Artificial Intelligence Assisted Surgical Coaching in Diverse Healthcare Contexts - A Multicentre Pilot Study

Introduction

The Lancet Commission on Global Surgery has highlighted a critical gap in the availability and quality of surgical training worldwide, noting that by 2030, an estimated five billion people will still lack access to safe, timely, and affordable surgical care. This disparity is particularly stark in low- and middle-income countries (LMICs), where shortages of skilled surgical providers and limited training opportunities persist. As health systems endeavor to meet these challenges, there is a clear need for innovative solutions that not only expand the surgeon workforce but also ensure that their operative skills are consistently refined and standardized.

Surgical education is undergoing a major change over the past years introducing video recording and data in-op and post-op. At the core of this transformation is the assumption that current technology can have a positive effect, when utilized correctly, on surgical education. Low Medium Income Countries (LMIC) are usually lacking both in recording technology and analytical modeling. Introduction of AI models into the surgical educational program can have a strong impact on surgical knowledge and execution, thus improving surgical outcomes.

Surgical coaching is one such innovation with significant potential. Drawing from educational theories and proven adult learning principles, surgical coaching goes beyond traditional apprenticeship models to foster a learning culture centered on feedback, reflection, and continuous improvement. Through structured, feedback-rich interactions, coaching encourages surgeons at all levels of training to enhance their technical proficiency, decision-making, and intraoperative confidence. By emphasizing openness, communication, and iterative skill refinement, coaching can help build a community of practice that nurtures excellence and peer support, ultimately improving both the educational yield for trainees and the quality of patient care.

However, despite its promise, surgical coaching faces notable barriers to implementation. Traditional coaching formats are often time- and labor-intensive, requiring substantial input from expert coaches who must dedicate significant portions of their schedules to reviewing recorded procedures and providing individualized feedback. This intensive human resource demand can limit scalability, especially in settings with already constrained professional capacity. Here, the application of artificial intelligence (AI) has emerged as a potential solution. AI-driven tools for video-based feedback analysis can streamline the coaching process by automating aspects of video review, identifying key surgical steps, and suggesting targeted improvements. Such tools can reduce coach workload, accelerate the feedback cycle, and democratize access to high-quality coaching across multiple centers and contexts.

A recent single-center pilot study in pituitary surgery has demonstrated the feasibility of employing video-based coaching as a mechanism for performance improvement and skill standardization. Pituitary surgery, which requires nuanced endoscopic techniques and delicate anatomical manipulations, is an ideal example for exploring the benefits of structured feedback and guidance. Insights gained from this pilot underscore the potential for expanding these methods to a multicenter, international scale. Utilizing AI computer vision models and an interactive application, the Surgical Video Platform (SVP), offers a novel solution for incorporating video-based analytics in surgical education programs

This project aims to adapt and implement AI-assisted coaching programs for use across multiple centers, including in LMICs, where the need for scalable surgical training solutions is greatest. By refining AI-driven feedback tools and validating their impact on learning curves, surgical competencies, and patient outcomes, this work seeks to advance both the science and practice of surgical coaching. Ultimately, the goal is to identify strategies that facilitate sustainable skill transfer, promote an educationally rich surgical culture, and improve global neurosurgical care—beginning with pituitary surgery as a model.

Methods

Study design

Prospective multicentre pilot cohort study with pre- and post-intervention data to evaluate the implementation of the coaching program, perceived local educational value and the comparative impact on surgical training and performance. This will be based on a pre-existing single centre experience (summarised via this video), but will be adapted to local healthcare contexts. Local ethical and governance approvals will be necessary.

Coaching programme intervention

Prior to study launch, each centre will undergo a series of meetings to tailor the coaching programme to their local training culture and needs, whilst maintaining the core coaching principles: deliberate review, post op feedback, and collaborative discussion. Similarly, a technological and human resources gap analysis will be performed using structured surveys (Gowda et al, unpublished) and necessary training or infrastructure set-up non-monetary support will be offered to enable the coaching programme. Access to the surgical video recording and uploading software will be offered in kind, and necessary hardware (e.g. hard drives and study laptops) will be provided through grant funding. The frequency and mode of delivery (e.g. face to face vs virtual) of each programme will be tailored to each context in order to maximise feasibility and participation (e.g. dovetailed with local departmental teaching programmes).

After necessary adaptations and training, each coaching programme will be launched and will involve: 1) consecutive pituitary surgery recording, 2) selection of videos for coaching session, 3) AI-assisted video curation and analysis, 4) execution of coaching session with recording of session outcomes, and 5) recording of potential clinical impacts of coaching programme.

Case selection will be at random from cases with surgical videos available from the preceding month as a default, unless local surgeons would like to discuss particular cases, which would then supersede the randomly selected video. Each monthly video will have anonymised clinical context added as free text (Figure 1a), and will then analysed by embedded AI models on the SDSC platform, which will index them for core surgical phrases such that they are easy to toggle, and will extract quantitative metrics of surgical workflow and instrument use (Figure 1b). Each indexed video will be shared via SDSC to the local pituitary surgery team, who will be asked to anonymously comment on the surgical strengths and learning points based on the video.

Coaching sessions will include pituitary surgery consultants and residents at a minimum but may include the wider surgical team. Sessions will be encouraged on a monthly basis, with one surgical case and video reviewed per month. If this is not feasible, then maintaining an average rate of one case per month will be encouraged (e.g. 2 cases discussed in a 2 monthly meeting). Each programme will run for 6 months before initial pilot analysis.

Each coaching session will consist of a presentation of case background and imaging (de-identified), followed by review of AI generated quantitative metrics (see below), qualitative feedback submitted by the team and then targeted team video review. During and after video review, semi-structured discussion of strengths and learning points will follow.

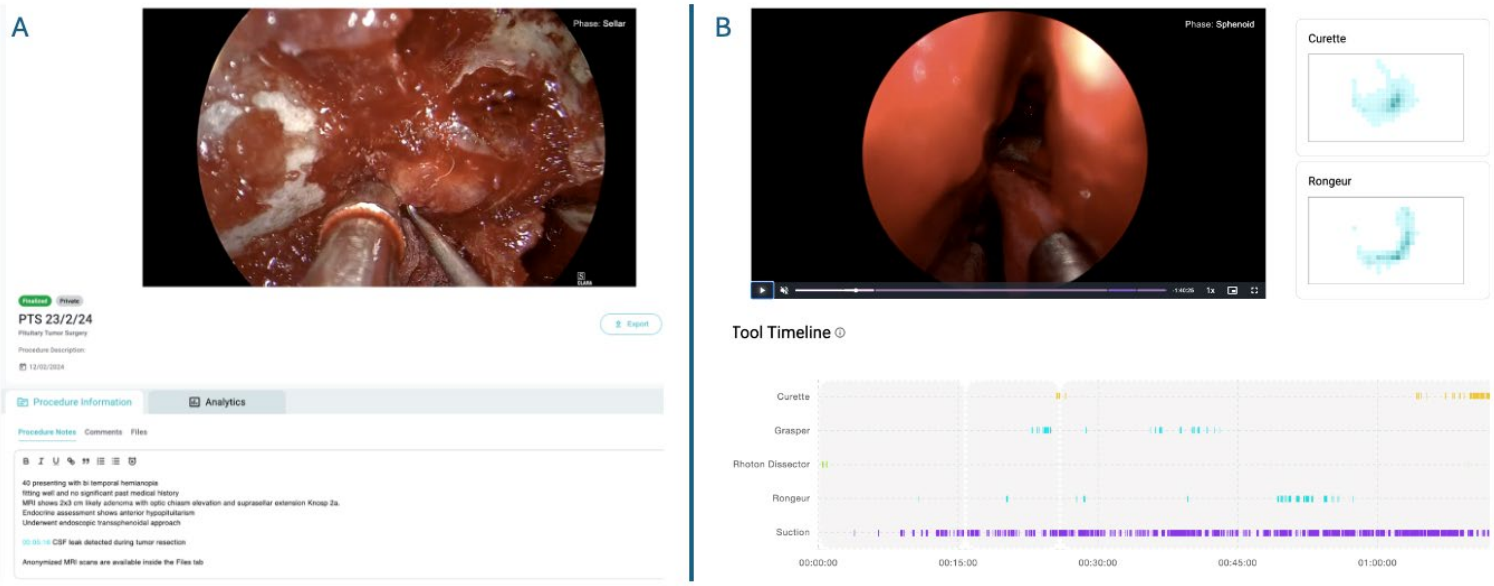

Figure 1: a) Example clinical context which can be added for each discussed case, with the ability to add relevant anonymised patient data (e.g. key images from MRI scans), and the ability to add surgical strengths and learning points as comments by users. B) Quantitative metrics available to supplement coaching discussions.

Centre selection

Partnerships will be developed with LMIC centres that meet the following criteria:

- Has >/=2 consultant neurosurgeons performing pituitary adenoma surgeries
- Performs >/= 1 transsphenoidal pituitary adenoma operations per month
- Are based in low- or middle-income countries according to World Bank classifiers
- Do not have an existing video-based surgical coaching programme

Partners will be approached via convenience sampling and will co-design an adapted version of the coaching programme to align with local contextual factors.

Data collection

Data collection will occur in 2 key phases. Control cohort phase (6 months) and intervention cohort phase (6 months). The former will occur after local approvals but prior to study launch (i.e. during study launch and set-up).

Baseline partner data will include hospital setting, catchment area, typical pituitary surgery volume, existing training structured initiatives, technology capacity and research capacity. Baseline clinical data will be collected via a prospective database, with metrics derived from previous multicenter studies on pituitary surgery outcomes. These included length of stay, hyponatremia, syndrome of inappropriate antidiuretic hormone, hypernatremia, diabetes insipidus, cerebrospinal fluid leak (biochemically confirmed and/or requiring operative intervention), postoperative visual deterioration, new suspected anterior pituitary hormone deficit requiring hydrocortisone supplementation on discharge (started if day 2 cortisol <350 nmol/L, except for patients with Cushing's disease), and new suspected posterior pituitary hormone deficit requiring desmopressin supplementation on discharge. Of note, only inpatient outcomes were harvested in this pilot study for practical purposes, therefore outcomes reliant on longer follow-up data (gross total resection, functional remission, etc.) were excluded.

The feasibility, appropriateness, and acceptability of the coaching program were assessed at the end of the study period. We defined feasibility as the extent to which a new treatment, or an innovation, can be successfully used or carried out within a given agency or setting – measured via the Feasibility of Intervention Measure (FIM). Acceptability was defined as the perception that a given intervention is agreeable, palatable, or satisfactory – measured via the Acceptability of Intervention Measure (AIM). Appropriateness was defined as the perceived fit, relevance, or compatibility of the innovation or evidence-based practice for a given practice setting, provider, or consumer; and/or perceived fit of the

innovation to address a particular issue or problem – measured via the Intervention Appropriateness Measure (IAM). These were assessed through a structured questionnaire with Likert scales at the end of the coaching program. Furthermore, the perceived educational yield of each session was collected using a structured qualitative questionnaire. Finally, user experience surveys on the SDSC platform will be gathered at the end of the programme to improve the platform.

The clinical impact of the program was assessed by comparing the 6-month coaching intervention period with a control cohort from the preceding 6 months (August 2022 to February 2023). For every case during the coaching and control, the surgical video was recorded, and surgical performance was scored using a modified Objective Structured Assessment of Technical Skills (mOSATS) scale by 3 blinded expert (consultant pituitary surgeons) assessors contracted via an independent 3rd party service.

AI model development

All videos from the control cohort will be annotated for surgical phases, steps and instruments in order to adapt existing supervised learning based neural networks hosted in SDSC for local contexts. Steps are goal-oriented sequences of actions (e.g. tumor resection) and are grouped into 3 overarching phases (nasosphenoid, sellar, closure) – guided by an international consensus-based framework. This annotation will be performed by the same 3rd party expert annotation service as above.

To ensure robust performance in the surgical setting, our model development will prioritize adaptation to diverse contexts. We will fine-tune a pre-trained convolutional neural network using a curated dataset of surgical videos representative of varying procedures, equipment, and patient anatomies. Transfer learning will leverage the model's existing image understanding capabilities while specializing it for surgical scene analysis. Validation will utilize context-specific metrics, including Dice coefficient for segmentation tasks and precision/recall for object detection, ensuring the model's accuracy and reliability within the target surgical environment.

During model deployment in the intervention cohort, AI model outputs of phase, step and instrument classifications will be reviewed and corrected as necessary by these expert data annotators. Downstream quantitative video metrics will be derived from these corrected annotations and presented at the monthly meeting for the candidate case. Pre and post correction data will be collected to diagnose and address AI correctable AI output errors.

Data analysis

Descriptive statistics of the FIM, AIM, and IAM questionnaires were generated using Excel (v16.8, Microsoft). Perceived educational yield data was qualitatively summarized using thematic analysis. Descriptive statistics were generated for the baseline characteristics, performance, and outcomes in the coaching and comparative cohorts using Excel.

Comparative statistics between coaching and comparative cohorts were performed using R statistical programming language, with packages *dplyr* for data preparation; glmmTMB for mixed-effect regressions; and *ggplot2*, *ggsci*, and patchwork for data visualization.

For analysis of the impact on surgical performance, mOSATS scores from each rater and for each sub-item were mean-averaged across each phase and for the operation as a whole. Beta regressions using logit link function were then conducted with the named attending and skull base fellow specified as crossed random effects. The scores were scaled using the Smithson and Verkuilen method to avoid values that are exactly 0 or 1 (a precondition for beta regression). The values were back-transformed for reporting.

For analysis of binary outcomes, logistic regressions were conducted with the logit link function. For analysis of length of stay, negative binomial regression with log link function was conducted.

Results & Anticipated outputs

- Iterative adaptation of coaching programs to local needs
- Iterative improvement of underpinning computer vision models & output metrics
- Foundational data to inform larger longer-term studies