STROBE Statement—checklist of items that should be included in reports of observational studies

	Item No.	Recommendation	Page No.	Relevant text from manuscript
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	2	- In this prospective observational study, 666 audiovisual recordings from 100 patients were categorized as no pain, mild pain, or moderate-to-severe pain.
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	2	- PAINAD-supervised audiovisual models demonstrated promising internal ordinal performance in classifying observable postoperative pain behaviors.
Introduction				
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	4	- Pain is inherently subjective, and patient self-report such as the Numeric Rating Scale remains the preferred approach to assessing postoperative pain when reliable communication is possible. In the postanesthesia care unit, however, residual anesthesia, sedation, disorientation, cognitive impairment, or communication difficulty may temporarily make self-reported measures unreliable or unavailable.

Objectives	3	State specific objectives, including any prespecified hypotheses	5	- As a step toward passive assessment of postoperative pain behavior when self-report is unreliable, we prospectively developed and internally validated audiovisual AI models using PAINAD as the primary supervision target in PACU patients. - We hypothesized that PAINAD-defined pain behaviors could be classified from short audiovisual recordings; we additionally explored whether incorporating temporal and auditory information improved classification performance.
Methods				
Study design	4	Present key elements of study design early in the paper	6	- This single-center prospective observational study was conducted at Asan Medical Center, a tertiary hospital in Seoul, Republic of Korea, from 24 February to 8 May 2025.
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	6-7	- This single-center prospective observational study was conducted at Asan Medical Center, a tertiary hospital in Seoul, Republic of Korea, from 24 February to 8 May 2025.

				- Each participant underwent a preoperative baseline assessment in the general ward, followed by repeated postoperative assessments in the PACU. PACU assessments were conducted upon admission, before and after analgesic administration, and immediately prior to discharge.
Participants	6	(*a*) *Cohort study*—Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up *Case-control study*—Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls *Cross-sectional study*—Give the eligibility criteria, and the sources and methods of selection of participants	6-7	- Adult patients undergoing urological surgery under general anesthesia were recruited through convenience sampling either the day before or the day of surgery. Eligible patients were between 19 and 80 years old, recovered in the PACU after surgery, were able to understand the NRS and communicate a self-reported pain score, and were subsequently transferred from the PACU to a general ward. Exclusion criteria included facial deformities or prior facial surgeries that could interfere with facial expression analysis, neurological or psychiatric conditions impairing communication or facial expression, and physical

				impairments significantly affecting observable movement. - Each participant underwent a preoperative baseline assessment in the general ward, followed by repeated postoperative assessments in the PACU. PACU assessments were conducted upon admission, before and after analgesic administration, and immediately prior to discharge.
		(*b*) *Cohort study*—For matched studies, give matching criteria and number of exposed and unexposed *Case-control study*—For matched studies, give matching criteria and the number of controls per case	N/A	Not applicable. This was not a matched cohort study, and no exposed and unexposed groups were defined.
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	7, 8, 10, 11	- For model development, NRS and PAINAD scores were categorized as no pain (0), mild pain (1–3), and moderate-to-severe pain (4–10). - PAINAD-supervised models constituted the main model-development analysis. Parallel NRS-supervised models were developed using identical architecture, patient partitions, preprocessing, and training

				procedures as a secondary exploratory clinical benchmark. - The principal performance metrics were balanced accuracy and quadratic weighted kappa (QWK). - Secondary performance metrics included class-specific recall and F1 score, mean absolute error (MAE), and row-normalized confusion matrices.
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	7	- At each assessment, a 10–20-second audiovisual recording was obtained using an HDR-PJ340 camcorder (Sony, Tokyo, Japan), capturing the patient's face and upper body within the frame. Audio was recorded simultaneously using the camcorder's built-in microphone. Observable pain behavior was evaluated using the Korean version of the PAINAD (PAINAD-K), after which the patient was asked to report pain intensity using the NRS. Hereafter, PAINAD refers to the PAINAD-K as implemented in this study. Two PACU nurses independently

				reviewed the recordings while blinded to the corresponding NRS scores. Any disagreements were resolved by consensus.
Bias	9	Describe any efforts to address potential sources of bias	7, 10	- Two PACU nurses independently reviewed the recordings while blinded to the corresponding NRS scores. Any disagreements were resolved by consensus. - To prevent information leakage between the training and evaluation datasets, all recordings from the same patient were assigned to the same fold. - Class imbalance in the training data was addressed using inverse-frequency weighted random sampling.
Study size	10	Explain how the study size was arrived at	6	- No formal hypothesis-based sample size calculation was performed; the study size was determined by the number of eligible participants enrolled during the predefined recruitment period.

Continued on next page

Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	7, 11	- For model development, NRS and PAINAD scores were categorized as no pain (0), mild pain (1–3), and moderate-to-severe pain (4–10). A score of 4 was selected as the boundary between mild pain and pain of at least moderate intensity, consistent with commonly used severity interpretations for both the NRS and PAINAD. Moderate and severe scores were combined into a single category representing pain of at least moderate intensity. - For MAE calculation, no pain, mild pain, and moderate-to-severe pain were coded as 0, 1, and 2, respectively.
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding	10, 11, 12	- All statistical tests were two-sided, and a Holm-adjusted p-value of less than 0.05 was considered statistically significant. Analyses were performed using Python 3.11 with NumPy 1.26.1, Statsmodels 0.14.6, and scikit-learn 1.7.2.
		(*b*) Describe any methods used to examine subgroups and interactions	N/A	- Not applicable. No subgroup or interaction analyses were performed.
		(*c*) Explain how missing data were addressed	8	- If a cropped facial image was unavailable, the corresponding original frame was resized and retained

		(*d*) *Cohort study*—If applicable, explain how loss to follow-up was addressed *Case-control study*—If applicable, explain how matching of cases and controls was addressed *Cross-sectional study*—If applicable, describe analytical methods taking account of sampling strategy	N/A	- No follow-up beyond PACU discharge was planned. Participant disposition through completion of data collection is reported in the Results.
		(*e*) Describe any sensitivity analyses	N/A	- Not applicable. No sensitivity analyses were performed.
Results				
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	13	- Of the 104 patients who provided written informed consent, four were excluded due to surgery cancellation (n=1) or postoperative transfer to the intensive care unit (n=3). The final cohort comprised 100 patients who contributed 666 audiovisual recordings, each paired with NRS and PAINAD assessments.
		(b) Give reasons for non-participation at each stage	13	- four were excluded due to surgery cancellation (n=1) or postoperative transfer to the intensive care unit (n=3).
		(c) Consider use of a flow diagram	N/A	- No flow diagram was included; participant flow from consent to analysis is reported in the text.
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	13, 24 (Table 1)	- The mean age was 63.1 years (SD, 8.8), and 86 patients (86.0%) were male. The mean body mass index was 25.4 kg/m² (SD, 3.1). The mean durations of anesthesia, surgery, and PACU stay were 153.8

				minutes (SD, 46.2), 103.1 minutes (SD, 42.7), and 65.4 minutes (SD, 13.6), respectively.
		(b) Indicate number of participants with missing data for each variable of interest	24 (Table 1)	- No data were missing for the participant characteristics reported in this table.
		(c) *Cohort study*—Summarise follow-up time (eg, average and total amount)	6, 7, 13, 24 (Table 1)	- Each participant underwent a preoperative baseline assessment in the general ward, followed by repeated postoperative assessments in the PACU. PACU assessments were conducted upon admission, before and after analgesic administration, and immediately prior to discharge. The number of assessments varied depending on analgesic administration and the patient's clinical course, with each participant contributing between 3 and 13 recordings. - The mean durations of anesthesia, surgery, and PACU stay were 153.8 minutes (SD, 46.2), 103.1 minutes (SD, 42.7), and 65.4 minutes (SD, 13.6), respectively.
Outcome data	15*	Cohort study—Report numbers of outcome events or summary measures over time	13	- Among the 666 recordings, 124 (18.6%) were categorized as no pain, 306 (45.9%) as mild pain, and 236 (35.4%) as moderate-to-severe pain according to PAINAD.
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision	13	- Balanced accuracy was 0.785 for

	(eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included		the image-only model, 0.808 for the video-only model, and 0.817 for the video-plus-audio model. The corresponding quadratic weighted kappa values were 0.760, 0.793, and 0.801, respectively.
	(*b*) Report category boundaries when continuous variables were categorized	7	- For model development, NRS and PAINAD scores were categorized as no pain (0), mild pain (1–3), and moderate-to-severe pain (4–10).
	(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	N/A	N/A

Continued on next page

Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	14-15	- Exact categorical agreement between the NRS and PAINAD was 36.5%. Agreement within one ordinal category, including exact agreement, was observed in 82.7% of recordings, while two-category disagreements occurred in 17.3%. The quadratic weighted kappa between the NRS and PAINAD was 0.038, indicating minimal ordinal agreement. - The moderate-to-severe pain error analysis was based on predictions from the video-only models. False alerts for moderate-to-severe pain occurred in 56 of 213 recordings labeled as no or mild pain (26.3%) with NRS supervision and in 48 of 430 recordings (11.2%) with PAINAD supervision. The risk difference, calculated as the NRS-supervised rate minus the PAINAD-supervised rate, was 15.1 percentage points (95% CI, 7.7–22.0; Holm-adjusted $P < 0.001$). Moderate-to-severe pain was missed in 133 of 453 recordings (29.4%) with NRS supervision and in 53 of 236 recordings (22.5%) with PAINAD supervision. The corresponding risk difference was 6.9 percentage points (95% CI, −1.8

				to 15.3; Holm-adjusted P = 0.112), indicating that the difference was not statistically significant.
Discussion				
Key results	18	Summarise key results with reference to study objectives	16	- In this prospective study, we developed and internally validated PAINAD-supervised audiovisual AI models for behavior-based pain assessment in the PACU. The models demonstrated consistent ordinal classification performance across image-only, video-only, and video-plus-audio inputs, with the highest balanced accuracy and quadratic weighted kappa observed in the video-plus-audio configuration.
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	17, 18	- This study has several limitations. First, it was conducted at a single center using convenience sampling, and the cohort consisted predominantly of male patients undergoing urological surgery, which may limit the generalizability to more diverse surgical populations. Second, although the proposed approach is designed to operate without active patient self-reporting, all participants in the development cohort were able to understand the NRS and

communicate a pain score. Therefore, model performance in patients who are drowsy, delirious, cognitively impaired, or otherwise unable to communicate was not directly evaluated. Third, the PAINAD was originally developed for patients with advanced dementia 14. Although its Korean version has been evaluated in older PACU patients unable to communicate 18, its use as a supervisory target in a broader postoperative population requires further validation. Moreover, PAINAD remains an observer-rated behavioral scale rather than an objective ground truth for pain. Fourth, the dataset included a preoperative baseline recording obtained in the general ward in addition to repeated PACU recordings; contextual differences between recording environments may have influenced model predictions. Fifth, although patient-level cross-validation prevented information leakage, performance metrics were calculated at the recording level, meaning that patients who contributed more recordings had a greater influence on the estimates. Sixth, categorizing

				NRS and PAINAD scores into three ordinal classes may have reduced the information contained in the original scores 27. The brief audiovisual inputs and facial cropping may also have incompletely represented certain PAINAD domains, particularly consolability and broader body language. Finally, the models underwent internal validation only; external validation in independent institutions, more diverse patient populations, and different audiovisual recording environments is required before clinical implementation.
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	16	- PAINAD was selected not because it represents an objective or definitive measure of pain, but because it provides a structured, observer-rated assessment of pain-related behavior. Several PAINAD domains—including facial expression, negative vocalization, breathing, and body language—conceptually correspond to information that can be captured through video and audio. In the present study, PAINAD therefore served as a behavioral supervision framework for model development

				rather than as a substitute for routine patient-reported pain assessment.
Generalisability	21	Discuss the generalisability (external validity) of the study results	17	- First, it was conducted at a single center using convenience sampling, and the cohort consisted predominantly of male patients undergoing urological surgery, which may limit the generalizability to more diverse surgical populations. - Therefore, model performance in patients who are drowsy, delirious, cognitively impaired, or otherwise unable to communicate was not directly evaluated.
Other information				
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	19	- This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2026-25504160) to S.H.K. The funder had no role in the study or publication decision.

*Give information separately for cases and controls in case-control studies and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at www.strobe-statement.org.