after cerebrospinal fluid drainage is related to levels of N-acetyl-aspartate in idiopathic normal pressure hydrocephalus. *Neurosurgery*. 2008;62(1):135-142.

81. Townley RA, Botha H, Graff-Radford J, et al. 18F-FDG PET-CT pattern in idiopathic normal pressure hydrocephalus. *NeuroImage Clin*. 2018;18:897-902.

82. Ohmichi T, Kondo M, Itsukage M, et al. Usefulness of the convexity apparent hyperperfusion sign in 123I-iodoamphetamine brain perfusion SPECT for the diagnosis of idiopathic normal pressure hydrocephalus. *J Neurosurg*. 2019;130(2):398-405.

83. Mihalj M, Dolić K, Kolić K, Ledenko V. CSF tap test — Obsolete or appropriate test for predicting shunt responsiveness? A systemic review. *J Neurol Sci*. 2016;362:78-84.

84. Wikkelsö C, Andersson H, Blomstrand C, Lindqvist G, Svendsen P. Predictive value of the cerebrospinal fluid tap-test. *Acta Neurol Scand*. 1986;73(6):566-573.

85. Feletti A, d'Avella D, Wikkelsø C, et al. Ventriculoperitoneal Shunt Complications in the European Idiopathic Normal Pressure Hydrocephalus Multicenter Study. *Oper Neurosurg*. 2019;17(1):97-102.

86. Mahr CV, Dengl M, Nestler U, et al. Idiopathic normal pressure hydrocephalus: diagnostic and predictive value of clinical testing, lumbar drainage, and CSF dynamics. *J Neurosurg*. 2016;125(3):591-597.

87. Scully AE, Lim ECW, Teow PP, Tan DML. A systematic review of the diagnostic utility of simple tests of change after trial removal of cerebrospinal fluid in adults with normal pressure hydrocephalus. *Clin Rehabil*. 2018;32(7):942-953.

88. Raneri F, Zella MAS, Di Cristofori A, Zarino B, Pluderi M, Spagnoli D. Supplementary Tests in Idiopathic Normal Pressure Hydrocephalus: A Single-Center Experience with a Combined Lumbar Infusion Test and Tap Test. *World Neurosurg*. 2017;100:567-574.

89. Koivisto AM, Alafuzoff I, Savolainen S, et al. Poor Cognitive Outcome in Shunt-Responsive Idiopathic Normal Pressure Hydrocephalus. *Neurosurgery*. 2013;72(1):1-8.

90. Hamilton R, Patel S, Lee EB, et al. Lack of shunt response in suspected idiopathic normal pressure hydrocephalus with Alzheimer disease pathology. *Ann Neurol*. 2010;68(4):535-540.

91. Malm J, Graff-Radford NR, Ishikawa M, et al. Influence of comorbidities in idiopathic normal pressure hydrocephalus — research and clinical care. A report of the ISHCSF task force on comorbidities in INPH. *Fluids Barriers CNS*. 2013;10(1):22.

92. Kang K, Yoon U, Lee JM, Lee HW. Idiopathic normal-pressure hydrocephalus, cortical thinning, and the cerebrospinal fluid tap test. *J Neurol Sci*. 2013;334(1-2):55-62.

93. Kang K, Han J, Lee SW, et al. Abnormal cortical thickening and thinning in idiopathic normal-pressure hydrocephalus. *Sci Rep*. 2020;10(1):21213.

94. Lang S, Dimond D, Isaacs AM, et al. Use of cortical volume to predict response to temporary CSF drainage in patients with idiopathic normal pressure hydrocephalus. *J Neurosurg*. 2023;139(6):1776-1783.

95. Frisoni GB, Fox NC, Jack CR, Scheltens P, Thompson PM. The clinical use of structural MRI in Alzheimer disease. *Nat Rev Neurol*. 2010;6(2):67-77.

96. Du AT, Schuff N, Kramer JH, et al. Different regional patterns of cortical thinning in Alzheimer's disease and frontotemporal dementia. *Brain*. 2006;130(4):1159-1166.

97. Dickerson BC, Bakkour A, Salat DH, et al. The Cortical Signature of Alzheimer's Disease: Regionally Specific Cortical Thinning Relates to Symptom Severity in Very Mild to Mild AD Dementia and is Detectable in Asymptomatic Amyloid-Positive Individuals. *Cereb Cortex*. 2009;19(3):497-510.

98. Danti S, Toschi N, Diciotti S, et al. Cortical thickness in *de novo* patients with Parkinson disease and mild cognitive impairment with consideration of clinical phenotype and motor laterality. *Eur J Neurol*. 2015;22(12):1564-1572.

99. Kim HJ, Ye BS, Yoon CW, et al. Cortical thickness and hippocampal shape in pure vascular mild cognitive impairment and dementia of subcortical type. *Eur J Neurol*. 2014;21(5):744-751.

100. Sotoudeh H, Sadaatpour Z, Rezaei A, et al. The Role of Machine Learning and Radiomics for Treatment Response Prediction in Idiopathic Normal Pressure Hydrocephalus. *Cureus*. 2021;13(10):e18497.

101. Mládek A, Gerla V, Skalický P, et al. Prediction of Shunt Responsiveness in Suspected Patients With Normal Pressure Hydrocephalus Using the Lumbar Infusion Test: A Machine Learning Approach. *Neurosurgery*. 2022;90(4):407-418.

102. Van Der Laan MJ, Polley EC, Hubbard AE. Super Learner. *Stat Appl Genet Mol Biol*. 2007;6(1).

103. Ashburner J, Friston KJ. Unified segmentation. *NeuroImage*. 2005;26(3):839-851.

104. Dahnke R, Yotter RA, Gaser C. Cortical thickness and central surface estimation. *NeuroImage*. 2013;65:336-348.

105. Giordan E, Palandri G, Lanzino G, Murad MH, Elder BD. Outcomes and complications of different surgical treatments for idiopathic normal pressure hydrocephalus: a systematic review and meta-analysis. *J Neurosurg*. Published online November 1, 2018:1-13.

106. Klassen BT, Ahlskog JE. Normal pressure hydrocephalus: how often does the diagnosis hold water? *Neurology*. 2011;77(12):1119-1125.

107. Klinge P, Hellström P, Tans J, Wikkelsø C, European iNPH Multicentre Study Group. One-year outcome in the European multicentre study on iNPH. *Acta Neurol Scand*. 2012;126(3):145-153.

108. Toma AK, Papadopoulos MC, Stapleton S, Kitchen ND, Watkins LD. Systematic review of the outcome of shunt surgery in idiopathic normal-pressure hydrocephalus. *Acta Neurochir (Wien)*. 2013;155(10):1977-1980.

109. Belgrado E, Tereshko Y, Tuniz F, et al. MDS-UDPRS-III in the diagnosis of idiopathic Normal Pressure Hydrocephalus and identification of candidates for Ventriculo-Peritoneal Shunting surgery. Results from a retrospective large cohort of patients. *J Neurol Sci*. 2023;445:120536.

110. Stupp R, Weller M, Belanger K, et al. Radiotherapy plus Concomitant and Adjuvant Temozolomide for Glioblastoma. *N Engl J Med*. Published online 2005.

111. Louis DN, Perry A, Wesseling P, et al. The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro-Oncol*. 2021;23(8):1231-1251.

112. Verhaak RGW, Hoadley KA, Purdom E, et al. Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*. 2010;17(1):98-110.

113. Patel AP, Tirosh I, Trombetta JJ, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*. 2014;344(6190):1396-1401.

114. Neftel C, Laffy J, Filbin MG, et al. An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. *Cell*. 2019;178(4):835-849.e21.

115. Puchalski RB, Shah N, Miller J, et al. An anatomic transcriptional atlas of human glioblastoma. *Science*. 2018;360(6389):660-663.

116. Pati S, Verma R, Akbari H, et al. Reproducibility analysis of multi-institutional paired expert annotations and radiomic features of the Ivy Glioblastoma Atlas Project (Ivy GAP) dataset. *Med Phys*. 2020;47(12):6039-6052.

117. Park JE, Oh JY, Park DH, et al. Mapping tumor habitats in isocitrate dehydrogenase - wild type glioblastoma: Integrating MRI, pathologic, and RNA data from the Ivy Glioblastoma Atlas Project. *Neuro-Oncol*. 2025;27(1):291-301.

118. Le NQK, Hung TNK, Do DT, Lam LHT, Dang LH, Huynh TT. Radiomics-based machine learning model for efficiently classifying transcriptome subtypes in glioblastoma patients from MRI. *Comput Biol Med*. 2021;132:104320.

119. Hu LS, D'Angelo F, Weiskittel TM, et al. Integrated molecular and multiparametric MRI mapping of high-grade glioma identifies regional biologic signatures. *Nat Commun*. 2023;14(1):6066.

120. Beig N, Bera K, Prasanna P, et al. Radiogenomic-Based Survival Risk Stratification of Tumor Habitat on Gd-T1w MRI Is Associated with Biological Processes in Glioblastoma. *Clin Cancer Res Off J Am Assoc Cancer Res*. 2020;26(8):1866-1876.

121. Zhang Z, Liu Y, Zhang Z, et al. MRI-based radiomic clustering identifies a glioblastoma subtype enriched for neural stemness and proliferative programs. *Front Oncol*. 2025;15:1662401.

122. Hsu JBK, Lee GA, Chang TH, et al. Radiomic Immunophenotyping of GSEA-Assessed Immunophenotypes of Glioblastoma and Its Implications for Prognosis: A Feasibility Study. *Cancers*. 2020;12(10):3039.

123. Dextraze K, Saha A, Kim D, et al. Spatial habitats from multiparametric MR imaging are associated with signaling pathway activities and survival in glioblastoma. *Oncotarget*. 2017;8(68):112992-113001.

124. Grossmann P, Gutman DA, Dunn WD, Holder CA, Aerts HJWL. Imaging-genomics reveals driving pathways of MRI derived volumetric tumor phenotype features in Glioblastoma. *BMC Cancer*. 2016;16:611.

125. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545-15550.

126. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*. 2013;14:7.

127. Kuhn M. Building Predictive Models in *R* Using the **caret** Package. *J Stat Softw*. 2008;28(5).

128. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser B Methodol*. 1995;57(1):289-300.

129. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using **lme4**. *J Stat Softw*. 2015;67(1).

130. Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest Package: Tests in Linear Mixed Effects Models. *J Stat Softw*. 2017;82:1-26.

131. Nakagawa S, Schielzeth H. A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods Ecol Evol*. 2013;4(2):133-142.

132. Lüdecke D, Ben-Shachar MS, Patil I, Waggoner P, Makowski D. performance: An R Package for Assessment, Comparison and Testing of Statistical Models. *J Open Source Softw*. 2021;6(60):3139.

133. Holm S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand J Stat*. 1979;6(2):65-70.

134. Friedman JH, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw*. 2010;33:1-22.

135. Zou H, Hastie T. Regularization and Variable Selection Via the Elastic Net. *J R Stat Soc Ser B Stat Methodol*. 2005;67(2):301-320.

136. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. *J R Stat Soc Ser B Methodol*. 1996;58(1):267-288.

137. Harrell , FE. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer International Publishing; 2015.

138. Meinshausen N, Bühlmann P. Stability Selection. *J R Stat Soc Ser B Stat Methodol*. 2010;72(4):417-473.

139. Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat Med*. 2019;38(7):1276-1296.

140. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ*. 2020;368:m441.

141. Kocak B, Baessler B, Bakas S, et al. CheckList for EvaluAtion of Radiomics research (CLEAR): a step-by-step reporting guideline for authors and reviewers endorsed by ESR and EuSoMII. *Insights Imaging*. 2023;14(1):75.

142. Hambardzumyan D, Gutmann DH, Kettenmann H. The role of microglia and macrophages in glioma maintenance and progression. *Nat Neurosci*. 2016;19(1):20-27.

143. Sharma P, Aaroe A, Liang J, Puduvalli VK. Tumor microenvironment in glioblastoma: Current and emerging concepts. *Neuro-Oncol Adv*. 2023;5(1):vdad009.

144. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. *PLOS ONE*. 2019;14(11):e0224365.