Supplementary Materials

Navigating artificial intelligence in spine surgery: implementation and optimization across the care continuum

Antony A. Fuleihan[1,2], Arjun K. Menta[1], Tej D. Azad[1], Kelly Jiang[1], Carly Weber-Levine[1], A. Daniel Davidar[1], Andrew M. Hersh[1], Nicholas Theodore[1]

[1]Department of Neurosurgery, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA.

[2]Sidney Kimmel Medical College, Thomas Jefferson University, Philadelphia, PA 19107, USA.

Correspondence to: Dr. Nicholas Theodore, Department of Neurosurgery, Johns Hopkins University School of Medicine, 600 N. Wolfe St., Meyer 7-113, Baltimore, MD 21287, USA. E-mail: theodore@jhmi.edu

Supplementary Table 1. Overview of referenced artificial intelligence studies

Title	Dataset	Key AI Results	Main Findings	Algorithms
Automated Detection of Postoperative Surgical Site Infections Using Supervised Methods with Electronic Health Record Data	6258 procedures from the National Surgical Quality Improvement Project Database were analyzed.	The generated models had high specificity (ranging from 0.788 to 0.988) and high negative predictive values (>0.98).	The automated detection models developed in the study were able to reliably eliminate the majority of patients who did not have a surgical site infection (SSI), thereby reducing the workload of the surgical clinical reviewers. The factors the models used to detect SSIs, such as diagnosis codes, antibiotic use, and lab value changes, are consistent with clinical knowledge about SSI detection. Patterns in longitudinal lab results, such as increases in platelet count and glucose after surgery, were useful indicators for detecting SSIs.	1. Multivariate logistic regression models, including one model for total SSI and one model for each of the three SSI subtypes (superficial, deep, and organ/space). 2. Other machine learning algorithms like Random Forest and Support Vector Machine, which were found to be outperformed by the logistic regression models for detecting postoperative SSI events.
MANDY: Towards A Smart Primary Care Chatbot Application	~20,000 web pages from Wikipedia and ~10,000 web pages from Healthline.	The question accuracy of Mandy's generated questions ranged from 67% to 100% across 6 test cases, with an average of 92%. The prediction accuracy of Mandy's diagnosis module ranged from 33% to 100% across 11 test cases, with an average of 72%.	Mandy is an AI assistant designed to help doctors, not to replace them in making diagnoses. The paper used word2vec, a natural language processing technique, to understand patient symptoms, and this approach worked well based on their evaluation. The current proof-of-concept has limitations and needs further development, such as expanding the number of diseases it can handle.	1. Word2vec for natural language processing and symptom extraction (Analysis Engine module) 2. Positive-Negative Matching Feature Count (P-N)MFC algorithm for mapping symptoms to hypothesized diseases (S2C Mapping module) 3. Question generation algorithm to ask patients follow-up questions based on the hypothesized diseases (Question Generator module)
Validity of AI-Based Gait Analysis for Simultaneous Measurement of Bilateral Lower Limb Kinematics Using a Single Video Camera	The dataset used consists of gait data collected from 21 healthy young participants, including 10 males and 11 females, with an average age of 20.7 ± 1.0 years.	The mean absolute errors for the limb on the camera side ranged from 2.3 to 3.1 degrees, while for the limb on the opposite side they ranged from 3.1 to 4.1 degrees, indicating acceptable accuracy on both sides. The coefficient of multiple correlation values, which measure waveform pattern similarity, ranged from 0.936 to 0.994 on the camera side and from 0.890 to 0.988 on the opposite side, indicating very good to excellent similarity.	Gait analysis using a single video camera and AI-based pose estimation had acceptable accuracy (mean absolute errors less than 5 degrees) and very good to excellent waveform similarity compared to 3D motion analysis. Measurement accuracy was slightly superior on the camera side compared to the opposite side. The precision on both sides was sufficiently robust for clinical evaluation.	1. OpenPose version 1.7.0 - a deep learning-based pose estimation algorithm used to estimate the positions of joint centers in the video footage. 2. Lens distortion correction algorithm - used to eliminate the impact of lens distortion in the video footage by calibrating the video camera using the 3D motion analysis system and a T-shaped wand with optical markers. 3. Fourth-order Butterworth low-pass filter with a cutoff frequency of 6 Hz - used to filter the trajectories of the joint positions.
Bridging The Literacy Gap For Surgical Consents: An AI-Human Expert Collaborative Approach	The dataset used in this study consists of surgical consent forms from 15 large academic medical centers across the United States, including both publicly owned and private institutions.	The LLM-based simplification significantly reduced the length and reading time of the consent forms (P <0.05). The consent forms were originally written at a college-level reading difficulty, but the LLM-based simplification reduced the reading level to an 8th grade level (P=0.004). The researchers used GPT-4 to generate procedure-specific consent forms for 5 different surgical procedures, and these forms had an average 6th grade reading level.	GPT-4 was able to significantly improve the readability of generic surgical consent forms, reducing the reading level from college freshman to 8th grade level. GPT-4 was able to generate de novo procedure-specific consent forms that met expert-level scrutiny and scored perfectly on a validated consent form quality rubric. The study demonstrates a novel AI-human expert collaborative framework to enhance surgical consent forms while ensuring content accuracy and legal sufficiency.	1. The GPT-4 language model, which was used to simplify the surgical consent forms. 2. Readability metrics such as Flesch-Kincaid Reading Level and Flesch Reading Ease score, which were used to quantify the readability of the consent forms before and after simplification by GPT-4. 3. An 8-item rubric developed by Spatz et al. to evaluate the quality and comprehensiveness of the procedure-specific consent forms generated by GPT-4.
Automation Of Surgical Skill Assessment Using a Three-Stage Machine Learning	The dataset used in this study consists of 242 laparoscopic cholecystectomy videos, from which 101 video segments containing 13,823 frames	The instrument detection and localization model achieved 78% average precision and 82% average recall for the clipper instrument on the test set. The linear regression model was able to predict good	The three-stage machine learning algorithm achieved an accuracy of 87 ± 0.2% in distinguishing good versus poor surgical skill. The algorithm had limitations in predicting the exact surgical skill level, with an accuracy of 70 ± 0.2% in predicting the skill level within ±1 point. The authors	1. A Convolutional Neural Network (CNN) for detecting and localizing surgical instruments (graspers and clippers) in laparoscopic cholecystectomy videos.

Algorithm	were randomly selected and partitioned into training, validation, and test sets.	vs poor surgical skill with 87% accuracy, and predict the exact skill level within ±1 point with 70% accuracy.	note that further refinement of the algorithm and a larger training database are required to improve automated surgical skill assessment.	2. Extraction of motion features from the detected instrument locations over time, including metrics like centroid position, movement range, direction changes, and position changes. 3. A linear regression model trained on the extracted motion features to predict surgical skill.
Prediction Model For Outcome After Low-Back Surgery: Individualized Likelihood of Complication, Hospital Readmission, Return To Work, and 12-Month Improvement In Functional Disability	The dataset used in this study is a prospective, longitudinal spine registry from Vanderbilt University Medical Center, which includes clinical data collected over 4 years from 1,803 consecutive patients undergoing lumbar spine surgery.	The multivariate prediction model demonstrated an R-squared of 0.51 and 0.47 for the development and validation study, respectively. The AUC values of the different predictive models ranged from 0.72-0.84 and 0.79-0.84 for the development and validation study, respectively.	The study developed and validated novel multivariate prediction models to estimate individual patients' unique risks and benefits of elective lumbar spine surgery for degenerative spine disease. The models were able to predict 12-month Oswestry Disability Index (ODI) scores, as well as the probability of complications, hospital readmission, need for inpatient rehabilitation, and return to work.	1. Linear regression with Bayesian model averaging to model the 12-month Oswestry Disability Index (ODI) outcome 2. Multiple logistic regression with Bayesian model averaging to model the categorical outcomes of complications, readmission, inpatient rehabilitation, return to work, and a composite measure of unplanned outcome 3. An 80/20 split of the data for model development and validation, respectively
Automated prediction of the Thoracolumbar Injury Classification and Severity Score from CT using a novel deep learning algorithm	The dataset used in this study consists of 111 patients with traumatic spine injuries or degenerative spine pathologies who underwent neurosurgical consultations at a single tertiary center from January 2018 to December 2019. The dataset includes CT scans, injury classifications, and demographic information for these patients.	The morphology network demonstrated 95.1% accuracy in binary classification (injured vs. non-injured) and 86.3% accuracy in TLICS morphology classification. The PLC network demonstrated 86.8% accuracy in classifying PLC integrity (no injury vs. suspected injury) and a 90.0% true positive rate.	The deep learning algorithm was able to predict the TLICS morphology with 86.3% accuracy and the PLC integrity with 86.8% accuracy. The algorithm can provide a TLICS score per vertebral level, which can aid in clinical decision-making. The algorithm has the potential to reduce the need for expensive MRI and provide physicians a risk assessment level regardless of their expertise in calculating TLICS.	1. Faster R-CNN, a state-of-the-art object detection region-based convolutional neural network 2. Two separate neural networks, one for classifying vertebrae into TLICS morphology classes and another for classifying vertebrae into binary PLC integrity scores 3. A 90%/10% train-test split to ensure no patient case appeared in both the training and testing sets
MySurgeryRisk: Development and Validation of a Machine-learning Risk Algorithm for Major Complications and Death After Surgery	The dataset consists of 51,457 adult patients who underwent major inpatient surgery requiring longer than 24 hours of inpatient admission at a quaternary-care academic center.	The algorithim presented with AUC values between 0.82 and 0.94 and accuracy between 0.74 and 0.86 for the 8 postoperative complications of interest. The model also predicts death with monthly intervals up to 24 months with AUC values between 0.77 and 0.83.	The study developed and validated an automated machine-learning algorithm called "MySurgeryRisk" that uses existing clinical data in electronic health records to forecast the risk for major complications and death after any type of surgery with high sensitivity and specificity. The algorithm can calculate probabilistic risk scores for 8 postoperative complications with AUC values ranging between 0.82 and 0.94, and predict the risk for death at 1, 3, 6, 12, and 24 months with AUC values ranging between 0.77 and 0.83. The algorithm has the potential to be implemented in real-time clinical workflow to improve preoperative risk assessment and guide patient management.	1. Algorithms to define and identify postoperative complications, including AKI, MV, and ICU admission. 2. A generalized additive model (GAM) to calculate patient-level risk scores for each postoperative complication. 3. A random forests classifier to calculate patient-level mortality scores at different time points after surgery. 4. Data preprocessing algorithms to handle missing data and outliers. 5. Feature transformation algorithms to reduce dimensionality and handle categorical variables with multiple levels.
Development And Validation of Risk Stratification Models for Adult	The dataset includes 1612 ASD patients treated surgically by 57 surgeons at 23 sites across 5 countries (17	Kaplan-Meier estimates showed that 12.1% of patients had at least one major complication within 10 days of surgery, 21.5% within 90 days, and 36% within 2	The study created accurate predictive models for major complications, hospital readmissions, and unplanned reoperations following adult spinal deformity surgery. The models showed a wide range of individual risk profiles,	The algorithm used in this study was the random survival forest algorithm, which was used to create predictive models for major complications, hospital readmission, and unplanned reoperation

Spinal Deformity Surgery	sites in the US, 2 in Spain, 2 in Turkey, 1 in France, and 1 in Switzerland).	years. The predictive models demonstrated a concordance statistic of up to 71.7% in the development sample.	with cumulative risk estimates at 2 years ranging from 3.9% to 74.1% for major complications, 3.17% to 44.2% for readmissions, and 2.67% to 51.9% for reoperations. The most important predictors of complications were surgical invasiveness, age, magnitude of deformity, and patient frailty.	following adult spinal deformity surgery.
Analyzing Surgical Technique In Diverse Open Surgical Videos With Multitask Machine Learning	The dataset used in this study is a sample of 68 patients undergoing elective noncardiac surgery, primarily oncologic gastrointestinal procedures, at a single tertiary academic center in the Netherlands. The patients were randomized to either an intervention group using a machine learning-derived early warning system or a control group receiving standard care.	The AVOS data set comprised 1,997 open surgical videos, annotated at a 1-second resolution, enabling a multitask neural network to achieve a mean precision of 0.71 and mean recall of 0.73 for action recognition, and mean average precisions of 0.89 for hand detection and 0.46 for tool detection. The neural network successfully identified unique surgical signatures for procedures like appendectomy, pilonidal cystectomy, and thyroidectomy, allowing for the characterization of surgical skill through metrics such as hand motion and pose, which revealed that experienced surgeons exhibit more localized movements compared to trainees. In a proof-of-concept study, the model demonstrated a significant correlation between localized hand movements and higher surgical skill levels, with an odds ratio of 3.6 (95% CI, 1.67-7.62; P = .001).	The study presented a multitask neural network model that can analyze open surgical videos and introduced a large, diverse data set of open surgical videos used to train the model. The model was able to generate procedure-specific signatures and identify kinematic elements of surgical skill from prospectively collected videos. The model provided artificial intelligence-deduced insights into open surgical technique.	1. A multitask neural network model for simultaneous, spatiotemporal analysis of hands, tools, and actions in open surgical videos 2. An alternating task training strategy to optimize the spatial and temporal model branches 3. Principal component analysis to create a single compound skill feature from kinematic metrics like hand translation and pose change 4. Logistic regression to predict surgeon skill level (experienced vs trainee) based on the compound skill feature
Effect Of A Machine Learning-Derived Early Warning System For Intraoperative Hypotension Vs Standard Care On Depth And Duration Of Intraoperative Hypotension During Elective Noncardiac Surgery: The HYPE Randomized Clinical Trial	The dataset used in this study is a sample of 68 patients undergoing elective noncardiac surgery, primarily oncologic gastrointestinal procedures, at a single tertiary academic center in the Netherlands. The patients were randomized to either an intervention group using a machine learning-derived early warning system or a control group receiving standard care.	In the preliminary RCT, the AI-driven intervention resulted in a significantly lower median time-weighted average of hypotension at 0.10 mm Hg compared to 0.44 mm Hg in the control group (P = .001). The intervention group experienced fewer hypotensive episodes (3.00 per patient) compared to the control group (8.00 per patient, P = .004) and had a substantially reduced total duration of hypotension, averaging 8.00 minutes versus 32.67 minutes in the control group (P < .001). Additionally, the AI model facilitated faster treatment initiation, with a median time from alarm to first treatment of 53 seconds in the intervention group, compared to 87 seconds in the control group (P < .001), demonstrating its effectiveness in enhancing clinical responsiveness.	The use of a machine learning-derived early warning system for intraoperative hypotension, compared to standard care, resulted in a significantly lower time-weighted average of hypotension (0.10 mm Hg vs 0.44 mm Hg). The total time spent in hypotension was significantly lower in the intervention group compared to the control group (8.00 minutes vs 32.67 minutes). There were no serious adverse events resulting in death in the intervention group, compared to 2 (7%) in the control group.	1. The Hypotension Prediction Index (HPI), which is a machine learning-derived early warning system that can predict hypotension before it occurs. This algorithm was developed and validated in previous studies. 2. The early warning system, which uses 23 variables extracted from the arterial pressure waveform to detect deteriorations that could lead to hypotension. 3. A hemodynamic diagnostic guidance and treatment protocol that was developed to help anesthesiologists interpret the variables provided by the early warning system and determine the appropriate treatment.
Development and Validation of an Artificial Intelligence-Based	The dataset used in this study consisted of 9 participants with chronic spinal cord injury (SCI) for over 6	During the 8-week intervention, the AI-assisted group (n = 4) showed significant improvements in strength across all tested items, with increases of over 40% in metrics	The AI-based motion analysis system led to increased muscle strength in the experimental group compared to the control group. Participants and the instructor found the system easy to use but identified areas for improvement,	1. The MediaPipe algorithm developed by Google to detect 33 human joint points and provide real-time motion analysis and visual feedback. 2. Algorithms to count the repetitions of key upper

Motion Analysis System for Upper Extremity Rehabilitation Exercises in Patients with Spinal Cord Injury: A Randomized Controlled Trial	months, aged 49-80 years, who were able to independently perform upper extremity exercises. The participants were randomly assigned to an Experimental Group (EG, n=4) or a Control Group (CG, n=5), and participated in 24 exercise sessions over 8 weeks.	such as the chest press and arm curl, while the control group (n = 5) exhibited minimal changes or declines in most measures. The AI-assisted approach was associated with positive user feedback on ease of use and real-time feedback. Usability scores varied, with participants finding the program easy to use but desiring enhancements in program diversity and feedback.	such as increasing the variety of exercise programs and improving system stability for independent home use.	extremity exercises like the chest press, shoulder press, and arm curl based on the detected joint positions and orientations.
Unsupervised Machine Learning on Motion Capture Data Uncovers Movement Strategies in Low Back Pain	The dataset used in this study consists of biomechanical data collected from 111 participants across three cohorts: 43 individuals with non-specific low back pain, 42 individuals with adult spinal deformity, and 26 healthy controls. The biomechanical data was collected using a marker-less motion capture system (Microsoft Kinect) while the participants performed sit-to-stand movements. The dataset includes a comprehensive set of kinematic, kinetic, and dynamic variables derived from the joint angles, velocities, accelerations, torques, and powers.	The nonlinear principal component analysis (NLPCA) identified two primary movement strategies, with the first principal component (PC1) accounting for 36.8% of variance and differentiating the standing-to-sit (STS) strategy of adult spinal deformity patients from controls (F(2, 108) = 7.55, p = 0.0008). The second principal component (PC2), accounting for 19.3% of variance, revealed a more laborious leaning stand-up strategy in patient groups compared to controls (F(2, 108) = 30.091, p < 0.0001), suggesting that higher PC1 scores correlate with reduced pain and disability, while PC2 showed no significant associations with patient-reported outcomes.	Unsupervised machine learning on full body biomechanics during sit-to-stand movements identified divergent movement strategies between low back pain patients and healthy controls. The first principal component (PC1) captured a "vertical rise" STS strategy in controls versus a more forward-leaning strategy in low back pain patients. The second principal component (PC2) captured a "leaning forward-swing back" STS strategy that was more characteristic of low back pain patients and resulted in higher spinal loading.	1. Nonlinear principal component analysis (NLPCA) to analyze the biomechanical data and identify divergent movement strategies between low back pain patients and healthy controls. 2. Permutation testing to determine the number of significant principal components to retain for further analysis. 3. Bootstrapping to assess the stability of the NLPCA solution.
Examining the Ability of Artificial Neural Networks Machine Learning Models to Accurately Predict Complications Following Posterior Lumbar Spine Fusion	The dataset used in this study is the American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP) database, which is a prospective, risk-adjusted, multicenter quality improvement program that collects over 135 preoperative, intraoperative, and 30-day postoperative outcomes from over 258 participating hospitals in the United States. The authors queried this database for patients who underwent posterior lumbar spine fusion between 2010 and 2014, and	The artificial neural networks (ANN) and logistic regression (LR) models outperformed the ASA benchmark for predicting surgical complications, with ANN achieving the highest AUC of 0.85 for cardiac complications and LR achieving an AUC of 0.83 for VTE complications. For wound complications, LR had an AUC of 0.81, while ANN also showed strong performance with an AUC of 0.80 for mortality prediction. ANN demonstrated greater sensitivity than LR for predicting mortality (sensitivity: 78% vs. 72%) and wound complications (sensitivity: 75% vs. 70%), while LR exhibited higher specificity for these complications, indicating that ANN may be more effective for clinical applications due to its ability to handle class imbalances and detect nonlinear patterns in	Both ANN and LR models outperformed the benchmark ASA class for predicting complications following posterior lumbar spine fusion. ANN had the best performance for predicting cardiac complications, while LR had the best performance for predicting VTE, wound complications, and mortality. ANN was more sensitive than LR for detecting mortality and wound complications.	1. Artificial Neural Networks (ANNs) 2. Logistic Regression (LR)

	analyzed a total of 22,629 (70% training, 30% testing) patients after excluding those with missing data.	the data.		
Artificial Intelligent Virtual Assistant for Plastic Surgery Patient's Frequently Asked Questions A Pilot Study	The dataset used in this study used 30 participants who posed 10 common preoperative questions relevant to plastic surgery, with each question linked to standardized answers developed from 3 example variations.	In a survey of 30 healthcare administrative participants, 70% were women, with an average age of 27.76 years. The AI virtual assistant (AIVA) provided accurate standard answers 92.3% of the time, although participants perceived these answers as correct only 83.3% of the time, with lower accuracy noted for topics like nausea (60.0%) and pain (66.7%). While participants found the AIVA easy to use (100% agreed), there were mixed feelings about its potential to replace human assistants, with 40% neutral on the matter.	The AIVA was able to accurately answer 92.3% of the questions, though participants only perceived 83.3% of the answers as correct. The AIVA performed particularly well in understanding and accurately answering questions on certain topics like nausea, recovery time, suture, drain, and scar. Participants generally had a positive perception of the AIVA, finding it easy to use and potentially helpful for patients, but they were neutral on whether it could replace a human assistant.	A chatbot trained through the IBM Watson Assistant (International Business Machines Corp)
Artificial Intelligence Versus Surgeon Gestalt In Predicting Risk Of Emergency General Surgery	The dataset used in this study consists of 150 adult patients who underwent emergency laparotomy at an academic tertiary referral medical center between May 2018 and May 2019. The dataset includes information on the patients' demographics, clinical characteristics, and 30-day postoperative outcomes such as mortality, septic shock, ventilator dependence, bleeding, and pneumonia.	The AI model, POTTER, demonstrated superior predictive performance for postoperative outcomes, with an area under the curve (AUC) of 0.88 for mortality and 0.92 for ventilator dependence. Comparatively, the combined estimates from surgeons interacting with POTTER (SURG-POTTER) had AUCs of 0.87 for mortality and 0.83 for ventilator dependence, while traditional surgeon estimates (SURG) scored lower (AUC of 0.84 for mortality and 0.83 for ventilator dependence). Overall, POTTER consistently outperformed both SURG and SURG-POTTER across various outcomes, including bleeding and pneumonia.	The AI-based POTTER algorithm outperformed surgeons' gestalt in predicting most postoperative outcomes, including mortality, ventilator dependence, bleeding, and pneumonia. When surgeons used POTTER, it improved their own risk prediction abilities compared to when they did not have access to POTTER. AI-based tools like POTTER can be a useful adjunct to surgeons in preoperative risk assessment and counseling.	1. Predictive Optimal Trees in Emergency Surgery Risk (POTTER) - a nonlinear AI-based risk calculator that uses Optimal Classification Trees (OCTs) to predict postoperative outcomes for emergency general surgery patients. 2. Optimal Classification Trees (OCTs) - the machine learning algorithm used by POTTER to account for nonlinear interactions between variables and provide highly accurate risk estimates.
Artificial Intelligence Based Hierarchical Clustering of Patient Types and Intervention Categories in Adult Spinal Deformity Surgery Towards a New Classification Scheme that Predicts Quality and Value	The dataset used in the study was a merged dataset from two independent and compatible prospective multicenter ASD databases, one from the United States and the other from Europe. The dataset included 570 patients who met the inclusion criteria of being 18 years or older, having radiographically-confirmed ASD, planning for surgical treatment, and having complete data at baseline and 2-year follow-up.	The AI algorithm applied hierarchical clustering to patient parameters, identifying three distinct patient clusters with significant differences in demographics and outcomes. Patients in the Old Revision cluster, for example, experienced the poorest baseline scores and the highest complication rates, but demonstrated the greatest improvement in PROM scores following more complex surgeries, such as three-column osteotomies. The analysis highlights the AI's ability to segment patients effectively, providing insights into the risk-to-benefit ratio for different surgical interventions across homogeneous patient groups.	Unsupervised hierarchical clustering identified three optimal patient types: young with coronal plane deformity, older with prior spine surgeries, and older without prior spine surgeries. Hierarchical clustering also identified four surgical clusters: 3-column osteotomies, Smith-Peterson osteotomies, no osteotomy/no interbody fusion, and interbody fusion. The intersection of patient-based and surgery-based clusters yielded 12 subgroups with varying rates of major complications (0-51.8%) and 2-year normalized PROM improvements (-0.1% to 100.2%).	The main algorithm used in this study was unsupervised hierarchical clustering, specifically using Ward distances and the gap method to optimize the clustering.
Using ChatGPT to	The dataset used in this study	A simplified surgical consent form	The authors used GPT-4, a large language model, to	The main algorithm used in this study was

Facilitate Truly Informed Medical Consent	was the existing surgical consent form used by the Lifespan Health System in Rhode Island. The researchers fed this form into the GPT-4 language model and asked it to simplify the content while preserving the meaning, in order to convert it to the average American reading level.	generated by GPT-4 was approved with one modification to include "sleep medicine" alongside "anesthesia." Concerns about biases, approval authority, and revising other documents were addressed, and the form was deployed in 2023.	simplify a surgical consent form and reduce its reading level from 12.6 (college freshman level) to 6.7 (7th grade level). The simplified consent form was reviewed and approved by various institutional stakeholders, with only a minor modification. The authors highlight that using AI to simplify medical documents can help improve patient comprehension and achieve truly informed consent.	Generative Pretrained Transformer 4 (GPT-4), a large language model developed by OpenAI.
Development of machine learning and natural language processing algorithms for preoperative prediction and automated identification of intraoperative vascular injury in anterior lumbar spine surgery	The dataset used in this study consisted of adult patients 18 years or older who underwent anterior lumbar spine surgery. The data was obtained through a retrospective review of electronic health records and was split into a training set of 786 (75.9%) patients who underwent surgery before 2014, and an independent testing set of 249 (24.1%) patients who underwent surgery after 2014.	In a cohort of 1,035 anterior lumbar spine surgery patients, the rate of intraoperative vascular injury was 7.2%. The NLP algorithm for detecting vascular injury achieved an AUC-ROC of 0.92 and significantly outperformed CPT and ICD codes, which had an AUC-ROC of 0.64. At the Youden index threshold, the NLP algorithm identified 86% of vascular injury cases, compared to 29% using procedural and diagnostic codes.	The NLP algorithm was able to identify a much higher proportion of intraoperative vascular injuries compared to administrative codes (CPT and ICD). The NLP algorithm achieved good performance on temporal validation, with an AUC-ROC of 0.92 and AUC-PRC of 0.74. The preoperative prediction algorithm using elastic-net penalized logistic regression achieved fair discrimination (AUC-ROC of 0.73) but poor calibration for higher predicted probabilities.	1. An XGBoost supervised machine learning NLP algorithm for automated detection of intraoperative vascular injury from free-text operative notes. 2. Five supervised machine learning algorithms for preoperative prediction of vascular injury: stochastic gradient boosting, random forest, support vector machine, neural network, and elastic-net penalized logistic regression. 3. The best performing preoperative prediction algorithm (elastic-net penalized logistic regression) was deployed as an open-access digital application.
Fully Automatic Cervical Vertebrae Segmentation Framework For X-Ray Images	The dataset used in this study consists of 296 (90% training, 10% test) lateral cervical spine X-ray images collected from the Royal Devon and Exeter Hospital in association with the University of Exeter. The images have varying patient ages (17 to 96), imaging systems (Philips, Agfa, Kodak, GE), resolutions (0.1 to 0.194 mm per pixel), and sizes (1000 to 5000 pixels), and include examples of vertebrae with fractures, degenerative changes, and bone implants.	The algorithm achieved an average pixel-level accuracy of 99% on lower-resolution images and a sensitivity and specificity of 0.96 at the original resolution, covering 96% of the vertebral area compared to ground truth. The method demonstrated a 17.1% improvement in sensitivity over the previous state-of-the-art and was over 70 times faster, producing results in under a second per image.	The proposed fully automatic framework for segmenting cervical vertebrae in X-ray images achieved a Dice similarity coefficient of 0.84 and a shape error of 1.69 mm. The global localization algorithm outperformed the previous state-of-the-art by 17.1% in terms of sensitivity. The center localization framework achieved an average error of only 1.81 mm, which is near human-level performance.	1. A deep fully convolutional network (FCN) for global spine localization 2. A novel FCN-based probabilistic spatial regressor for vertebral center localization 3. A novel shape-aware loss function for vertebrae segmentation
Development of Machine Learning Algorithms for Prediction of Prolonged Opioid Prescription After Surgery for Lumbar Disc Herniation	The dataset used in this study is a retrospective chart review of 5,413 patients who underwent surgery for lumbar disc herniation between January 1, 2000 and March 1, 2018. The dataset was split into a training set (80%) and	Out of 5,413 patients undergoing lumbar disc herniation surgery, 416 (7.7%) had sustained opioid prescriptions at 90 to 180 days post-surgery. Predictors of prolonged opioid use included female sex, previous spine surgery, instrumentation, preoperative opioid use, and comorbid depression. In the holdout set (n=1,082), the elastic-net	Five machine learning models were developed to predict prolonged opioid prescription after lumbar disc herniation surgery, with the elastic-net penalized logistic regression model performing the best. The three most important predictors were instrumentation, duration of preoperative opioid prescription, and comorbidity of depression. The final model was made available as an open access web application to provide predictions and patient-specific	1. Elastic-net penalized logistic regression 2. Random forest 3. Stochastic gradient boosting 4. Neural network 5. Support vector machine The elastic-net penalized logistic regression model had the best performance in terms of discrimination, calibration, and overall

	an independent testing set (20%) for developing and evaluating the predictive models.	penalized logistic regression performed best, with an AUC of 0.81, Brier score of 0.064, and good calibration (slope=1.13). Instrumentation, preoperative opioid duration, and depression were the strongest predictors of sustained opioid use.	explanations.	performance.
AI-Powered Real-Time Annotations During Urologic Surgery: The Future of Training and Quality Metrics	The dataset used in this study is a large-scale dataset containing 7,768 (60/15/25 training/validation/test) surgical videos across multiple specialties, including Urology. The dataset includes surgical videos from partial/complete nephrectomy and radical prostatectomy procedures, and the videos were manually annotated by a group of specialists.	The AI platform identified key surgical annotations in RARP and RAPN procedures. For RARP, the system recognized 10 surgical phases, including Retzius space dissection and vesico-urethral anastomosis, and identified 7 specific surgical events such as ligation of vas deferens and specimen packaging. For RAPN, 6 surgical phases were noted, including lesion resection and site closure, alongside 7 events such as intraoperative ultrasound and notable hemorrhage.	The study demonstrates the first use of real-time AI annotation of surgical steps and safety milestones during two urologic robotic procedures. The AI system was able to identify key surgical steps, events, and safety milestones in real-time during the live procedures. This technology has the potential to provide real-time intraoperative decision support, improve surgical training and education, and enhance the quality of surgical care.	1. Video Transformer Network (VTN) technology, which uses a Vision Transformer (ViT) backbone and a bidirectional Gated Recurrent Unit (GRU) to process and annotate surgical videos in real-time and offline. 2. Multi-Task Learning (MTL) approach to train the model on a large dataset of over 7,700 surgical videos across multiple specialties, including urology.