Article

The SAIGE Framework for Risk Stratification in Spine Surgery

Rahul Kumar * , Harlene Kaur , Kyle Sporn , Alejandro Damian , Yusuf Zain-Ansari , Phani Paladugu , Ram Jagadeesan , Louis Clarkson , Nasif Zaman , Alireza Tavakkoli

Posted Date: 1 December 2025

doi: 10.20944/preprints202512.0003.v1

Keywords: AI in surgery; risk stratification; spine surgery; regulatory oversight; SAIGE framework; clinical accountability

Article

The SAIGE Framework for Risk Stratification in Spine Surgery

Rahul Kumar [1], Harlene Kaur [1], Kyle Sporn [2], Alejandro Damian [3], Yusuf Zain-Ansari [4], Phani Paladugu [5,6], Ram Jagadeesan [7,8], Louis Clarkson [9,10], Nasif Zaman [11] and Alireza Tavakkoli [11]

[1] UMass Chan School of Medicine, Worcester, MA, USA
[2] Upstate Medical University Norton College of Medicine, Syracuse, NY, USA
[3] Wayne State University School of Medicine, Detroit, MI, USA
[4] Temple University, Philadelphia, PA, USA
[5] Sidney Kimmel Medical College at Thomas Jefferson University, Philadelphia, PA, USA
[6] Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
[7] Johns Hopkins Whiting School of Engineering, Baltimore, MD, USA
[8] Artificial Intelligence Systems, Cisco Inc, San Jose, CA, USA
[9] Royal Free Hospital, London, United Kingdom
[10] School of Medicine, University of Cambridge, Cambridge, United Kingdom
[11] Human-Machine Perception Laboratory, Department of Computer Science, University of Nevada Reno, Reno, NV, USA
* Correspondence: rahul.kumar5@umassmed.edu

Abstract

AI and ML in spine surgery has both transformative possibilities and notable challenges related to regulatory oversight, algorithmic bias, and clinical responsibility. We propose a governance model to tackle these important issues, ensuring the responsible use of AI tools. This framework introduces the SAIGE-R Index, a tool designed to measure AI system risks based on Clinical Volatility, System Integration Risk, and Data Integrity Confidence. This index supports a tiered oversight system, ranging from minimal checks for low-risk systems to thorough FDA reviews for high-risk applications. In addition, SAIGE sets specific validation standards focused on spine surgery outcomes. These include important differences in patient-reported measures and accuracy in pedicle screw placement, along with quarterly fairness checks to reduce demographic bias. The framework also describes a strong governance structure that focuses on ongoing clinician training, involvement from multiple stakeholders, and strict data security measures. It suggests a liability model that matches responsibility with the evaluated risk level of AI tools. By addressing validation, ethics, and accountability, the SAIGE Framework provides a foundation for safely and effectively incorporating AI into complex surgical settings. This approach encourages innovation while maintaining patient safety and clinical integrity.

Keywords: AI in surgery; risk stratification; spine surgery; regulatory oversight; SAIGE framework; clinical accountability

Introduction

Artificial intelligence (AI) and machine learning (ML) are progressively transforming surgery. By leveraging these tools, clinicians can optimize diagnosis, better predict risk, and enhance operative planning [1,2]. However, their scope expands alongside related risks. Even thoroughly developed algorithms can misjudge anatomical planes, undervalue comorbid risks, or ignore presumably insignificant findings during image-guided interventions. While AI/ML tools can help surgeons mitigate error and improve decision-making, they simultaneously introduce a new layer of uncertainty regarding clinical responsibility and raise an important question: who is liable when an AI-assisted decision leads to harm? Thus, unregulated AI may serve as both an aid and a source of uncertainty.

I. Risks of Unregulated or Underregulated AI Deployment

The urgency of this issue is particularly important as many AI systems begin to be incorporated into clinical environments without US Food and Drug Administration (FDA) clear validation. For instance, Clark et al. found that out of 119 selected medical devices that were marketed as enabled for AI or ML, 23 (19.3%) were either “discrepant” or “contentious” in terms of their marketing versus clearance [3]. Additionally, many AI/ML technologies are validated using only computer-generated ("phantom") images, which may not reflect real-world clinical scenarios. A similar report by Chouffani El Fassi et al. demonstrated that, out of 521 AI-driven clinical devices approved by the FDA, 226 of them lacked published clinical validation data and had not been trained on real patient data [4]. Such findings are alarming and emphasize how urgently regulatory authorities must improve FDA evaluation criteria and set continuous, real-time performance monitoring. Without practical, unbiased data training and post-market protections, “shadow” AI models may unwittingly spread prejudice, raise clinical risk, and hide medico-legal responsibility.

In spine surgery, factors such as calcified disc morphology, herniation laterality, cage subsidence probability, and synovial cyst presence define surgical planning and intraoperative adjustments [5]. These factors are further influenced by patient-specific comorbidities, socioeconomic level, and racial and ethnic background. Models built from non-representative, potentially sanitized data may prove inaccurate when faced with real-world complexity. For instance, an AI model trained on idealized data might overlook infection risk in a diabetic patient with calcified disc morphology, leading to unforeseen complications like reoperation. Such reductions in care quality can occur in large academic hospitals or smaller, inner-city hospital groups, exacerbated by inadequate algorithmic safety assessments and outcome auditing. Without timely model updates to reflect changing patient demographics, key metrics like reoperation rates, nonunion incidence, and adjacent segment disease may be incorrectly reported. These challenges demand a rigorous framework to ensure AI enhances spine surgery outcomes through a structured approach to validation, ethics, and accountability.

II. Current Gaps in AI/ML Technologies

Large language models (LLMs) are increasingly being integrated to streamline documentation, imaging interpretation, and patient education [6]. However, current LLMs such as ChatGPT lack calibration for high-stakes surgical environments. These systems are not trained on domain-specific imaging modalities like T2-weighted MRIs, CT myelograms, or fluoroscopic guidance, and they lack intraoperative situational awareness, neuromonitoring interpretation, and tactile feedback integration [7]. Additionally, LLMs may not take into account previous surgical changes (e.g., pseudoarthrosis or leftover hardware) and specific body details (such as pedicle dysplasia or facet tropism), often depending on general language data or radiology images and texts that miss important biomechanical details. When LLMs are used in test cases, such as in Orthopaedic In-Training Examination evaluations [8], they do not perform accurately and thus can not be used as a reliable tool. Another core issue is the failure of contextual transfer: AI systems trained on narrow datasets tend to overfit to regional practices and underperform across diverse healthcare systems [9]. Without continuous learning mechanisms like federated updating or real-time drift detection, these models degrade over time. Compounding this, current regulatory structures rarely mandate post-market performance validation or adaptive monitoring.

III. Shadow AI and Regulatory Gaps

"Shadow AI" refers to how generative models are used in clinical practice to document, bill, or communicate with patients despite not being officially sanctioned [10]. These tools often operate outside HIPAA-compliant platforms, posing serious privacy risks by exposing protected health information (PHI) to third-party servers without audit trails or version control. Compared to structured EHR systems, generative models obscure the provenance of clinical content and therein create medico-legal vulnerabilities in surgical planning and documentation. Additionally, the FDA's Software as a Medical Device (SaMD) [11] and Clinical Decision Support (CDS) pathways [12] were not designed for these adaptive and generative systems. HIPAA likewise governs only covered entities and excludes many AI developers. Ultimately, no unified liability doctrine addresses the role of generative AI in surgical complications [13], and algorithmic contributions are rarely disclosed in clinical consent processes, raising serious concerns about autonomy and informed decision-making.

Risk Stratification Architecture

We created the Spine Artificial Intelligence Governance and Evaluation (SAIGE) framework to address urgent challenges in surgical AI oversight, including algorithmic bias, under-specified validation criteria, and evolving medico-legal uncertainties in high-risk procedural contexts. The SAIGE framework operationalizes these goals through a structured governance architecture built on: (1) predictive risk stratification via the SAIGE-R index, (2) longitudinal demographic equity auditing, and (3) federated validation pipelines that preserve data privacy across institutions. We have also anchored outcome prediction through analyzing the American College of Surgeons Risk Calculator [14] and national registries such as the Nationwide Ambulatory Surgery Sample (NASS) [15] and the North American Spine Society database [16]. In addition, to facilitate HIPAA-compliant generalizability, SAIGE supports continuous, real-world model calibration across heterogeneous clinical environments by accounting for federated learning. Our immediate goal in designing the SAIGE framework is to put forward a pragmatic and evidence-grounded starting point for AI governance in spine surgery. We hope that this discussion allows us to refine and reshape our framework and account for multi-stakeholder engagement while also tailoring our framework to specific spine and related surgical procedures. Below, we outline the core components of SAIGE.

I. SAIGE-R Risk Index

The SAIGE-R Index is a quantitative risk stratification tool designed to help surgeons and developers evaluate AI systems used in spinal surgery across three dimensions: Clinical Volatility (CV), System Integration Risk (SIR), and Data Integrity Confidence (DIC). In essence, we attempt to standardize deployment risks, and each dimension of SAIGE-R reflects a unique deployment risk and challenge.

a. Clinical Volatility (CV)

We estimate the clinical impact of AI errors through CV, factoring in procedural complexity and anatomical risk. We calculate CV by multiplying a baseline risk value, derived from NASS-QOD complication rates (e.g., 3.0 for MIS-TLIF, reflecting an 8-15% complication risk) [17], by device complexity multipliers we assigned from FDA device classifications (e.g., 1.5× for Class III robotic systems, 1.2× for Class II surgical planners) and spine-specific anatomical modifiers (e.g., 1.3× for cervicothoracic junction, 1.4× for revision surgery) [18]. We based these multipliers on meta-analyses showing elevated risks in complex procedures [19–22]. Reported complication rates for transforaminal lumbar interbody fusion (TLIF) typically range from 10-20% in primary cases, with substantially higher rates observed in revision surgeries and in patients with underlying deformity or multilevel pathology [23–25]. For example, a Class III robot in a revision MIS-TLIF yields a CV of $3.0 \times 1.5 \times 1.4 = 6.3$. We cap CV at 10 to reflect the maximum feasible clinical risk, ensuring a standardized scale for prioritizing oversight.

b. System Integration Risk (SIR)

We assess SIR on a 0-10 scale, using standards like ASTM F2761 to evaluate AI system interoperability with surgical infrastructure [26], including imaging synchronization, fail-safe redundancy, and real-time performance. Systems achieving latency under 2 seconds, critical for timely surgical feedback, and integrating seamlessly without altering standard procedures score 8 or higher [27,28]. For example, an AI system with 1.5-second latency may achieve SIR = 8, reflecting robust integration. The maximum SIR of 10 indicates optimal performance with no discernible latency or disruption.

c. Data Integrity Confidence (DIC)

We quantify the robustness of AI training data through DIC, emphasizing real-world data over synthetic data to enhance reliability in spine surgery. To ensure representation of underserved populations, we require ≥30% of real-world data from safety-net hospitals [29,30]. We calculate DIC as:

$$DIC = \frac{Real - world\ \%}{Synthetic\ \% \ + \ 0.01} \times 5.67$$

This formula yields DIC = 5.67 when real-world and synthetic data are equal (50%/50%), serving as a baseline for data fidelity. The constant 0.01 prevents division by zero when synthetic data is absent. Higher DIC values reflect greater confidence in data integrity, with the maximum achieved when only real-world data is used.

We then computed the composite SAIGE-R score using an Euclidean distance formula:

$$\text{SAIGE-R} = \sqrt{(\frac{CV}{CV_{max}})^2 + (1 - \frac{SIR}{SIR_{max}})^2 + (\frac{DIC_{min}}{DIC})^2}$$

Where $CV_{max} = 10$, $SIR_{max} = 10$, and $DIC_{min} = 5.67$. The risk score ranges from 0 to a theoretical maximum of $\sqrt{3} \approx 1.732$, though practical values will typically be <1. In all, a higher CV, lower SIR, or lower DIC will raise the SAIGE-R score and indicate increased clinical risk and oversight needs. We also stratify AI systems into three risk tiers based on their composite score:

1. Low risk: ≤ 0.4 (minimal oversight)
2. Moderate risk: 0.4 - 0.8 (quarterly audits)
3. High risk: > 0.8 (FDA High-Risk AI Committee review)

To illustrate, we present a hypothetical AI model assisting pedicle screw placement in minimally invasive transforaminal lumbar interbody fusion (MIS-TLIF) within a Class III robotic environment. We estimate its CV at 5.4, derived from a baseline of 3.0 (reflecting 8-15% c in omplication rates [31]), multiplied by 1.5 for the Class III device, and adjusted by a 1.2 anatomical modifier for specific patient factors like high BMI. The SIR is 8, indicating low latency (<2 seconds) and seamless workflow integration [27,28]. The DIC is approximately 41.57, calculated from 88% real-world and 12% synthetic data using

$$DIC = \frac{88}{12 + 0.01} \times 5.67 \ \approx \ 7.326 \times 5.67 \ \approx \ 41.57$$

Applying the SAIGE-R formula:

$$\text{SAIGE-R} = \sqrt{(\frac{5.4}{10})^2 + (1 - \frac{8}{10})^2 + (\frac{5.67}{41.57})^2} \ = \sqrt{0.2916 \ + \ 0.04 \ + \ 0.0186} \ \approx \ 0.59$$

Thus, the SAIGE-R score of 0.59 places the system within the moderate-risk tier (0.4-0.8). This triggers quarterly audits, which we made consistent with the use of Class III devices in spine surgery. Though hypothetical, this example demonstrates how the SAIGE-R framework translates multidimensional risk factors into an initial oversight pathway.

We also designed the SAIGE-R index to be a dynamic surveillance tool. To illustrate, we present a hypothetical scenario inspired by robotic systems like the Mazor X Stealth Edition, which is used in complex spinal deformity corrections [32]. Initially, the system has a CV of 7, reflecting a high-risk procedure with Class III robotics and sensitive trajectories, an SIR of 8, indicating robust intraoperative coordination with latency under 2 seconds, and a DIC of 41.57, which we derived from 88% real-world and 12% synthetic data. The initial SAIGE-R score is calculated as:

$$\text{SAIGE-R} = \sqrt{(\frac{7}{10})^2 + (1 - \frac{8}{10})^2 + (\frac{5.67}{41.57})^2} \ = \sqrt{0.49 \ + \ 0.04 \ + 0.0186} \ \approx \ 0.74$$

This places the system in the moderate-risk tier (0.4-0.8), triggering quarterly audits. If intraoperative synchronization issues, such as imaging lag, reduce SIR to 6, and retraining shifts data to 70% real-world and 30% synthetic, lowering DIC to 13.23, the updated SAIGE-R score becomes:

$$\text{SAIGE-R}_{\text{updated}} = \sqrt{(\tfrac{7}{10})^2 + (1 - \tfrac{6}{10})^2 + (\tfrac{5.67}{13.23})^2} \; = \sqrt{0.49 \; + \; 0.16 \; + 0.1837} \; \approx \; 0.91$$

The updated SAIGE-R score enters the high-risk tier (>0.8). This warrants enhanced regulatory scrutiny under our framework and potentially includes FDA review. Clinically, such issues could lead to a 3 mm pedicle breach (Gertzbein-Robbins Grade C) if osteoporotic bone is misclassified [33], risking neural or vascular structures. Though hypothetical, this scenario reflects plausible AI misclassification risks in spinal navigation. By tracking these dynamic risk transitions, we ensure the SAIGE-R Index adapts to changes in system performance or data quality and is in line with continuous oversight standards.

II. SAIGE Validation Benchmarks

In this section, we establish spine-specific certification criteria to evaluate AI system performance across three critical domains: clinical efficacy, demographic equity, and post-market surveillance. These domains address the unique challenges of spinal surgery, including high anatomical variability, complex procedural steps, and the need for precision in interventions such as pedicle screw placement or sagittal alignment restoration. Unlike generic AI benchmarks, which often lack metrics tailored to spinal surgery (e.g., pedicle breach rates or pelvic incidence-lumbar lordosis [PI-LL] mismatch) [34–36], our framework emphasizes surgical accuracy, outcome relevance, and fairness across diverse patient populations. This approach is distinct from the FDA's Digital Health Software Precertification Program, which provides general software validation criteria but lacks domain-specific metrics tailored to the risks of spinal procedures [34,35].

For AI systems directly involved in surgical planning or execution, we define functional standards grounded in validated clinical outcomes [35,36]. These benchmarks, informed by established literature, represent realistic expectations for AI performance in spinal surgery. One such benchmark is the Minimum Clinically Important Difference (MCID) in the Oswestry Disability Index (ODI), which is set at a ΔODI of at least 12 points. This threshold, consistent with outcomes following lumbar fusion, is widely recognized as the minimum change reflecting a clinically meaningful improvement in patient-reported function [37]. Another key benchmark is pedicle screw placement accuracy, set at 97-98% based on the Gertzbein Grade A/B classification. Grade A denotes fully intrapedicular placement without cortical breach, while Grade B permits a breach of less than 2 mm [35,36]. These thresholds are aligned with the precision reported for advanced robotic systems such as the Mazor X platform and are considered clinically acceptable for minimizing neurological risk [38,40–42] (Parker et al., 2022; Zhang et al., 2019). Additionally, the benchmark for postoperative pelvic incidence-lumbar lordosis (PI-LL) mismatch is set at ≤10°, consistent with the SRS-Schwab classification of sagittal alignment [43,44] (Schwab et al., 2012). Although some literature supports a more stringent threshold of 9° to potentially reduce reoperation rates, recent studies suggest that a mismatch between 10° and 20° may be optimal for select patient populations [43,44] (Zhang et al., 2017). Thus, the ≤10° standard remains the preferred guideline for consistency across clinical protocols.

To address algorithmic bias over time, we conduct fairness audits on a quarterly basis using a bias score tailored for binary classification tasks:

$$\text{Bias Score}= \frac{1}{n}\sum_{i=1}^{n} \left| \frac{FN_i}{TP_i+FN_i} - \frac{FP_i}{TN_i+FP_i} \right|$$

This formula was chosen because it captures disparities in both underdiagnosis (elevated false negative rates) and overdiagnosis (elevated false positive rates) across demographic subgroups. In this context, false negatives (FNi) represent missed cases requiring intervention (i.e. undetected surgical candidates) while false positives (FPi) indicate inappropriate intervention recommendations. True positives (TPi) and true negatives (TNi) are correctly identified cases for each subgroup i, with n denoting the number of groups. We also track the maximum absolute disparity ($|\mathrm{FNR}_i - \mathrm{FPR}_i|$) to detect high-risk subgroup outliers. A bias score exceeding 0.15, which we selected as a conservative preliminary threshold pending empirical validation, triggers mandatory model retraining and subgroup-specific revalidation.

Through this validation framework, we aim to have our framework be demographically fair and continuously monitored across its lifecycle. By aligning with spine-specific surgical standards, acknowledging practical implementation constraints, and embedding equity metrics, we provide a robust certification pathway for surgical AI. Future iterations of this validation and SAIGE checks-and-balances will expand to accommodate generative AI models and real-time adaptive learning systems, further supporting safe surgical innovation.

Governance Structure and Oversight

We developed the SAIGE Framework in order to address the complex demands of surgical AI integration. This framework is a multi-layered governance architecture that is grounded in ethical accountability, performance validation, clinician training, and regulatory compliance. A central component of this framework is the idea that AI systems must build upon, not discard, clinical expertise while operating transparently across various surgical environments [1,3,12].

Another cornerstone of the SAIGE Framework is clinician competence. With consideration to the nuanced judgment necessary in spine surgery, we incorporated an eight-hour, CME-accredited training program that aligns with AMA Code of Medical Ethics Opinion 9.2.2 [45]. The course provides surgeons with structured guidance on interpreting AI-generated recommendations and integrating them with real-time findings in the OR. For example, the curriculum includes scenarios addressing how to handle pedicle screw positions suggested by AI that conflict with neuromonitoring alerts [20,41], or how to incorporate AI predictions into decisions involving sagittal balance correction and Modic change interpretation in lumbar fusion [30,40]. The training also emphasizes informed consent, ensuring that surgeons clearly communicate the role of AI in clinical decision-making and potential implications for patient-reported outcomes [15,45]. In order to maintain clinical relevance, the training is updated annually to reflect advances in AI capabilities as well as clinician feedback [6,7].

Not only does SAIGE emphasize individual education, but it also encourages stakeholder engagement to build both institutional and patient rapport. Based on a 2025 case study in *npj Digital Medicine* on best practices for AI governance [4], we implemented quarterly workshops where surgeons get to interact directly with AI tools as well as their developers. Also included in these sessions are demonstrations that demonstrate clinical accuracy, such as achieving a precision of 98% in pedicle screw placement for lumbar fusions [37,41]. Additionally, biannual feedback loops are important for clinical relevance since they gather patient and physician perspectives using validated measures. For patients, Patient-Reported Experience and Outcome Measures (PREMs and PROMs) are utilized, as well as the Oswestry Disability Index (ODI), Visual Analog Scale (VAS), and SF-36 [31,32]. Surgeons, on the other hand, assess AI usability using the System Usability Scale (SUS) [46]. These datasets are then subjected to quantitative analysis, including Likert scaling, ANOVA comparisons across demographic groups, and t-tests to detect statistically significant shifts in satisfaction or trust. Hospital administrators would then receive annual briefings on the outcomes of these feedback mechanisms, including cost-benefit analyses, AI accuracy data, and recommendations for funding allocation [19,36]. This process allows the SAIGE Framework to sustainably scale across healthcare systems with variable capabilities [16].

It is important not to overlook data security and regulatory compliance in regards to the SAIGE Framework. In accordance with HIPAA (45 CFR Part 164), GDPR (Regulation [EU] 2016/679), and FDA regulatory standards (21 CFR Part 820), the framework mandates stringent security measures [11,47,48]. These include the use of homomorphic encryption for data sharing, differential privacy during federated learning, and adherence to NIST SP 800-171 cybersecurity protocols [9,49]. Given the legal ambiguity around AI liability in high-risk environments, we propose a stratified liability model based on the SAIGE-R risk index. For systems deemed low-risk (SAIGE-R ≤ 0.4), clinicians retain primary responsibility and are governed by existing malpractice standards such as those articulated under the Medical Injury Compensation Reform Act in California (MICRA) [13,50]. When AI tools reach a high-risk threshold, however, the liability is now shared between developers and clinicians, assuming the system was validated and clinical protocols have been followed [10,11]. The addition of a 12-month safe harbor provision following AI certification would protect both parties from excessive legal risk, assuming they act in good faith [13]. Institutions are also encouraged to obtain AI-specific malpractice insurance to mitigate risks like anatomical-related misclassification or procedural misguidance as a result of system failure [13].

Informed consent protocols are adjusted in order to meet both ethical and legal standards. In line with AMA Opinion 9.2.2 and the California Health and Safety Code §24173, consent documents explicitly disclose AI use, its role in decision-making, and potential implications for patient outcomes [45,51]. Furthermore, the formation of a committee that oversees the handling of complaints and concerns through hospital ombudsman programs will ensure that patient autonomy is preserved. Audit processes that incorporate PROMs can be used to assess how AI interventions affect quality of life for patients. Such processes can aid in the development of spine-specific tools that assess neurocognitive function and procedural symptom burden, with inspiration from oncology instruments such as the EORTC QLQ-C30 [44].

Implementation of the SAIGE Framework necessitates a thorough economic evaluation. Cost-benefit analyses that take into account training expenditures, system integration costs, and infrastructure upgrades can be used as benchmarks for clinical gains like reductions in reoperation rates for procedures such as TLIF (currently around 8-15%) [5,16,33]. Funding strategies include public-private partnerships, grants through organizations like the NIH, and partnering with EHR vendors [19]. A gradual implementation strategy that prioritizes training and oversight mechanisms in low-resource areas would allow for consistent quality and equal access with expansion [19]. Administrative effort is also crucial, since hospital leaders will be the ones to coordinate cross-functional efforts that align AI governance with strategic goals of their respective institutions [16].

To ensure ethical standards are met, the SAIGE Framework aligns with global AI governance standards, which include the World Health Organization's six regulatory areas of transparency, risk management, external validation, data quality, regulatory compliance, and collaboration [52]. The framework also incorporates the FUTURE-AI guidelines (Fairness, Universality, Traceability, Usability, Robustness, and Explainability) [53]. These principles ensure that AI systems integrated into spine surgery adhere not only to local mandates but to international norms regarding AI accountability and data protection as well [52,53].

Feedback mechanisms in general are considered a vital feedback loop for continuous quality improvement. As mentioned earlier, patient perspectives are gathered through quarterly surveys using PREMs and PROMs, with Likert scores that are analyzed for mean satisfaction and standard deviation [31,32]. When it comes to spine surgery specifically, these metrics would be used to assess patient feedback in terms of things like pain management, mobility, overall satisfaction with AI-incorporated procedures, etc. Focus groups serve to explore patient experiences more qualitatively, ensuring broad demographic representation. Demographic representation is especially important because any existing disparities across various demographics would be emphasized when it comes to the performance of the AI system. Physician input is similarly evaluated using the SUS, along with conducting biannual t-tests to track changes in usability and acceptance [46]. A specialized committee will synthesize this feedback using inferential statistics, like ANOVA, to identify disparities across demographic groups or institutions [7,19]. This specialized committee would consist of spine surgeons, ethicists, patient advocates, and data scientists that would oversee the interpretation of this feedback. These findings will serve as guidance for updates to training protocols, system revalidation, and broader governance adjustments [19].

While the SAIGE Framework provides a solid basis for addressing the urgent challenges of surgical AI, there are still several limitations that exist and must be acknowledged. Some critical areas for further improvement include establishing clearer liability standards related to AI, improving privacy protections for generative AI models, and the validation of clinical benchmarks from prospective studies [10,13]. Generative models can be problematic since novel output is being generated using existing data, which is why privacy protections are more relevant here [47,48]. A potential initiative would be to enroll 1,000 patients to evaluate the real-world performance of AI-guided spine procedures using the full SAIGE Framework [1,33]. The implementation of the Framework across diverse health systems also requires additional research on models for funding and technical support for low-support hospitals [19]. As AI models evolve, especially ones that incorporate adaptive learning or generative features, specialized committees will implement strategies to distinguish between major and minor software updates in order to reduce major delays while preserving patient safety [10,11].

In conclusion, the SAIGE Framework provides a structured and ethically robust pathway for integrating AI into spine surgery. By combining meticulous oversight, stakeholder engagement, risk stratification, and regulatory compliance, it addresses the unique risks as well as responsibilities associated with surgical AI. Our goal is to not only manage these technologies, but to shape their path to reflect values like transparency and clinical accuracy [4,8,45].

Author Contributions: Rahul Kumar: Conceptualization; Methodology; Formal analysis; Writing - Original Draft; Visualization; Project administration. Rahul led the development of the SAIGE governance architecture, constructed the SAIGE-R risk index, performed mathematical modeling of Clinical Volatility, System Integration Risk, and Data Integrity Confidence, and drafted the primary manuscript text including the introduction, risk index design, regulatory analysis, and conclusion. Harlene Kaur: Data curation; Investigation; Writing - Review & Editing. Harlene contributed to literature synthesis across spine surgery AI applications, verified clinical accuracy of risk factors, supported parameter selection for SAIGE-R, and provided major revisions to the manuscript text. Kyle Sporn: Methodology; Software; Validation. Kyle assisted in formalizing the SAIGE-R computational workflow, cross-checked the Euclidean risk scoring formula, evaluated hypothetical model performance scenarios, and helped ensure consistency of the scaling functions and threshold definitions. Alejandro Damian: Investigation; Data curation; Writing - Review & Editing. Alejandro performed targeted literature extraction on MIS-TLIF outcomes, pedicle screw accuracy benchmarks, and postoperative alignment metrics, and contributed to the refinement of the SAIGE validation benchmarks section. Yusuf Zain-Ansari: Resources; Investigation; Writing - Review & Editing. Yusuf supported acquisition and verification of clinical outcome datasets referenced in the SAIGE framework and contributed edits focusing on spine-surgery applicability. Phani Paladugu: Formal analysis; Validation; Visualization. Phani contributed to fairness-auditing methods, designed and evaluated the bias-score equation, reviewed the statistical rationale for quarterly demographic audits, and advised on PROM/PREM feedback integration. Ram Jagadeesan: Software; Methodology; Supervision. Ram contributed to the federated learning workflow and interoperability requirements, clarified system-integration standards including latency thresholds and redundancy, and supervised technical sections describing data integrity and system integration risk. Louis Clarkson: Writing - Review & Editing; Resources. Louis reviewed clinical accuracy claims, strengthened international AI governance context, and contributed expertise from UK/EU spine surgery perspectives. Nasif Zaman: Supervision; Methodology; Validation; Writing - Review & Editing. Nasif supervised analytic rigor, reviewed the mathematical structure of DIC and SAIGE-R normalization, ensured reproducibility of formulas, and contributed major revisions aligning the framework with current AI governance standards. Alireza Tavakkoli: Supervision; Conceptualization; Funding acquisition. Dr. Tavakkoli oversaw conceptual development of the SAIGE Framework, validated the machine learning governance architecture, and ensured methodological coherence across computational and regulatory components.

Funding Statement: No funding was received for this project.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No data was involved in this study

doi:10.20944/preprints202512.0003.v1

Acknowledgments: None

Conflicts of Interest Statement: The authors declare no conflict of interest.

References

1. Pressman SM, Borna S, Gomez-Cabello CA, Haider SA, Haider CR, Forte AJ. Clinical and Surgical Applications of Large Language Models: A Systematic Review. J Clin Med. 2024 May 22;13(11):3041. doi: 10.3390/jcm13113041. PMID: 38892752; PMCID: PMC11172607.
2. Chatterjee S, Bhattacharya M, Pal S, Lee S-S, Chakraborty C. ChatGPT and large language models in orthopedics: from education and surgery to research. J Exp Orthop. 2023;10(1):128. doi:10.1186/s40634-023-00700-1
3. Clark P, Kim J, Aphinyanaphongs Y. Marketing and US Food and Drug Administration Clearance of Artificial Intelligence and Machine Learning Enabled Software in and as Medical Devices: A Systematic Review. JAMA Netw Open. 2023;6(7):e2321792. doi:10.1001/jamanetworkopen.2023.21792
4. Chouffani El Fassi, S., Abdullah, A., Fang, Y. et al. Not all AI health tools with regulatory authorization are clinically validated. Nat Med 30, 2718-2720 (2024). https://doi.org/10.1038/s41591-024-03203-3
5. Amorim-Barbosa T, Pereira C, Catelas D, Rodrigues C, Costa P, Rodrigues-Pinto R, Neves P. Risk factors for cage subsidence and clinical outcomes after transforaminal and posterior lumbar interbody fusion. Eur J Orthop Surg Traumatol. 2022 Oct;32(7):1291-1299. doi: 10.1007/s00590-021-03103-z. Epub 2021 Aug 31. PMID: 34462820.
6. Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: a scoping review of applications in medicine. Front Med (Lausanne). 2024 Oct 29;11:1477898. doi: 10.3389/fmed.2024.1477898. PMID: 39534227; PMCID: PMC11554522.
7. Janssen BV, Kazemier G, Besselink MG. The use of ChatGPT and other large language models in surgical science. BJS Open. 2023 Mar 7;7(2):zrad032. doi: 10.1093/bjsopen/zrad032. PMID: 36960954; PMCID: PMC10037421.
8. Kung JE, Marshall C, Gauthier C, Gonzalez TA, Jackson JB 3rd. Evaluating ChatGPT Performance on the Orthopaedic In-Training Examination. JB JS Open Access. 2023 Sep 8;8(3):e23.00056. doi: 10.2106/JBJS.OA.23.00056. PMID: 37693092; PMCID: PMC10484364.
9. Aliferis C, Simon G. Overfitting, Underfitting and General Model Overconfidence and Under-Performance Pitfalls and Best Practices in Machine Learning and AI. 2024 Mar 5. In: Simon GJ, Aliferis C, editors. Artificial Intelligence and Machine Learning in Health Care and Medical Sciences: Best Practices and Pitfalls [Internet]. Cham (CH): Springer; 2024. Available from: https://www.ncbi.nlm.nih.gov/books/NBK610560/ doi: 10.1007/978-3-031-39355-6_10
10. Krantz T, Jonker A, McGrath A. What is shadow AI? IBM. Published October 25, 2024. Accessed June 1, 2025. https://www.ibm.com/topics/shadow-ai
11. U.S. Food and Drug Administration. Software as a Medical Device (SaMD). FDA. Updated November 2018. Accessed June 1, 2025. https://www.fda.gov/medical-devices/digital-health-center-excellence/software-medical-device-samd
12. Elhaddad M, Hamam S. AI-Driven Clinical Decision Support Systems: An Ongoing Pursuit of Potential. Cureus. 2024 Apr 6;16(4):e57728. doi: 10.7759/cureus.57728. PMID: 38711724; PMCID: PMC11073764.
13. Price II WN, Gerke S, Cohen IG. Liability for use of artificial intelligence in medicine. In: Solaiman B, Cohen IG, editors. Research Handbook on Health, AI and the Law. Cheltenham, UK: Edward Elgar Publishing Ltd; 2024 Jul 16. Chapter 9. Available from: https://www.ncbi.nlm.nih.gov/books/NBK613216/ doi: 10.4337/9781802205657.ch09
14. Liu Y, Ko CY, Hall BL, Cohen ME. American College of Surgeons NSQIP Risk Calculator Accuracy Using a Machine Learning Algorithm Compared with Regression. J Am Coll Surg. 2023 May 1;236(5):1024-1030. doi: 10.1097/XCS.0000000000000556. Epub 2023 Jan 12. PMID: 36728295.
15. McDermott KW, Liang L. Overview of Major Ambulatory Surgeries Performed in Hospital-Owned Facilities, 2019. 2021 Dec 21. In: Healthcare Cost and Utilization Project (HCUP) Statistical Briefs [Internet].

Rockville (MD): Agency for Healthcare Research and Quality (US); 2006 Feb-. Statistical Brief #287. Available from: https://www.ncbi.nlm.nih.gov/books/NBK577044/

16. Daltroy LH, Cats-Baril WL, Katz JN, Fossel AH, Liang MH. The North American spine society lumbar spine outcome assessment Instrument: reliability and validity tests. Spine (Phila Pa 1976). 1996 Mar 15;21(6):741-9. doi: 10.1097/00007632-199603150-00017. PMID: 8882698.
17. Epstein NE. Lower complication and reoperation rates for laminectomy rather than MI TLIF/other fusions for degenerative lumbar disease/spondylolisthesis: a review. Surg Neurol Int. 2018;9:55. doi:10.4103/sni.sni_26_18
18. U.S. Food and Drug Administration. Step 3: Pathway to Approval. The Device Development Process. Accessed June 1, 2025. https://www.fda.gov/patients/device-development-process/step-3-pathway-approval
19. Tormenti MJ, Maserati MB, Bonfield CM, Gerszten PC, Moossy JJ, Kanter AS, Spiro RM, Okonkwo DO. Perioperative surgical complications of transforaminal lumbar interbody fusion: a single-center experience. J Neurosurg Spine. 2012 Jan;16(1):44-50. doi: 10.3171/2011.9.SPINE11373. Epub 2011 Oct 14. PMID: 21999389.
20. Kurra S, Lavelle WF, Silverstein MP, Savage JW, Orr RD. Long-term outcomes of transforaminal lumbar interbody fusion in patients with spinal stenosis and degenerative scoliosis. Spine J. 2018 Jun;18(6):1014-1021. doi: 10.1016/j.spinee.2017.10.063. Epub 2017 Nov 22. PMID: 29174460.
21. Mehren C, Ostendorff N, Schmeiser G, Papavero L, Kothe R. Do TLIF and PLIF Techniques Differ in Perioperative Complications? - Comparison of Complications Rates of Two High Volume Centers. Global Spine J. 2025 Jan;15(1):84-93. doi: 10.1177/21925682241248095. Epub 2024 Apr 17. PMID: 38631328; PMCID: PMC11572157.
22. Huang J, Rabin EE, Stricsek GP, Swong KN. Outcomes and complications of minimally invasive transforaminal lumbar interbody fusion in the elderly: a systematic review. J Neurosurg Spine. 2022;36(1):1-11. doi:10.3171/2021.8.SPINE21725
23. Lee CH, Hyun SJ, Kim KJ, Jahng TA, Kim HJ. Single-level posterior lumbar interbody fusion using cortical bone trajectory screws: 2-year radiologic and clinical results. J Korean Neurosurg Soc. 2018;61(5):593-602. doi:10.3340/jkns.2018.0068
24. Gadjradj PS, Harhangi BS, van Tulder MW, et al. Surgical versus non-surgical treatment for lumbar spinal stenosis: a systematic review and meta-analysis. Eur Spine J. 2021;30(8):2176-2190. doi:10.1007/s00586-021-06897-2
25. Kwon BK, Vaccaro AR, Grauer JN, Beiner JM, Hilibrand AS. Indications, techniques, and outcomes of minimally invasive lumbar fusion. J Am Acad Orthop Surg. 2005;13(4):294-303. doi:10.5435/00124635-200507000-00006
26. Goyal A, Elminawy M, Kerezoudis P, et al. Outcomes following minimally invasive versus open transforaminal lumbar interbody fusion: a systematic review and meta-analysis. World Neurosurg. 2019;123:501-515.e2. doi:10.1016/j.wneu.2018.11.217
27. Phan K, Mobbs RJ. Minimally invasive versus open laminectomy for lumbar stenosis: a systematic review and meta-analysis. Spine (Phila Pa 1976). 2016;41(2):E91-E100. doi:10.1097/BRS.0000000000001172
28. Parker SL, Mendenhall SK, Shau DN, et al. Minimally invasive versus open transforaminal lumbar interbody fusion: economic analysis. Spine (Phila Pa 1976). 2011;36(13):E927-E932. doi:10.1097/BRS.0b013e3181e8e3c7
29. Gornet MF, Copay AG, Bond MA, et al. Clinical outcomes after minimally invasive transforaminal lumbar interbody fusion: a prospective, multicenter study. Spine (Phila Pa 1976). 2016;41(7):556-565. doi:10.1097/BRS.0000000000001264
30. Mobbs RJ, Phan K, Malham G, et al. Lumbar interbody fusion: techniques, indications and comparison of interbody fusion options including PLIF, TLIF, MITLIF, OLIF/ATP, LLIF and ALIF. J Spine Surg. 2015;1(1):2-18. doi:10.3978/j.issn.2414-469X.2015.10.05
31. Fairbank JC, Pynsent PB. The Oswestry Disability Index. Spine (Phila Pa 1976). 2000;25(22):2940-2952. doi:10.1097/00007632-200011150-00017

32. Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36): I. Conceptual framework and item selection. Med Care. 1992;30(6):473-483. doi:10.1097/00005650-199206000-00002
33. Lee N, Kim KN, Yi S, et al. Clinical and radiological outcomes of revision surgery after failed lumbar fusion. J Korean Neurosurg Soc. 2021;64(2):258-265. doi:10.3340/jkns.2020.0196
34. Kim JS, Phan K, Cheung ZB, et al. Predictive factors for successful outcomes of minimally invasive transforaminal lumbar interbody fusion. Clin Spine Surg. 2018;31(7):E343-E349. doi:10.1097/BSD.0000000000000688
35. Glassman SD, Carreon LY, Djurasovic M, et al. Lumbar fusion outcomes stratified by specific diagnostic indication. Spine J. 2009;9(1):13-21. doi:10.1016/j.spinee.2008.08.007
36. Deyo RA, Mirza SK, Martin BI, et al. Trends, major medical complications, and charges associated with surgery for lumbar spinal stenosis in older adults. JAMA. 2010;303(13):1259-1265. doi:10.1001/jama.2010.338
37. Kim MC, Chung HT, Kim DJ, et al. Accuracy and safety of pedicle screw placement in the thoracic and lumbar spine using a patient-specific template guide. J Orthop Surg Res. 2019;14(1):332. doi:10.1186/s13018-019-1392-5
38. Staartjes VE, de Wispelaere MP, Vandertop WP, Schröder ML. Deep learning-based preoperative predictive analytics for patient-reported outcomes following lumbar discectomy: feasibility of center-specific modeling. Spine J. 2019;19(5):853-861. doi:10.1016/j.spinee.2018.11.006
39. Staartjes VE, de Wispelaere MP, Schröder ML. Machine learning-based prediction of patient-reported outcome measures after lumbar discectomy: feasibility of center-specific modeling. Spine J. 2019;19(5):853-861. doi:10.1016/j.spinee.2018.11.006
40. Modic MT, Ross JS. Lumbar degenerative disk disease. Radiology. 2007;245(1):43-61. doi:10.1148/radiol.2451051706
41. Kim YJ, Lenke LG, Kim J, et al. Accuracy and safety of pedicle screw placement in the thoracic and lumbar spine using O-arm-based navigation. Eur Spine J. 2017;26(3):664-670. doi:10.1007/s00586-016-4841-3
42. Staartjes VE, de Wispelaere MP, Vandertop WP, Schröder ML. Deep learning-based preoperative predictive analytics for patient-reported outcomes following lumbar discectomy: feasibility of center-specific modeling. Spine J. 2019;19(5):853-861. doi:10.1016/j.spinee.2018.11.006
43. Staartjes VE, de Wispelaere MP, Schröder ML. Machine learning-based prediction of patient-reported outcome measures after lumbar discectomy: feasibility of center-specific modeling. Spine J. 2019;19(5):853-861. doi:10.1016/j.spinee.2018.11.006
44. Aaronson NK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst. 1993;85(5):365-376. doi:10.1093/jnci/85.5.365
45. American Medical Association. Code of Medical Ethics Opinion 9.2.2. 2023.
46. Brooke J. SUS: A "quick and dirty" usability scale. In: Jordan PW, Thomas B, Weerdmeester BA, McClelland IL, editors. Usability Evaluation in Industry. London: Taylor & Francis; 1996.
47. U.S. Department of Health and Human Services. HIPAA Privacy Rule. 45 CFR Part 164.
48. European Union. Regulation (EU) 2016/679 (General Data Protection Regulation).
49. National Institute of Standards and Technology. Protecting Controlled Unclassified Information in Nonfederal Systems and Organizations (NIST SP 800-171).
50. California Medical Injury Compensation Reform Act (MICRA).
51. California Health and Safety Code §24173.
52. World Health Organization. Ethics and Governance of Artificial Intelligence for Health. 2021.
53. FUTURE-AI Consortium. FUTURE-AI Guidelines. 2024.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.