npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01383-3

Human AI collaboration for unsupervised categorization of live surgical feedback

Rafal Kocielnik[1], Cherine H. Yang[2], Runzhuo Ma[3], Steven Y. Cen[4], Elyssa Y. Wong[5], Timothy N. Chu[2], J. Everett Knudsen[2], Peter Wager[2], John Heard[2], Umar Ghaffar[2], Anima Anandkumar[1] & Andrew J. Hung[2]

Formative verbal feedback during live surgery is essential for adjusting trainee behavior and accelerating skill acquisition. Despite its importance, understanding optimal feedback is challenging due to the difficulty of capturing and categorizing feedback at scale. We propose a Human-AI Collaborative Refinement Process that uses unsupervised machine learning (Topic Modeling) with human refinement to discover feedback categories from surgical transcripts. Our discovered categories are rated highly for clinical clarity and are relevant to practice, including topics like *"Handling and Positioning of (tissue)"* and *"(Tissue) Layer Depth Assessment and Correction [during tissue dissection]."* These AI-generated topics significantly enhance predictions of trainee behavioral change, providing insights beyond traditional manual categorization. For example, feedback on *"Handling Bleeding"* is linked to improved behavioral change. This work demonstrates the potential of AI to analyze surgical feedback at scale, informing better training guidelines and paving the way for automated feedback and cueing systems in surgery.

In recent years, it has become increasingly evident that formative verbal feedback to surgical trainees in the operating room (OR) plays a critical role in enhancing surgical education and outcomes[1]. High-quality feedback during surgical training is linked with better intraoperative performance[2], faster acquisition of technical skills[3], and greater trainee autonomy[4]. Feedback in the OR is intended to modify trainee behavior or thinking and, as depicted in Fig. 1, is generally triggered by a trainer's observation of a trainee's performance. Quantifying, understanding, and addressing the quality of such feedback is essential for improving both immediate surgical practices and long-term educational outcomes. However, effectively quantifying and enhancing the value of live surgical feedback remains a substantial challenge.

Despite the recognized benefits, existing literature lacks a universally accepted system for categorizing and evaluating this type of feedback effectively. Prior approaches have struggled with meaningful categorization, clinical validation, and scalability. They have also been limited by the substantial manual effort required from human annotators[1]. On top of that, no consensus exists on a method for categorizing and evaluating OR feedback to optimize its impact. Some approaches focused on categorizing teaching behaviors (e.g., informing, questioning, responding, or tone setting)[5] or tried to describe intraoperative communication (e.g., explaining, commanding, and questioning)[6]. Yet, these offer an incomplete view at best and have not been rigorously validated for clinical relevance. The latest work, which involved substantial manual effort from experts, developed broader feedback categorization into three core components (Anatomic, Procedural, Technical) and auxiliary delivery aspects (Praise, Criticism, and Visual Aid). This categorization has been shown to exhibit some association with behavioral outcomes[7] and the potential for automation to deliver feedback[8]. Yet, such broad categorization to date has provided limited value for the fine-grained understanding of feedback effectiveness[3] and is not sufficiently detailed for the development of automated surgical guidance systems[9]. On top of that, the approach is limited by the human capability to recognize complex patterns in the data[10], inherent cognitive biases of human annotators[11], and limitations of human attention span[12]. These challenges affect the scalability of manual approaches and fundamentally limit the discovery of more complex patterns in the data.

To address these challenges, we introduce a semi-automated surgical feedback analysis framework (Fig. 2). The framework takes as input raw text transcripts of surgical feedback delivered during live surgical cases (a) to perform a semi-automated discovery of feedback categories (b), which are then evaluated for their interpretability and ability to predict clinical outcomes (c). The core of our framework is the novel *Human-AI Collaborative Topic Refinement Process* that facilitates the discovery of meaningful categorization of live surgical feedback with minimum human effort. Our

[1]Computing+Mathematical Sciences, California Institute of Technology, Pasadena, CA, USA. [2]Department of Urology, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [3]Department of Urology, New York Presbyterian Hospital, Weill Cornell Medicine, New York, NY, USA. [4]Keck School of Medicine, University of Southern California, Los Angeles, CA, USA. [5]Department of Urology, University of Texas Southwestern Medical Center, Dallas, TX, USA. ✉e-mail: ajhung@gmail.com

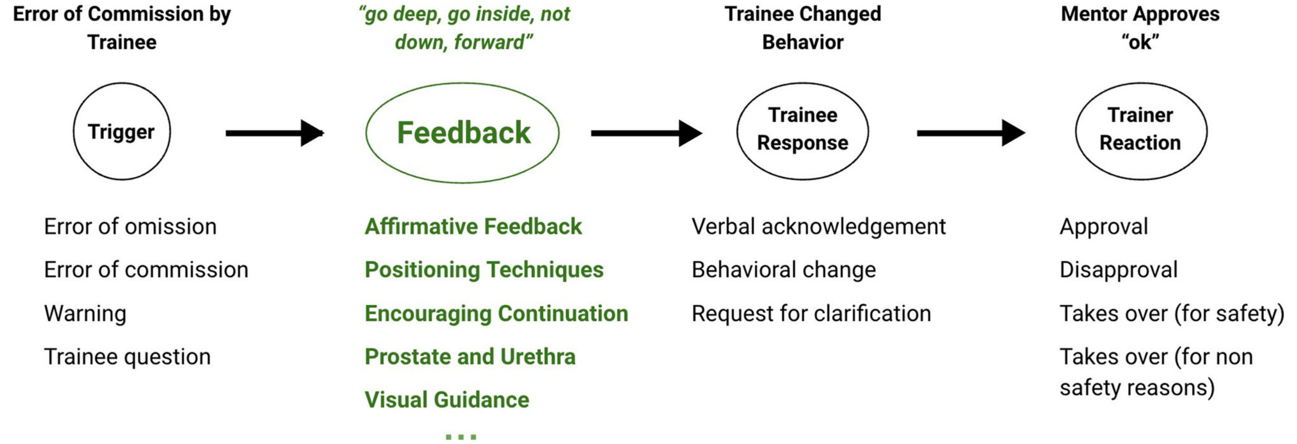


Fig. 1 | Informal feedback delivery process in live robot-assisted surgery. The need for feedback is often triggered by trainee behavior, while the feedback itself can relate to various aspects; we highlight some feedback content categories discovered with our framework. Feedback can impact the trainee's behavior, as well as result in verbal acknowledgment, or a request for clarification. Trainee behavior is sometimes also met with a subsequent reaction from the trainer. The categorization in black has been provided manually, while the feedback categories in green have been automatically discovered using our AI-based framework.

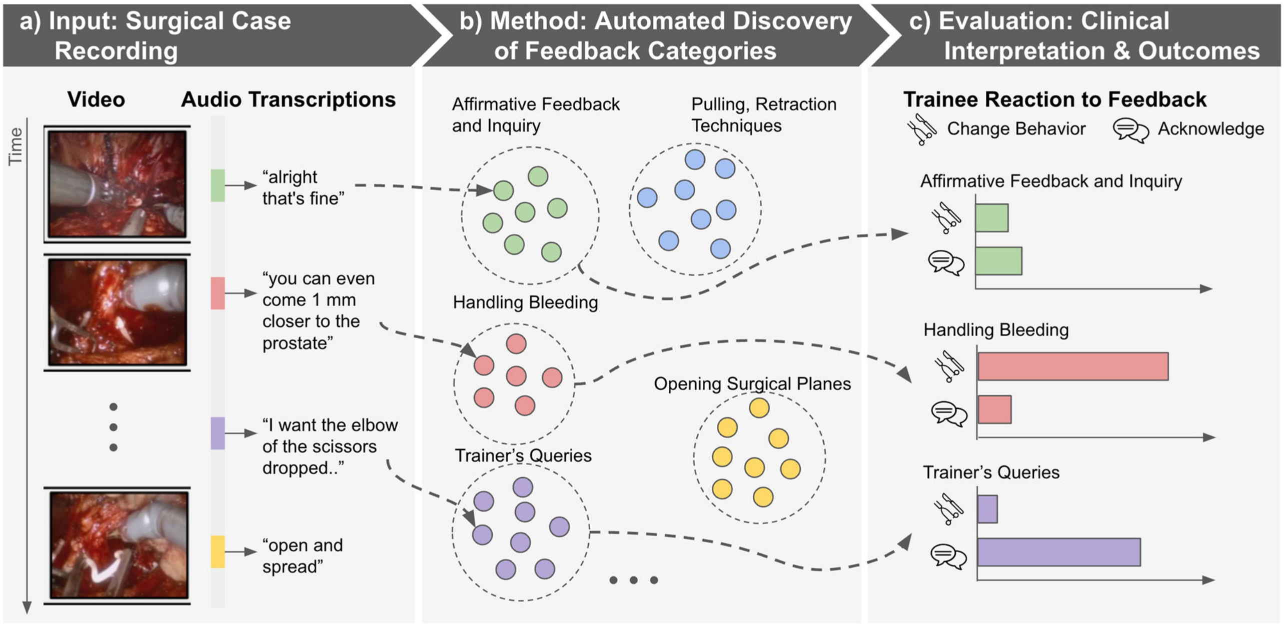

a) Input: Surgical Case Recording
b) Method: Automated Discovery of Feedback Categories
c) Evaluation: Clinical Interpretation & Outcomes
Video
Audio Transcriptions
Time
"alright that's fine"
"you can even come 1 mm closer to the prostate"
"I want the elbow of the scissors dropped.."
"open and spread"
Affirmative Feedback and Inquiry
Pulling, Retraction Techniques
Handling Bleeding
Opening Surgical Planes
Trainer's Queries
Trainee Reaction to Feedback
Change Behavior
Acknowledge
Affirmative Feedback and Inquiry
Handling Bleeding
Trainer's Queries

Fig. 2 | Overview of the surgical feedback analysis framework applied in this work. a We take as input raw transcripts of feedback (video is shown only for context) delivered during live surgical cases[7]. **b** We apply our novel *Human-AI collaboration process* to discover the categorization of this raw feedback into clinically meaningful categories. **c** We perform rigorous interpretation and evaluation steps, including the ability of the discovered categorization to predict clinical outcomes.

process involves 3 main steps: 1) Automated Topic Clustering, 2) Human Interpretation, and 3) Automated Topic Refinement. In the first step, we apply an automated unsupervised topic modeling technique called BERTopic, which extracts a representation of text (embedding) using a pre-trained large language model—BERT. This representation captures the semantic "meaning" of the text and groups feedback instances with similar meanings into the same topic. In the second step, we collect human interpretation and topic refinement suggestions, which are then automatically applied in the third step to produce refined topic clusters. Our process is designed to minimize the manual effort of categorizing surgical feedback at scale, while supporting human raters with initial discovery and ensuring the incorporation of valuable human insights.

We perform a rigorous evaluation of our discovery process and the clinical relevance of the resulting categorization of surgical feedback into topic clusters. Initial automated topic modeling resulted in the discovery of 28 topics, which were consolidated into 20 final topics following human interpretation and refinement. These topics were evaluated by two trained human raters in terms of "clinical clarity" defined as "meaningfulness for clinical practice". The top-scoring topics related to aspects of high clinical relevance such as "Controlling and Addressing Bleeding", "Sweeping Techniques", and "Trainer's Intervention & Visual Verification". Subsequent competitive evaluation against manually annotated categories from prior work[7] revealed that our AI-discovered topics offer a statistically significant and independent contribution to the prediction of trainee *Behavioral Response*, *Verbal Acknowledgment*, as well as *trainer Taking Control for non-safety related reasons*. We further show which discovered novel topics significantly contribute to outcome prediction, revealing categories related to high urgency feedback around *"Handling Bleeding"* and *"Trainer's Intervention"* as well as fine-grained categorization of feedback around instrument handling, such as *"Sweeping Techniques"* and *"Pulling,*

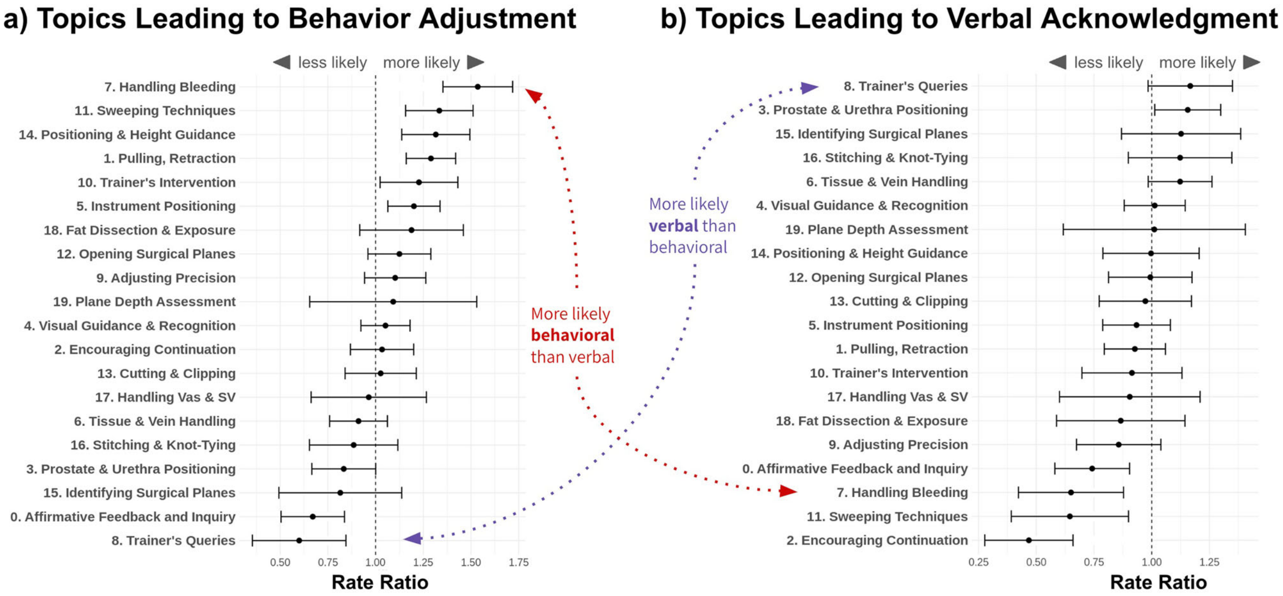


Fig. 3 | Discovered feedback topic clusters and their associations with different aspects of behavioral change of a trainee. a Association of discovered topics with trainee *Behavior Adjustment*, representing an observable adjustment made by a trainee that corresponds directly with the preceding feedback. **b** Association of discovered topics with trainee *Verbal Acknowledgment* representing audible reaction from the trainee confirming that they have heard the feedback. The strength of association was quantified as the Rate Ratio (RR) calculated as the rate of behavioral adjustment when feedback on a given topic was present over the rate when it was absent. The more the mean RR for a topic is to the right, the stronger the positive association. Whereas RR closer to the left denotes a negative association. We can see that some high-urgency feedback such as *"Handling Bleeding'* is much more likely to result in immediate behavior adjustment and less likely to lead to verbal acknowledgment (i.e., trainee just saying he/she understood the feedback). At the same time, intuitively, *"Trainer's Queries"* are much more likely to be met with just a verbal response from a trainee rather than a behavior change. We note that some of the topic titles were shortened for display purposes.

Retraction". Finally, the statistical analysis of the association of the discovered topics with trainee behavior offers clinically meaningful insights into which types of feedback are more likely to lead to behavior adjustment or verbal acknowledgment in the context of real-world surgeries.

To the best of our knowledge, we are the first to propose a Human-AI collaboration for the semi-automated discovery of surgical feedback categorization. We further rigorously demonstrate the utility of our approach, showing its statistically significant higher predictive power compared to manual human categorization, while requiring much less human effort. We further show the ability of our approach to reveal novel and clinically meaningful surgical feedback categories. Our findings reveal that certain feedback topics are particularly predictive of positive trainee outcomes, while others are less effective, offering practical tools for trainers to refine their feedback techniques, ultimately aiming to improve patient care and surgical proficiency across various specialties. Ultimately, our process can lead to automated quantification of live surgical feedback at scale and the development of automated coaching systems.

Results: clinical interpretation and validation

Our process led to the discovery of 20 AI topics capturing various aspects of surgical feedback (Fig. 3) based on a raw transcribed text of 3740 instances of live surgical feedback collected in prior work[7]. We evaluated these topics for their ability to independently contribute (on top of manual human categories[7]) to the prediction of clinical outcomes in the form of annotated Trainee Behavior and Trainer Reaction also provided in prior work (Table 1). We further analyze the statistical associations of the individual discovered topics with Trainee Behavior Adjustment (Fig. 3a) and Trainee Verbal Acknowledgment (Fig. 3b). Further details of the dataset, the *Human-AI collaborative refinement process*, and the evaluation can be found in the "Methods" section.

Clinical outcomes prediction

Table 1 presents the results of prediction of various aspects of Trainee Behavior using a Random Forest (RF) model. The full model includes both *human* manually-derived categories from prior work[7] and AI-discovered topics. The subsequent 2 variations of the full model remove firstly the AI-discovered topics or secondly the human manually-derived categories. The change in the AUROC (*ΔAUC*) describes the impact of the removal of these variables on model performance. Any negative values indicate a reduction in performance, and further statistically significant reductions for which a 95% confidence interval (95% CI) does not overlap with 0.0 are underscored. Further details of the analysis can be found in the Methods section. Similar results obtained with other supervised models can be found in Supplementary Note E.

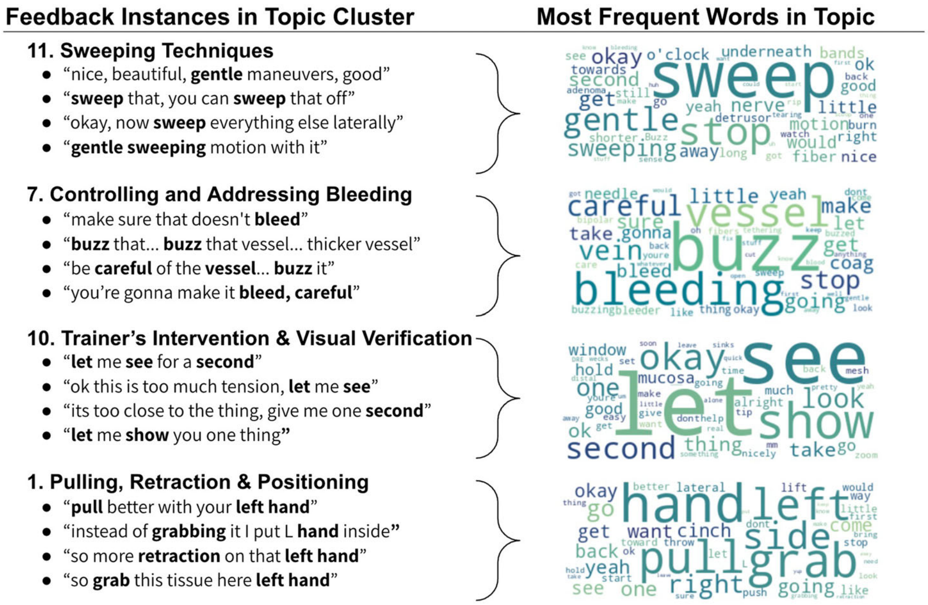
Without AI-discovered topics, the model accuracy statistically decreased in two of the three behavioral adjustment measurements with AUROC change of −0.03 (95% CI: −0.05, −0.02) and −0.04 (95% CI: −0.05, −0.03) in *"Verbal Response"* and *"Behavioral Response"* respectively. In contrast, without human-rated categories, only the decrease in *"Verbal Response"* was statistically significant with AUROC of −0.02 (95% CI: −0.03, −0.01) as in Table 1. Further investigation of the variables of importance in the RF model predicting *"(Trainee) Behavioral Adjustment"* revealed that seven of the AI-discovered topics were ranked among the top twelve predictive features with the out-of-bag Gini (OOBGini) score > 0 (Supplementary Note B). The positive OOBGini score suggests that the variable contributes positively to the model's accuracy. Specifically, the AI-discovered topics related to *"Pulling, Retraction"*, *"Trainer's Queries"*, and *"Sweeping Techniques"* ranked as the 2nd, 5th, and 6th most predictive variables. Similarly, four of the AI-discovered topics ranked among the eight top predictive variables for *"(Trainee) Verbal Acknowledgment"* (Supplementary Note C). Specifically, the AI-discovered topics related to *"Encouraging Continuation"* and *"Sweeping Techniques"* ranked as the 2nd and 3rd most predictive variables.

A large statistically significant reduction in AUROC of −0.10 (95% CI: −0.16, −0.04) was found with removal of AI-discovered topics for anticipating *"(Trainer) Taking Control (non-safety reasons)"*, which refers to a situation when the trainer takes full control of the surgical robot away from the trainee for demonstrative teaching purposes. An even larger reduction of

Table 1 | Independent contributions of AI Discovered Topics and Manual Human Labeled Categories from prior work to prediction of outcomes following surgical feedback

Category	Behavior Outcome	Full (AI+Human)		Without AI Topic Clusters		Without Human Categories	
		AUC	95% CI	Δ AUC	95% CI	Δ AUC	95% CI
Trainee Behavior	Verbal Acknowledgment	0.70	(0.69, 0.74)	−0.03↓4.3%	(−0.05, −0.02)	−0.02↓2.9%	(−0.03, −0.01)
	Behavioral Adjustment	0.74	(0.73, 0.76)	−0.04↓5.4%	(−0.05, −0.03)	−0.01 ↓1.4%	(−0.02, 0.0)
	Ask for Clarification	0.53	(0.47, 0.58)	0.01 ↑1.9%	(−0.05, 0.08)	−0.02 ↓3.8%	(−0.08, 0.04)
Trainer Reaction	Approval	0.66	(0.64, 0.69)	−0.01 ↓1.5%	(−0.04, 0.01)	−0.01 ↓1.5%	(−0.02, 0.01)
	Disapproval	0.63	(0.57, 0.70)	−0.04 ↓6.4%	(−0.11, 0.03)	−0.04 ↓6.4%	(−0.08, 0.0)
	Taking Control (safety)	0.82	(0.70, 0.94)	−0.15 ↓18.3%	(−0.30, 0.0)	0.03 ↑3.7%	(−0.02, 0.09)
	Taking Control (non-safety)	0.82	(0.76, 0.87)	−0.10↓12.2%	(−0.16, −0.04)	−0.03↓3.7%	(−0.06, −0.01)

A higher drop after the removal of *AI Topic Clusters* compared to the removal of *Human Categories* indicates the higher importance of these predictors. The analysis was performed using fivefold cross-validation. Significant changes at $\alpha = 0.05$ level are underlined.



11. Sweeping Techniques

- "nice, beautiful, **gentle** maneuvers, good"
- "**sweep** that, you can **sweep** that off"
- "okay, now **sweep** everything else laterally"
- "**gentle sweeping** motion with it"

7. Controlling and Addressing Bleeding

- "make sure that doesn't **bleed**"
- "**buzz** that... **buzz** that vessel... thicker vessel"
- "be **careful** of the **vessel**... **buzz** it"
- "you're gonna make it **bleed, careful**"

10. Trainer's Intervention & Visual Verification

- "**let** me **see** for a **second**"
- "ok this is too much tension, **let** me **see**"
- "its too close to the thing, give me one **second**"
- "**let** me **show** you one thing**"**

1. Pulling, Retraction & Positioning

- "**pull** better with your **left hand**"
- "instead of **grabbing** it I put L **hand** inside**"**
- "so more **retraction** on that **left hand**"
- "so **grab** this tissue here **left hand**"

Fig. 4 | Sample of AI discovered topics grouping trainers' feedback delivered live during surgery. We provide examples of feedback instances (left) and word clouds (right) summarizing the most frequent words in the topic.

AUROC of −0.15 (95% CI: −0.30, 0.0) was found for *"(Trainer) Taking Control (safety)"* without AI-discovered topics, which represents a rare situation when the trainer has to take control of the surgical robot out of concern for patient safety. Yet this difference did not reach statistical significance. For the removal of Human Categories, the reduction of −0.03 (95% CI: −0.06, −0.01) for *"(Trainer) Taking Control (non-safety reasons)"* also reached statistical significance, though with a much smaller effect size. There are no remarkable changes in model accuracy in other measurements after dropping either AI or human-rated categories, which indicates no unique contribution from either (Table 1)

Association analysis

Figure 3 depicts the associations of the AI-discovered topics with trainee behavioral outcomes after feedback: a) Behavioral Adjustment and b) Verbal Acknowledgment. Each AI topic is represented on a y-axis. Examples of the feedback instances under the discovered AI topics can be found in Fig. 4. The strength of association was quantified as the Rate Ratio (RR) calculated as the rate of behavioral adjustment when feedback on a given topic was present over the rate when it was absent. The more the mean RR for a topic is to the right, the stronger the positive association. Whereas RR closer to the left denotes a negative association. The vertical line at 1.0 represents a reference denoting average association strength. The error bars around average RR values for each topic represent 95% confidence intervals. Topics for which confidence intervals don't overlap with the reference line denote statistically significant association. This analysis has been performed using a multivariate generalized linear mixed-effects model (GLMM). Further details can be found in the "Methods" section.

Figure 3a shows AI discovered feedback topics associated with chances of trainee *Behavioral Adjustment*. As shown, the *"Handling Bleeding"* topic was the most strongly associated with *positive behavioral adjustment* with an RR of 1.54 (95% CI: 1.35, 1.72). Other strongly positively associated topics for trainee behavior change included *"Sweeping Techniques"* (RR: 1.33; 95% CI: 1.16, 1.51), *"Positioning & Height Guidance"* (RR: 1.32; 95% CI: 1.13, 1.49), *"Pulling, Retraction"* (RR: 1.28; 95% CI: 1.16, 1.42), and *"Instrument Positioning"* (RR: 1.20; 95% CI: 1.06, 1.34). Several topics were also *negatively associated* with the chance of trainees' *Behavioral Adjustment*, including *"Trainer's Queries"* (when a trainer asks a question to a trainee) (RR: 0.60; 95% CI: 0.35, 0.84) followed by *"Affirmative Feedback and Inquiry"* (RR: 0.67; 95% CI: 0.50, 0.84), and *"Prostate & Urethra Positioning"* (RR: 0.83; 95% CI: 0.67, 1.00).

Figure 3b shows discovered feedback topics associated with chances of *Verbal Acknowledgment* from a trainee. We can see that the topics capturing trainers' guidance on *"Prostate & Urethra Positioning"* as well as *"Trainer's Queries"* (when the trainer asks a question to a trainee) are *positively associated* with verbal acknowledgment from a trainee, with RRs of 1.16 (95% CI: 1.01, 1.30) and 1.17 (95% CI: 0.98, 1.35) respectively. At the same time, *"Encouraging Continuation"* represented the topic most *negatively associated* with verbal acknowledgment with an RR of 0.47 (95% CI: 0.28, 0.66). Other topics with such negative association included feedback around *"Sweeping Techniques"* (RR: 0.65; 95% CI: 0.39, 0.90), *"Handling Bleeding"* (RR: 0.65; 95% CI: 0.42, 0.88), and *"Affirmative Feedback and Inquiry"* (RR: 0.74; 95% CI: 0.58, 0.90).

Discussion

The critical importance of verbal feedback during surgical training has been increasingly recognized[1,13], playing a key role in both immediate trainee performance adjustments[2] and long-term surgical skill acquisition[3]. Interpersonal communication challenges have been indicated as responsible for many inefficiencies and errors in the operating room (OR)[14]. Yet, understanding how feedback is delivered in live surgeries and what constitutes optimal feedback requires a clinically meaningful categorization scheme[15]. Existing efforts focus on categorizing narrow aspects of feedback such as teaching behaviors (e.g., informing, questioning, responding)[5], communication type (e.g., commanding, explaining, questioning)[6], or coarse-grained content (e.g., anatomic, procedural, technical) and delivery aspects (e.g., praise, criticism)[7]. They also require substantial manual effort and ultimately suffer from limited human ability to recognize complex patterns in the data[7]. Existing clinical evaluations of prior categorization approaches often lack examination of impact on trainee behavior[5] and merely provide descriptive statistics[6]. The most thorough evaluation to date only analyzed associations between certain categorization subsets and trainee behavior[7]. Our work addresses these gaps by providing a comprehensive evaluation.

In this work, we introduce a novel *Human-AI Collaborative Refinement Process* for categorizing live real-world surgical feedback, which leverages unsupervised learning techniques complemented by human expertise. By first automatically discovering feedback categories from raw surgical transcripts, we reduce the reliance on expensive manual annotation, enhancing scalability and improving the precision of feedback analysis. Leveraging machine learning addresses the inherent challenges related to human attention span[12], the subjectivity of individual raters[11], knowledge limitations[16], as well as the limitations of human raters to recognize complex patterns in the data at scale[10]. At the same time, with support for subsequent low-effort human interpretation and refinement, we provide crucial supervision from human raters in this high-stakes domain[17]. The benefit of our approach lies in hybrid Human-AI collaboration, which leverages the strengths of both unsupervised machine learning and human refinement to achieve high interpretability and clinical relevance of the discovered feedback categorization into topics. This is a novel application of Human-AI collaboration in surgical space. Our method also has substantial practical importance, as it offers a scalable, efficient way to analyze and categorize surgical feedback, which is crucial for training and assessment in medical education[18].

For unsupervised topic discovery, we employ BERTopic[19], an embedding-based topic modeling framework that offers substantial benefits over traditional methods like Latent Dirichlet Allocation (LDA)[20], Hierarchical Dirichlet Processes (HDP)[21], and Non-negative Matrix Factorization (NMF)[22]. Traditional keyword-based models depend on word co-occurrence and do not perform well with brief texts comprising only a few words[23], typical in our data. They also necessitate extensive preprocessing (e.g., stop word removal, stemming), which can impact the results[24]. In contrast, BERTopic leverages pretrained text embeddings that capture semantic meaning, making it especially effective for shorter texts[23], and capable of accurately representing polysemous words and synonyms[25]. It groups similar feedback instances into topics effectively without relying on identical keywords, producing more coherent and diverse topic representations. Furthermore, embedding-based methods can directly process raw texts[26], benefiting from a broader linguistic context. We also enhance topic discovery with GPT-4 prompting-based topic title generation. This method utilizes both frequent keywords and representative feedback instances to craft contextually meaningful titles without depending solely on dominant keywords. It adapts especially well to changes in topic composition following human feedback.

Our study successfully demonstrated three key results: a) the effective unsupervised categorization of surgical feedback into clinically relevant topics, b) the significant and independent contribution of these AI-discovered topics to predicting trainee behavioral outcomes in addition to prior human categorization, and c) the ability of our method to reduce the human effort significantly compared to manual categorization. The AI-discovered topics highlight several intuitive associations with trainee behavior, which provide additional support for their meaningfulness and practical value. We can see that feedback on topics such as *"Handling Bleeding"*, *"Sweeping Techniques,"*, and *"Positioning & Height Guidance"* is most likely to result in a change in trainee behavior. The effectiveness of these feedback topics aligns well with findings from prior studies that emphasize the importance of targeted, actionable, and specific feedback in surgical training environments[9,27]. Furthermore, feedback on *"Handling Bleeding"* is especially likely to result in the subsequent change in trainee behavior due to its urgency. The trainee is much more likely to try to address the bleeding as soon as possible, rather than simply acknowledge the reception of feedback or engage in verbal discussion. On the other hand, also intuitively, the topic capturing *"Trainer's Queries"* (when a trainer asks a question to a trainee) is much more likely to result in a verbal acknowledgment from a trainee rather than an immediate behavior response. Feedback topics such as *"Adjusting Precision"* and *"Cutting & Clipping"* result in an average rate of behavioral adjustment, likely due to their lower urgency.

The analysis of the variables of importance for the prediction of trainee behavior reveals which discovered AI topics are of particular importance for predicting trainee behavior and supplementing prior human categorization[7]. Even more importantly, we can see that these components offer meaningful discoveries likely generalizable to other types of surgeries. Specifically, the topics around *"Handling Bleeding"* and *"Trainer's Intervention"* highlight the importance of capturing feedback expressing high-stakes interventions that require immediate reaction or can incur safety concerns. Furthermore, the predictive importance of having separate topics around *"Pulling, Retraction"* and *"Sweeping Techniques"* suggests the value of capturing more fine-grained aspects around the technical and procedural execution of surgical tasks. These technical aspects are fairly universal across surgery types[28,29]. Finally, the independent predictive value of *"Trainer's Queries"* and *"Trainer's Interventions"* suggests that capturing the teaching style of a trainer is also important, which aligns with selective categorization schemes from prior work around teaching behavior[4] and intraoperative communication[6].

We have rigorously evaluated our approach following several human interpretation and statistical validation steps. Two trained human raters independently interpreted the initially discovered topics, rating them in terms of *clinical clarity* and *consistency*. These interpretations have been repeated at each step of the discovery and refinement process. We further evaluated the statistical association of the discovered topics with trainee behavioral responses and trainer reaction dimensions annotated independently in prior work[7]. Finally, a biostatistician who was not involved in the development of the Human-AI collaborative approach served as the independent evaluator to assess the AI model performance. The independent biostatistician also further inspected the variables of importance in these prediction models, to understand which of the AI-discovered topics provided the novel categorization on top of manual categories. This multifaceted evaluation with independent qualitative interpretability and statistical validity checks performed by different analysts enhanced the credibility of our findings and further improved on the evaluation rigor from prior work in this space.

While our method marks a significant advancement in surgical feedback analysis, it is not without limitations. The initial unsupervised learning approach relies on the quality of the extracted text meaning representations (embeddings), which may be affected by medical vocabulary and professional slang[30]. While we followed the best practices, any automated clustering method has inherent instability[31] which may lead to potentially better starting points for the initial step of topic discovery. We currently also don't incorporate the multimodal (i.e., video and audio) nature of feedback, relying only on transcribed text (Fig. 2 includes video only for context). We note that there are techniques such as multimodal transformer[32] which can facilitate that, but this also leads to the additional burden of interpretation and refinement for the human raters. Additionally, our approach, while paving the way for automated quantification of live surgical feedback at scale, still benefits from human-AI collaboration for clinical interpretation and topic refinement. However, as we show in the additional analysis in Supplementary Note D, human refinement is not crucial for the good performance of our method in behavior prediction tasks but rather serves to enhance interpretability and clinical clarity.

The broader impact of this research is multifaceted, contributing to advancements in real-world surgical feedback analysis, effective Human-AI collaboration in health, and the development of automated training tools in surgery.

Our approach crucially enables the *analysis of real-world surgical feedback at scale* by leveraging unsupervised machine learning techniques coupled with minimal yet strategic human supervision. As such, it can substantially aid in understanding the complex dynamics of communication in the operating room (OR), which has been a long-standing challenge[15].

We also *enhance the completeness of feedback categorization* by using AI methods that discover complex patterns in the data, which might otherwise be overlooked by human raters. The initial automation step also helps to minimize the subjective interpretation biases typically associated with individual institutions or specific training backgrounds.

Additionally, the Human-AI collaboration approach we propose is likely to *improve acceptance among medical professionals* compared to a fully automated system without human input[33]. Human supervision and refinement enhance interpretability and empower human raters by incorporating their expert judgment, ensuring that the AI complements rather than replaces the human element in surgical training.

Our approach can also *inform controlled studies and lead to the development of real-time warning and feedback systems*. Our process results in a higher granularity of categorization associated with statistically significant associations with trainee behavior, which can facilitate personalized training and assessment approaches[3]. For example, the absence of a specific category of feedback in particular training contexts, when expected, could signal a level of subjectivity or simple omission that can diminish the quality of training[34]. Such granularity can help formulate data-driven hypotheses that inform the design of specific targeted clinical studies investigating different properties of feedback in an experimental setting. Ultimately, our approach can lead to the development of an automated real-time warning and feedback system in a surgical context. Such a system could warn the trainer about potentially ineffective feedback in real-time or suggest what feedback might be valuable to consider in particular contexts, serving as an AI-driven assistant much needed in surgical settings[35].

Further research could explore the integration of multimodal data, such as video, and audio, into the AI analysis framework. Multimodal integration could lead to more nuanced insights into surgical performance and training needs. There is significant potential for developing real-time feedback and warning systems based on the categorization and analysis achieved in this study. Future work could focus on the design and implementation of these systems in live surgical settings to provide immediate guidance and correction, thereby enhancing the safety and effectiveness of surgical training and operations.

Methods

We aim to categorize the recorded instances of feedback delivered during surgery into clinically meaningful and interpretable categories with minimal manual effort from experts. In this section, we describe the dataset on which we applied our process, as well as the details of our *Human-AI Collaborative* approach.

Ethics approval

All datasets used in this study were collected following rigorous ethical standards under the approval of the Institutional Review Board (IRB) of the University of Southern California, ensuring the protection of participants' rights and privacy. Written informed consent was obtained from all individuals who participated in the dataset collection (HS-17-00113). Furthermore, to safeguard the privacy and confidentiality of the participants, the datasets were de-identified prior to model development or analysis.

Surgical feedback dataset

We use a dataset of real-life feedback delivered during the course of robot-assisted surgeries[7]. The dataset contains audio recordings of conversations in the operating room (OR) captured by wireless microphones worn by the surgeons. Our data also contains video captured from an endoscopic camera representing the surgeon's point of view. Video and audio were captured simultaneously using an external recording device. Each surgery utilized the da Vinci Xi robotic surgery system[36].

A subset of the conversations recorded in the OR represents surgical feedback. Such feedback is defined as any trainer utterance intended to modify trainee thinking or behavior. Trainers were defined as those providing feedback, while the trainees were the recipients of such feedback. Utterances meeting this definition were timestamped and manually transcribed from audio recordings by medical residents with surgical knowledge. Only feedback instances meeting these criteria were transcribed, while other types of dialog were excluded. The dataset contains 3740 individual feedback instances as shown in Table 2.

Further, the dataset contains annotations of several dimensions related to the outcomes (Table 2). The annotations related to *Trainee Behavior* capture any form of reaction from a trainee to the feedback provided by the trainer. The annotated trainee reactions include *Verbal Acknowledgment*, which captures the trainee's verbal confirmation of understanding the feedback, *Behavioral Adjustment*, which captures the change in behavior of a trainee aligned with the feedback, and *Ask for Clarification*, which captures instances where feedback was not clear and triggered the trainee to require further explanation.

Additionally, the dataset includes annotations of *Trainer Reaction*, which capture trainers' responses to changes, or the absence thereof, in trainee behavior following feedback. These include: *Approval*, which indicates trainers' satisfaction with trainee behavior, and *Disapproval*, which captures the opposite. Situations when the trainer needs to take over control were also annotated with *Taking Control (non-safety)* capturing instances when the trainer takes over the console for non-safety related reasons, while *Taking Control (safety)* captures instances when the trainer had to take over control out of concern for patient safety. The precise definitions as well as the prevalence of the annotation in our dataset are reported in Table 2. Further details around data collection and annotation can be found in ref. 7.

Human-AI collaboration for unsupervised feedback categorization

We introduce a novel *Human-AI Collaboration* process for unsupervised categorization of surgical feedback, as depicted in Fig. 5. This process first leverages the unsupervised clustering technique to automatically categorize the unstructured feedback text into topics (*1. Automated Topic Clustering*). Such initial categorization is then inspected by trained human raters (*2. Human Interpretation*). This step involves low-effort human feedback about interpretation and adjustments needed to the discovered topics. The human-suggested adjustments are then leveraged to automatically reorganize the topics (*3. Automated Topic Refinement*). This process can be repeated several times as needed. By applying this process, we discovered feedback topics that align with clinically meaningful categorization while requiring minimal human effort compared to fully manual annotation of each feedback instance as performed in prior work[7]. In this section, we provide the details of the unsupervised topic clustering method used as a starting point, and the human interpretation and refinement steps.

Table 2 | Statistics of the annotations in our dataset per behavior categories

Category	Dimension	Definition	Count	Freq	Count/Case
Feedback	Instances	*Trainer utterance intended to modify trainee thinking or behavior*	3740	100.0%	129.0 ± 77.4
Trainee Behavior	Verbal Acknowledgment	*Verbal or audible reaction from the trainee confirming that they have heard the feedback*	1691	45.2%	58.3 ± 31.4
	Behavioral Adjustment	*Behavioral adjustment made by a trainee that corresponds directly with the preceding feedback*	1666	44.6%	57.5 ± 39.3
	Ask for Clarification	*Trainee asks for feedback to be restated due to lack of understanding*	77	2.1%	3.2 ± 2.4
Trainer Reaction	Approval	*Trainer verbally demonstrates that they are satisfied with the trainee behavioral change*	552	14.8%	19.7 ± 16.9
	Disapproval	*Trainer verbally demonstrates that they are not yet satisfied with the trainee behavioral change*	71	1.9%	3.6 ± 2.6
	Taking Control (non-safety)	*Trainer takes control of the robot for non-safety related reasons*	98	2.6%	3.6 ± 2.3
	Taking Control (safety)	*Trainer takes control of the robot out of concern for patient safety*	18	0.5%	1.3 ± 0.5

We provide absolute total counts (*Count*), the relative frequency of behavior per feedback instance counts (*Freq*) as well as the prevalence per surgical case (*Count/Case*).

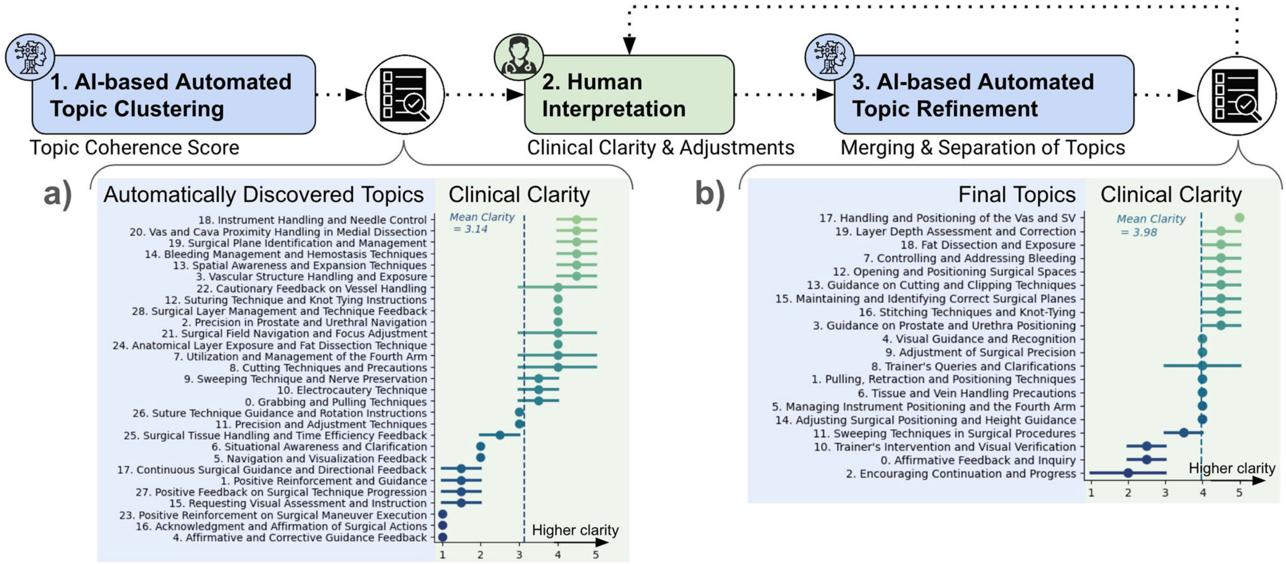


Fig. 5 | Human-AI collaboration process for unsupervised surgical feedback topic discovery and refinement. The first step involves the automated discovery of **a** an initial categorization using unsupervised topic modeling techniques. These topics are then evaluated by human raters, who also provide suggestions for refinement. These suggestions are then automatically applied to **b** refine the discovered topics further.

We applied minimal preprocessing of feedback text instances to remove any identifying mentions of trainers or trainees. We have further normalized the representation of some short phrases (e.g., "*k*" with "*ok*"). We also replaced some common abbreviations for surgical terms with their full-text equivalents, such as *"DVC"* replaced with *"Dorsal Vein Complex"*. This minimal preprocessing has been applied to facilitate the extraction of pre-trained representations (embeddings) for feedback instance text. Contrary to keyword-based methods (e.g., LDA[37]), neural techniques perform best without removal of *"stop words"* as transformer-based embedding models need the full context to create an accurate embedding[38].

AI-based automated topic modeling

We employ BERTopic[19], a modern topic modeling technique, to analyze textual data and extract meaningful topic clusters. As shown in Fig. 6 the BERTopic pipeline follows a distinct workflow to model topics. **Step 1** it converts all feedback instances in the dataset into numerical representations (i.e., *embeddings*) that represent their "meaning". These embeddings can be used to numerically score the semantic similarity between feedback text instances. **Step 2** involves projecting the embeddings to a lower-dimensional space, facilitating the grouping of feedback instances that share similar content. **Step 3** involves clustering the feedback instances into distinct topics, such that feedback instances with similar meaning are grouped into the same *Topic*. The final step, **Step 4**, involves summarizing the "meaning" of each topic. This involves providing a concise description capturing the pattern of the feedback instances grouped under a particular *Topic*. To decide on the initial number of topics, we relied on a widely used topic coherence metric—normalized pointwise mutual information (NPMI)[39]. This coherence measure has been shown to emulate human judgment with reasonable performance[40]. For our data, this metric achieved the best score for a range between 20 and 28 topics. We further provide a more detailed description of the exact techniques used in each step.

Step 1: Numerical Representation of Text (Sentence Embeddings): BERTopic leverages the pre-trained embeddings that represent textual data in a numerical format that captures the semantic "meaning" of the text. Such embeddings can be used to calculate a score for the similarity in meaning between feedback instances (cosine similarity) as shown using a few examples in Table 3. Using this similarity score, a very similar pair of feedback instances receives a score close to 1.0, while very different instances receive a score close to 0.0. These embeddings are extracted from the BERT model, a transformer-based language model pre-trained on a large corpus of text[41]. The embeddings are further specialized for capturing sentence similarities using the *Sentence Transformer* framework called SBERT[42]. We

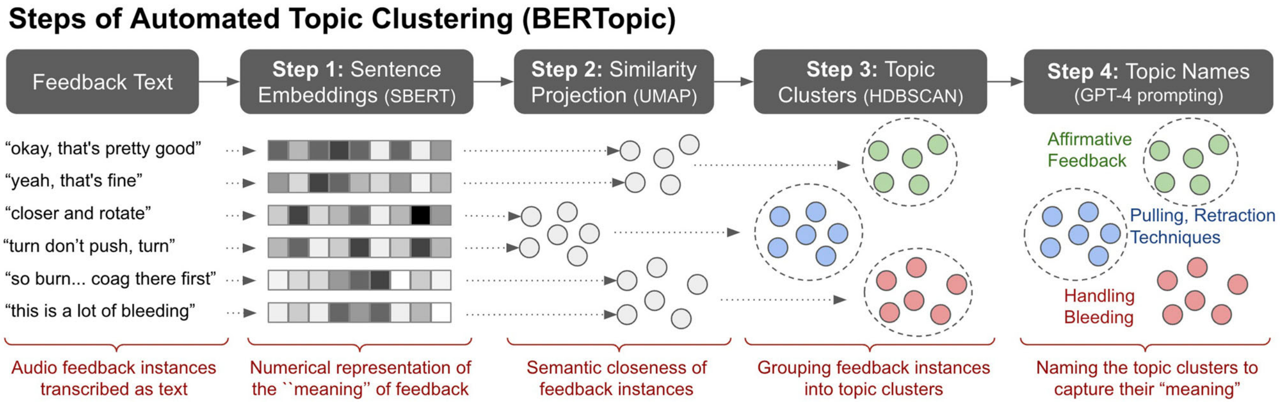

Steps of Automated Topic Clustering (BERTopic)
Feedback Text
Step 1: Sentence Embeddings (SBERT)
Step 2: Similarity Projection (UMAP)
Step 3: Topic Clusters (HDBSCAN)
Step 4: Topic Names (GPT-4 prompting)
"okay, that's pretty good"
"yeah, that's fine"
"closer and rotate"
"turn don't push, turn"
"so burn... coag there first"
"this is a lot of bleeding"
Affirmative Feedback
Pulling, Retraction Techniques
Handling Bleeding
Audio feedback instances transcribed as text
Numerical representation of the ``meaning" of feedback
Semantic closeness of feedback instances
Grouping feedback instances into topic clusters
Naming the topic clusters to capture their "meaning"

Fig. 6 | Processing steps involved in the discovery of topics from surgical feedback instances. These steps are provided under the BERTopic framework[19]. *Step 1* involves obtaining numerical representations of feedback texts using pretrained embeddings. *Step 2* applies dimensionality reduction and projection of the mebeddings into lower-dimensional space. *Step 3* clusters the feedback instances into topics based on similarity metric for text represenations. *Step 4* summarizes the contents of each topic cluster using concise titles provided by GPT-4 prompting.

Table 3 | Examples of feedback instances from our dataset along with their semantic similarity scores calculated using pretrained *'all-MiniLM-L12-v2'* sentence-embeddings (SBERT)[42]

Feedback Instance Example	Feedback for Comparison	Similarity Score
"okay, that's pretty good"	"yeah, that's fine"	0.506
	"closer and rotate"	0.119
	"so burn... coag there first"	0.121
"closer and rotate"	"turn don't push, turn"	0.377
	"yeah, that's fine"	0.058
	"this is a lot of bleeding"	0.042
"so burn... coag there first"	"this is a lot of bleeding"	0.257
	"yeah, that's fine"	0.056
	"closer and rotate"	0.094

used pretrained embeddings using the "*all-MiniLM-L12-v2*"architecture[43]. The core idea behind BERTopic is to use these embeddings to capture the contextual relationships between whole, even very short sentences, thereby enabling the identification of more coherent and semantically rich topics compared to traditional methods[23].

Step 2: Similarity Projection: BERTopic applies dimensionality reduction techniques, specifically UMAP (Uniform Manifold Approximation and Projection)[44], to project the extracted sentence embeddings into a lower-dimensional space. This step is essential for mitigating the "curse of dimensionality," enhancing clustering performance by making distances between high-dimensional text embeddings more meaningful in a lower-dimensional space. Moreover, by projecting data into a lower-dimensional space, it retains the integrity of semantic similarities, crucial for accurately reflecting the thematic structures within the text corpus[45]. The UMAP settings that resulted in the highest coherence score in our settings were the following: No. of neighbors = 15, No. of components = 5, min-dist = 0.05, metric="*cosine*".

Step 3: Detecting Topic Clusters: For the clustering step, BERTopic utilizes the HDBSCAN algorithm (Hierarchical Density-Based Spatial Clustering of Applications with Noise)[46], which is capable of identifying clusters of varying densities and shapes, making it well-suited for the diverse nature of text data[45]. HDBSCAN constructs a hierarchy of clusters by examining the dataset across different scales, calculates mutual reachability distances based on core distances to maintain density consistency, and then uses these distances to build a minimum spanning tree (MST)[47]. From this MST, it generates a dendrogram representing the hierarchical structure of clusters. Clusters are then selected based on their stability within this hierarchy, with more stable clusters (based on the excess-of-mass algorithm[45]) deemed significant, while points not belonging to any stable cluster are classified as noise. This approach allows HDBSCAN to effectively handle datasets with complex structures and varying densities without pre-specifying the number of clusters, making it especially suited for real-world data analysis where the true cluster structure is unknown. The parameter values used for this step were min-cluster-size = 50, metric="*euclidean*", cluster-selection-method="*excess of mass*"[48]. We further applied outlier reduction by calculating the Class-based Term Frequency-Inverse Document Frequency (c-TF-IDF)[49] representations of outlier documents and assigning them to the best matching c-TF-IDF representations of non-outlier topics. This is the default method supported by BERTopic[19].

Step 4: Summarized Topic Representation (Topic Names): The final step of the process is providing concise and meaningful titles for the topic clusters based on the underlying feedback instances in the cluster. The framework provides several methods to accomplish this[50]. In this work, we combined the KeyBERTInspired[51] representation for topic cluster keywords extraction with GPT-4-based instruction prompting to determine the final title for each topic. KeyBERTInspired method extracts the candidate keywords from a topic cluster using frequency and importance weighting based on the uniqueness of the keywords for the cluster using c-TF-IDF. It then calculates the semantic similarity of these keywords to the representative feedback instances to select the most relevant ones. Unlike traditional keyword extraction methods that rely on statistical frequency, KeyBERT focuses on the semantic significance of words and phrases, selecting those with the highest similarity to the document's overall content. We then combine these extracted keywords with a few representative feedback instances for a topic to prompt GPT-4 for a short label for the topic using the instruction prompt provided in Supplementary Note A.

Human interpretation

Following automated topic modeling, we facilitate human interpretation (Fig. 5).

To evaluate the clinical value of the initially discovered topics, we performed a human evaluation with two trained human raters knowledgeable about the aspects of surgeries. Each rater was provided a spreadsheet with the raw text of the transcribed feedback instances, together with a column representing the title of the main topic to which the instance was automatically assigned. The raters had access to both keyword-based

Table 4 | Example refinement rules based on human interpretation and guidance

Initial Topic Name(s)	Initial Topic Keyword(s)	Rule Type	Refinement Outcome
10. Electrocautery Technique	“buzz”, “sweep”, “that”, “it”	Split	“buzz” from “sweep”
9. Sweeping Technique and Nerve Preservation	“sweep”, “gentle”, “off”, “do”	Split	“coag” from “sweep”
19. Surgical Plane Identification and Management	“plane”, “this”, “travel”, “easy”	Split	“travel” from “plane” and “easy”
18. Instrument Handling and Needle Control	“needle”, “your”, “instrument”, “driver”	Split	“needle” from “instrument” and “driver”
14. Bleeding Management and Hemostasis	“bleed”, “stop”, “there”, “gone”	Split	“bleed” from “stop”
27. Positive Feedback on Surgical Technique	“work”, “job”, “good”, “nice”	Merge	Combine into a new topic: “0. Affirmative Feedback and Inquiry”
23. Positive Reinforcement on Maneuver Execution	“nice”, “move”, “beautiful”, “very”		
1. Positive Reinforcement and Guidance	“good”, “okay”, “great”, “that”		
17. Surgical Guidance and Directional Feedback	“keep”, “go”, “let;, “okay”	Merge	Combine into a new topic: “2. Encouraging Continuation and Progress”
16. Acknowledgment and Affirmation of Actions	“mhm”,“mmhm,“mhmm”,“lucky”		
4. Affirmative and Corrective Guidance Feedback	“yup”, “yeah”, “uh”, “huh”		

The topic refinement rules were categorized into three types: splitting, merging, and approval. Where approval required no adjustment to the originally discovered topic cluster.

representations of the topics and their titles, generated via GPT-4 prompting.

- **Suggested Reorganization:** As part of the quality evaluation of the topics, the raters were also asked to provide suggestions for refining the discovered topics by suggesting whether a topic could be combined with another existing topic (merging) as well as whether the feedback instances grouped under one topic should be separated into two or more topics (splitting). These suggestions were aggregated across the two raters and applied in the topic refinement round.
- **Rating Topics:** The raters were asked to evaluate the quality of each topic based on two criteria: *“Clinical clarity,”* defined as *“the meaningfulness for clinical practice,”* as well as *“Consistency,”* defined as *“the uniformity of feedback instances representing the same aspect.”* These dimensions were rated for each topic using a 5-point Likert scale from *“1 - least clear”* to *“5 - most clear”* for *“Clinical clarity”* and from *“1 - least consistent”* to *“5 - most consistent”* for *“Consistency”*. The agreement was measured using a 2-way mixed intra-class correlation with absolute agreement (ICC)[52].

The interpretation round was performed twice with the same set of raters. First, after the initial fully automated categorization of feedback instances into 28 topics (Fig. 5a) and then in the second round, after implementing the suggested reorganizations from the first round resulting in 20 refined topics (Fig. 5b). In each round, raters had access to individual feedback instances assigned to the discovered topics to provide their interpretation. The ratings between rounds were statistically compared using an independent samples t-test based on a random sample of 30 feedback instances under each topic.

Refinement impact on clarity and consistency. The initial automatically extracted topics received a human-derived average *clinical clarity* score of 3.14 (SD = 1.36) and *consistency* score of 3.35 (SD = 1.28). After clinicians' input, the topics were consolidated into 20, which scored higher on *clinical clarity* with an average score of 3.98, (SD = 0.92, $p < 0.05$) as well as *consistency* with an average score of 4.35 (SD = 0.77, $p < 0.01$). Clinical clarity scores exhibited good agreement among raters in both rounds (Round 1: ICC = 0.85, 95% CI: 0.67, 0.93; Round 2: ICC = 0.72, 95% CI: 0.30, 0.89), while consistency exhibited moderate agreement (Round 1: ICC = 0.74, 95% CI: 0.44, 0.88; Round 2: ICC=0.62, 95% CI: 0.04, 0.85)

AI-based automated topic refinement

Following the suggestions from the first round, the topics were reorganized. In total, human Rater 1 provided 4 suggestions for topic refinement, while Rater 2 provided 9 suggestions. These suggestions were reconciled with the raters based on a follow-up discussion to formulate a final set of adjustments. The disagreements between the raters were on the desired level of abstraction, rather than on substantially different or conflicting grouping suggestions. The human refinement rules were then grouped into three categories. Examples of concrete rules can be seen in Table 4. We also summarize the rule types below:

- **Splitting Rules**, which captured suggestions for separating the feedback instances automatically categorized under one topic into separate topics. There were seven rules of this kind.
- **Merging Rules**, which captured suggestions for combining feedback instances originally automatically assigned to separate topics under one topic. There were five rules of this kind.
- **Approval Rules**, which captured approving the original automatically detected topic and the feedback instances assigned to it. There were 15 rules of this kind.

These 27 rules were applied sequentially starting with *Splitting*, followed by *Merging* and then *Approval*. This was intentional, as some of the intermediate topics created via splitting were subsequently merged. The rules were applied via a simple keyword-based filtering of the feedback instances. This process reduced the number of topics from the initially discovered 28 (see Fig. 5a) to the final set of 20 topics (Fig. 5b with examples in Supplementary Note F).

The resulting new set of topics with reorganized feedback instances was again passed through Step 4 of the BERTopic framework to extract automated keywords and prompt GPT-4 for topic names. These representations were used in the second round of human interpretation and subsequent validation steps. Examples of final AI topics along with concrete feedback instances under these topics as well as the word clouds with the most frequent words for a topic can be found in Fig. 4.

Refinement impact on downstream tasks. We evaluated the impact of the Human Refinement of AI-detected topics on the prediction of outcomes following surgical feedback. Supplementary Note D compares AI topics before refinement (“AI Initial” as in Fig. 5a) to AI topics after Human Refinement ("AI+Human Refinement” as in Fig. 5b). The AUC differences were tested using Delong's z-test and Random Forest model. We used a fivefold cross-validation procedure, which produces very narrow confidence intervals. We can see that the differences are very small showing that human refinement did not have much impact on the prediction of behavioral outcomes. We emphasize that this comparable performance has been achieved with 20 instead of 28 topics as predictors. Furthermore, human inspection and refinement helped validate the topics in terms of clinical meaning and helped improve "clinical clarity" and "consistency".

Evaluation setup for discovered topic clusters

We first evaluated the independent contributions of AI-discovered topics to prediction of trainee behavioral outcomes. Random Forest (RF) in a fivefold cross-validation setup was used to build models for predicting trainee behavioral outcomes (Table 1). We first split the data into five pieces using a stratified sampling procedure to ensure each piece has an equal case and control sample. Then we train the model with 80% training sample five times. Within each training process, we used a fivefold cross-validation for hyperparameter turning. The hyperparameter setting for RF was based on grid search using an interval of (5, 10, 25, 50, 100) number of variables to enter and 200 to 1000 trees by an interval of 100. The other hyperparameters were set as a maximal depth of 50, and leaf size of 5. The optimized hyperparameters are selected by minimizing the out-of-bag misclassification rate. The Gini impurity index was used as the loss function. We used King's method to address the imbalanced data[53]. We have built and compared three models using Delong's Z test[54]. The full model is defined as using both AI topics and human-rated categories[7] as the input. The two reduced models dropped either AI topics or human-rated categories. The difference in prediction accuracy AUROC between full and reduced models represents the independent contribution of either AI topics or human rating in predicting trainee behavioral outcomes. For example, the AUROC difference between the full model and the reduced model without AI topics represents the performance loss without AI topics as the input. If this loss is statistically significant, it means the AI topics had irreplaceable contributions to model prediction accuracy. For the full model, we also exported the variables of importance ranked by the Out-of-bag Gini index (OOBGini). It is another evidence of contribution from either AI or human rating in the association with trainee behavioral outcomes.

Aside from RF, we have also explored the use of other supervised models, specifically Elastic-Net and Adaboost (see Supplementary Note E). RF and AdaBoost are considered non-parametric approaches while ElasticNet is considered a parametric approach in case of strong linear predictors. All the models use AI Topic Clusters after the Human Refinement step as predictors. For the Adaboost, since it is more efficient, only 25 trees were built with a depth of 3 as recommended by ref. 55. For ElasticNet, the α value was optimized by gradient descent in the range of 0.1 to 0.02 by 0.001, and the hyperparameter tuning for L1 ratio was searched between 0 to 1 by increment of 0.1 to minimize cross-validation predicted residual sum of squares (CVPRESS).

We further investigated the adjusted association between individual topics and trainee behavioral outcomes. The independent influence of discovered topics (coded as 0 or 1 for the absence or presence of a topic, respectively) on binary-coded trainee *Behavioral Reaction* and *Verbal Acknowledgment* was investigated using a multivariate generalized linear mixed-effects model (GLMM)[56]. The model included fixed effects for each topic and a random intercept, to accommodate the clustering of data by surgical cases. This acknowledges that observations within the same surgical case are likely to be more similar than those from different cases. The supplementary Variance Inflation Factor[57] analysis indicated no multicollinearity among the predictors, affirming that the linear dependencies between predictors do not inflate the estimated coefficients.

After model fitting, the fixed effect coefficients were exponentiated to obtain the RRs, offering an interpretation of how the presence of each topic (versus its absence) affects the trainee reaction, within the context of the Poisson distribution's log-linear relationship. We also calculated Wald confidence intervals for the fixed effects.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Code availability

All unsupervised models were developed using Python and standard deep-learning libraries such as PyTorch[58] and BERTopic[19]. The code is available from the corresponding author upon reasonable request. SAS Enterprise Miner 15.1[59]: High-performance procedures were used for machine learning evaluation. SAS 9.4[60] was used for all other statistical analysis. SAS is a global software; we did not build the custom code for machine learning evaluation.

Received: 12 May 2024; Accepted: 10 December 2024;
Published online: 20 December 2024

References

1. Agha, R. A., Fowler, A. J. & Sevdalis, N. The role of non-technical skills in surgery. *Ann. Med. Surg.* **4**, 422–427 (2015).
2. Bonrath, E. M., Dedy, N. J., Gordon, L. E. & Grantcharov, T. P. Comprehensive surgical coaching enhances surgical skill in the operating room. *Ann. Surg.* **262**, 205–212 (2015).
3. Ma, R. et al. Tailored feedback based on clinically relevant performance metrics expedites the acquisition of robotic suturing skills—an unblinded pilot randomized controlled trial. *J. Urol.* **208**, 414–424 (2022).
4. Haglund, M. M. et al. The surgical autonomy program: a pilot study of social learning theory applied to competency-based neurosurgical education. *Neurosurgery* **88**, E345–E350 (2021).
5. Hauge, L. S., Wanzek, J. A. & Godellas, C. The reliability of an instrument for identifying and quantifying surgeons' teaching in the operating room. *Am. J. Surg.* **181**, 333–337 (2001).
6. Blom, E. et al. Analysis of verbal communication during teaching in the operating room and the potentials for surgical training. *Surg. Endosc.* **21**, 1560–1566 (2007).
7. Wong, E. Y. et al. Development of a classification system for live surgical feedback. *JAMA Netw. Open* **6**, e2320702 (2023).
8. Kocielnik, R. et al. Deep multimodal fusion for surgical feedback classification. In *Machine Learning for Health (ML4H)*, 256–267 (PMLR, 2023).
9. Laca, J. A. et al. Using real-time feedback to improve surgical performance on a robotic tissue dissection task. *Eur. Urol. Open Sci.* **46**, 15–21 (2022).
10. Chan, J. Y.-C. & Mazzocco, M. M. Integrating qualitative and quantitative methods to develop a comprehensive coding manual: Measuring attention to mathematics in play contexts. *Methods Psychol.* **4**, 100044 (2021).
11. Chinh, B., Zade, H., Ganji, A. & Aragon, C. Ways of qualitative coding: a case study of four strategies for resolving disagreements. In *Proc. Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–6 (2019).
12. Baralt, M. Coding Qualitative Data. *Research Methods in Second Language Acquisition: A Practical Guide* 222–244 (Wiley, 2011).
13. McCulloch, P. et al. The effects of aviation-style non-technical skills training on technical performance and outcome in the operating theatre. *BMJ Qual. Saf.* **18**, 109–115 (2009).
14. Helmreich, R. L. & Schaefer, H.-G. Team performance in the operating room. In *Human Error in Medicine*, 225–254 (CRC Press, 2018).
15. SCHWARTZ, R. W. et al. Undergraduate surgical education for the twenty-first century. *Ann. Surg.* **216**, 639–647 (1992).
16. Chen, N.-C., Drouhard, M., Kocielnik, R., Suh, J. & Aragon, C. R. Using machine learning to support qualitative coding in social science: Shifting the focus to ambiguity. *ACM Trans. Interact. Intell. Syst. (TiiS)* **8**, 1–20 (2018).
17. Reverberi, C. et al. Experimental evidence of effective human–AI collaboration in medical decision-making. *Sci. Rep.* **12**, 14952 (2022).

18. Kiyasseh, D. et al. A multi-institutional study using artificial intelligence to provide reliable and fair feedback to surgeons. *Commun. Med.* **3**, 42 (2023).
19. Grootendorst, M. Bertopic: neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794* (2022).
20. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003).
21. Teh, Y., Jordan, M., Beal, M. & Blei, D. Sharing clusters among related groups: hierarchical Dirichlet processes. *Adv. Neural Inf. Process. Syst.* **17**, 1385–1392 (2004).
22. Lee, D. & Seung, H. S. Algorithms for non-negative matrix factorization. *Adv. Neural Inf. Process. Syst.* **13**, 556–562 (2000).
23. Egger, R. & Yu, J. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify Twitter posts. *Front. Sociol.* **7**, 886498 (2022).
24. Zhou, L., Pan, S., Wang, J. & Vasilakos, A. V. Machine learning on big data: opportunities and challenges. *Neurocomputing* **237**, 350–361 (2017).
25. Mendonça, M. & Figueira, Á. Topic extraction: Bertopic's insight into the 117th Congress's Twitterverse. *Informatics*, **11**, 8 (MDPI, 2024).
26. Thielmann, A., Weisser, C., Kneib, T. & Säfken, B. Coherence based document clustering. In *Proc. 17th International Conference on Semantic Computing (ICSC)*, 9–16 (IEEE, 2023).
27. Atkinson, R. B. et al. Real-time student feedback on the surgical learning environment: use of a mobile application. *J. Surg. Educ.* **80**, 817–825 (2023).
28. Inouye, D. A. et al. Assessing the efficacy of dissection gestures in robotic surgery. *J. Robot. Surg.* **17**, 597–603 (2023).
29. Ma, R. et al. Surgical gestures as a method to quantify surgical performance and predict patient outcomes. *NPJ Digit. Med.* **5**, 187 (2022).
30. Tripathy, J. K. et al. Comprehensive analysis of embeddings and pre-training in NLP. *Comput. Sci. Rev.* **42**, 100433 (2021).
31. Liu, T., Yu, H. & Blair, R. H. Stability estimation for unsupervised clustering: a review. *Wiley Interdiscip. Rev. Comput. Stat.* **14**, e1575 (2022).
32. Xu, P., Zhu, X. & Clifton, D. A. Multimodal learning with transformers: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* (2023).
33. Kocielnik, R., Amershi, S. & Bennett, P. N. Will you accept an imperfect AI? Exploring designs for adjusting end-user expectations of AI systems. In *Proc. CHI Conference on Human Factors in Computing Systems*, 1–14 (2019).
34. Haque, T. F. et al. An assessment tool to provide targeted feedback to robotic surgical trainees: development and validation of the end-to-end assessment of suturing expertise (ease). *Urol. Pract.* **9**, 532–539 (2022).
35. Park, S. J. et al. Clinical desire for an artificial intelligence-based surgical assistant system: electronic survey-based study. *JMIR Med. Inform.* **8**, e17647 (2020).
36. Freschi, C. et al. Technical review of the da Vinci surgical telemanipulator. *Int. J. Med. Robot. Comput. Assist. Surg.* **9**, 396–406 (2013).
37. Jelodar, H. et al. Latent Dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimed. tools Appl.* **78**, 15169–15211 (2019).
38. Suryadjaja, P. S. & Mandala, R. Improving the performance of the extractive text summarization by a novel topic modeling and sentence embedding technique using sbert. In *Proc. 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, 1–6 (IEEE, 2021).
39. Bouma, G. Normalized (pointwise) mutual information in collocation extraction. *Proc. GSCL* **30**, 31–40 (2009).
40. Lau, J. H., Newman, D. & Baldwin, T. Machine reading tea leaves: automatically evaluating topic coherence and topic model quality. In: *Proc. 14th Conference of the European Chapter of the Association for Computational Linguistics*, 530–539 (2014).
41. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: pre-training of deep bidirectional transformers for language understanding. In *Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186 (2019).
42. Reimers, N. & Gurevych, I. Sentence-bert: sentence embeddings using siamese bert-networks. In: *Proc. Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992 (2019).
43. Wang, W. et al. Minilm: deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Adv. Neural Inf. Process. Syst.* **33**, 5776–5788 (2020).
44. McInnes, L., Healy, J., Saul, N. & Großberger, L. Umap: uniform manifold approximation and projection. *J. Open Source Softw.* **3**, 861 (2018).
45. Campello, R. J., Moulavi, D. & Sander, J. Density-based clustering based on hierarchical density estimates. In *Proc. Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 160–172 (Springer, 2013).
46. Stewart, G. & Al-Khassaweneh, M. An implementation of the hdbscan* clustering algorithm. *Appl. Sci.* **12**, 2405 (2022).
47. Wang, Y., Yu, S., Gu, Y. & Shun, J. Fast parallel algorithms for Euclidean minimum spanning tree and hierarchical spatial clustering. In: *Proc. International Conference on Management of Data*, 1982–1995 (2021).
48. McInnes, L. & Healy, J. Accelerated hierarchical density-based clustering. In: *Proc. International Conference on Data Mining Workshops (ICDMW)*, 33–42 (IEEE, 2017).
49. Grootendorst, M. Maartengr/ctfidf: creating class-based tf-idf matrices. https://github.com/MaartenGr/cTFIDF. Accessed 25 Mar 2024 (2020).
50. Grootendorst, M. 6a. representation models - bertopic. https://maartengr.github.io/BERTopic/getting_started/representation/representation.html. Accessed 18 Mar 2024 (2024).
51. Issa, B., Jasser, M. B., Chua, H. N. & Hamzah, M. A comparative study on embedding models for keyword extraction using keybert method. In *Proc.13th International Conference on System Engineering and Technology (ICSET)*, 40–45 (IEEE, 2023).
52. Chaturvedi, S. & Shweta, R. Evaluation of inter-rater agreement and inter-rater reliability for observational data: an overview of concepts and methods. *J. Indian Acad. Appl. Psychol.* **41**, 20–27 (2015).
53. King, G. & Zeng, L. Logistic regression in rare events data. *Political Anal.* **9**, 137–163 (2001).
54. Sun, X. & Xu, W. Fast implementation of Delong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Process. Lett.* **21**, 1389–1393 (2014).
55. Hastie, T., Tibshirani, R., Friedman, J. H. & Friedman, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, **2** (Springer, 2009).
56. Bates, D. et al. Package 'lme4'. *Convergence* **12**, 2 (2015).
57. Thompson, C. G., Kim, R. S., Aloe, A. M. & Becker, B. J. Extracting the variance inflation factor and other multicollinearity diagnostics from typical regression results. *Basic Appl. Soc. Psychol.* **39**, 81–90 (2017).
58. Paszke, A. et al. Pytorch: an imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32**, 8026–8037 (2019).
59. Inc., S. I. Sas help center: Sas®enterprise miner™ 15.1: reference help. https://documentation.sas.com/doc/en/emref/15.1/titlepage.htm. Accessed 2 Oct 2024 (2020).
60. Inc., S. I. Sas 9.4 software overview for the customer. https://support.sas.com/software/94/. Accessed 2 Oct 2024 (2021).

Acknowledgements

This study was supported in part by the National Cancer Institute under Award Numbers R01CA251579 and R01CA273031. The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Author contributions

R.K.: conceptualization, methodology, evaluation, experimental analysis and visualization, writing original draft, review, and editing. S.Y.C.: evaluation, experimental analysis, writing review, and editing. C.H.Y., R.M., E.Y.W., T.N.C.: data curation and evaluation. J.E.K., P.W., J.H., and U.G.: conceptualization, supervision, writing review, and editing. A.J.H. and A.A.: conceptualization, funding acquisition, supervision, writing review, and editing. All authors have read and approved the manuscript.

Competing interests

R.K., C.H.Y., R.M., S.Y.C., E.Y.W., T.N.C., J.E.K., P.W., J.H., U.G., A.A. declare no competing interests. A.J.H. declares no competing interests, but reports financial disclosures with Intuitive Surgical, Inc. and Teleflex, Inc.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-024-01383-3.

Correspondence and requests for materials should be addressed to Andrew J. Hung.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.