Supplemental Information

EXPERIMENTAL PROCEDURES

Video Analysis

Timestamps were manually identified for the beginning and end of the 5-minute ongoing pain event and 10-second transient pain windows. Ongoing and transient pain video segments from each study visit were extracted from recorded visit videos based on these timestamps.

Videos were analyzed by CERT.[1] CERT automatically detects faces and codes each frame of video for the presence and movement intensity of a set of AUs from FACS. CERT operates in real time, with an optimal pixel resolution of 96 × 96 face size and frontal view face angles of ±15° of frontal in yaw, pitch, and roll. Fourteen AUs previously related to pain were selected for analysis. The following were selected for analysis: brow lower (AU 4), cheek raiser (AU 6), lid tightener (AU 7), nose wrinkle (AU 9), upper lip raiser (AU 10), lip corner puller (AU 12), lips part (AU 25), jaw drop (AU 26), mouth stretch (AU 27), and eye closure (AU 43) (for reviews see Prkachin[2,3] and Larochette et al[4]). In addition, 3 dimensions of head pose (yaw, pitch, and roll) were tracked to measure if they were pain-related head movements.[5] Measured AUs are listed in Table 1 of the main text.

Three statistics (mean, 75th percentile, and 25th percentile) for each AU were computed across each pain event probe (ongoing and transient). For each event, 14 AUs × 3 statistics were concatenated into a 42-dimensional vector (the facial behavior data vector). Video samples were excluded when <10% of frames did not contain a detected face because of occlusion or nonfrontal head orientation. Evaluated video samples were $n = 150$ for ongoing pain and $n = 297$ for transient pain. Three samples were excluded from the transient pain condition because of insufficient frames with a detected face. Nondetections were typically due to occlusion by the researcher.

ML Methods for Estimating Pain From Facial Expressions

We used ML methods to generate 2 types of pain estimation models: (1) binary classification of clinically significant pain versus no pain, and (2) pain-intensity estimation across a scale of 0 to 10.

For the binary classification of clinically significant pain versus no pain, a logistic regression model was used to learn the mapping from the facial behavior measures to the binary pain labels. The input to the model consisted of the 42-dimensional AU data vector. The training output consisted of 1 for pain and 0 for no pain, and after training, returned a classification probability for each test sample.

For pain-intensity estimation, a linear regression model was used to learn a linear combination of the measures in the AU data vector to predict pain-intensity self-ratings. Both regression models used L1 regularization, which yields a sparse regression model and reduces model overfitting.[6] Overfitting was primarily addressed through cross-validation, as described in the following sections.

Separate models were trained for ongoing and transient pain.

Statistical Evaluation of the Performance of ML-Generated Pain Estimation Models

Performance of the trained binary pain classification and pain severity estimation models were evaluated on novel data using cross-validation.[7] Cross-validation was introduced by Tukey[7] and is a standard method in ML for estimating model performance on new subjects who were not used to develop the model. AUC and Cohen's κ were used to measure the performance of the binary pain classification models. Pearson correlations (both within and across subjects) were computed to evaluate pain-intensity estimation models.

Evaluation of the Binary Pain Estimation Model

Performance of the trained binary pain classification model was evaluated on data from new participants by using cross-validation.[7] The cross-validation procedure for the binary model was as follows. The data were divided into 10 partitions of 5 participants' data each. The model parameters were computed by using data from 9 partitions, and the 10th was held out for testing. Model parameters were then deleted, and the process was repeated, each time holding out a different 1 of the 10 partitions for testing. Performance was then aggregated over the 10 tests. This procedure measures generalization to novel participants, where data from a test participant is never used for training. The number of training participants was 45 and the number of test participants was 50.

The following 2 metrics were used to measure the performance of the binary pain classification model: (1) AUC, and (2) Cohen's κ. AUC ranges from 0.5 (chance) to 1 (perfect discrimination) and can be interpreted as the probability of the system ranking an instance from the positive class higher than an instance from the negative class. Cohen's κ is

a standard method form experimental psychology to measure categorical agreement between raters that takes into account chance agreements. The decision threshold for calculating κ was 0.5.

Evaluation of the Pain-Intensity Estimation Model

Performance of the pain-intensity model was evaluated by using leave-1-person-out cross-validation. (Withholding 1 participant at a time enabled evaluations of within-participant correlations. For the binary pain classification, we used larger partitions of 5 participants because the AUC metric requires data from both classes to return a nondegenerate value.) This is the same procedure as used to measure performance of the binary pain estimation model, except there were 50 partitions, and 1 participant was held out for testing in each partition. In this cross-validation procedure, the model was developed 50 times, each time leaving out 1 of the 50 participants for testing, and the previous model was deleted between tests. Test performance was then aggregated over the 50 tests.

We computed the correlation between the reported labels and the ML prediction in the following 2 ways: (1) within-participant correlation, and (2) across-participant (overall) correlation. Within-participant correlation was computed as the correlation for 1 participant's data at a time, and then the mean correlation coefficient was computed individually across all participants. This removed within-participant differences in AU mean and scale between study conditions (ie, normalized data across observations within each participant). Across-participant correlation (overall correlation) was calculated by concatenating the test outputs for all test participants before computing correlations with self-report labels. We used Pearson correlation coefficients as our correlation measure.

ML Model Versus Current Methods of By-Proxy Pain Assessment

We evaluated performance of the ML-generated pain estimation models in comparison with human observers' (parents' and nurses') pain ratings. Because inpatient nurse ratings were not available for the final study visit, which was performed after patient discharge, and because most patients with laparoscopic appendectomy are no longer in pain by the time of a follow-up visit,[8–10] which occurred at least 12 days postoperatively in our study, all nurse ratings of the final visit were assumed to be 0. Median (minimum, maximum) pain ratings by children at the final visit were 0 (0, 0) for ongoing pain and 0 (0, 2) for transient pain.

For binary decisions on presence/absence of clinically significant pain, proxy ratings were treated as agreeing with child ratings when proxy ratings were ≥ 4 on pain trials (child rating ≥ 4), and < 4 for no-pain trials (child rating $= 0$). Categorical agreement was measured with Cohen's κ. Signal detection (the ability to separate the pain and no-pain trials regardless of decision threshold) was measured by AUC.

For pain-intensity estimation, Pearson correlations were computed between observer ratings and child ratings. Comparisons between overall correlations were made by using Fisher's z-tests. Because parents and nurses had access to information about elapsed time since surgery when making their judgments, performance of the model was compared with human observers 2 ways: once using a model trained only on the facial AUs, and again using a model that included time since surgery as a predictive variable along with facial AUs.

REFERENCES

1. Littlewort G, Whitehill J, Wu T, et al. The computer expression recognition toolbox (CERT). *Proc IEEE Conference on Automatic Face and Gesture Recognition*; 2011; 299–305.
2. Prkachin KM. The consistency of facial expressions of pain: a comparison across modalities. *Pain*. 1992;51(3):297–306
3. Prkachin KM. Assessing pain by facial expression: facial expression as nexus. *Pain Res Manag*. 2009;14(1):53–58
4. Larochette AC, Chambers CT, Craig KD. Genuine, suppressed and faked facial expressions of pain in children. *Pain*. 2006;126(1-3):64–71
5. Lucey P, Cohn JF, Matthews I, et al. Automatically detecting pain in video through facial action units. *IEEE Trans Syst Man Cybern B Cybern*. 2011;41(3):664–674
6. Murphy KP. *Machine Learning, a Probabilistic Perspective*. Cambridge, MA: MIT Press; 2012
7. Tukey JW. Bias and confidence in not-quite large samples. *Ann Math Stat*. 1958;29:614
8. Lejus C, Delile L, Plattner V, et al. Randomized, single-blinded trial of laparoscopic versus open appendectomy in children: effects on postoperative analgesia. *Anesthesiology*. 1996; 84(4):801–806
9. Miyauchi Y, Sato M, Hattori K. Comparison of postoperative pain between single-incision and conventional laparoscopic appendectomy in children. *Asian J Endosc Surg*. 2014;7(3):237–240
10. Paya K, Fakhari M, Rauhofer U, Felberbauer FX, Rebhandl W, Horcher E. Open versus laparoscopic appendectomy in children: a comparison of complications. *JSLS*. 2000; 4(2):121–124