Research Article

MOLECULAR PAIN

Molecular Pain
Volume 14: 1–9

DOI: 10.1177/1744806918763658
journals.sagepub.com/home/mpx
SAGE

A deep neural network to assess spontaneous pain from mouse facial expressions

Alexander H Tuttle[1], Mark J Molinaro[1], Jasmine F Jethwa[1], Susana G Sotocinal[2], Juan C Prieto[3], Martin A Styner[3], Jeffrey S Mogil[2] and Mark J Zylka[1]

Abstract

Grimace scales quantify characteristic facial expressions associated with spontaneous pain in rodents and other mammals. However, these scales have not been widely adopted largely because of the time and effort required for highly trained humans to manually score the images. Convoluted neural networks were recently developed that distinguish individual humans and objects in images. Here, we trained one of these networks, the InceptionV3 convolutional neural net, with a large set of human-scored mouse images. Output consists of a binary pain/no-pain assessment and a confidence score. Our automated Mouse Grimace Scale integrates these two outputs and is highly accurate (94%) at assessing the presence of pain in mice across different experimental assays. In addition, we used a novel set of "pain" and "no pain" images to show that automated Mouse Grimace Scale scores are highly correlated with human scores (Pearson's $r = 0.75$). Moreover, the automated Mouse Grimace Scale classified a greater proportion of images as "pain" following laparotomy surgery when compared to animals receiving a sham surgery or a post-surgical analgesic. Together, these findings suggest that the automated Mouse Grimace Scale can eliminate the need for tedious human scoring of images and provide an objective and rapid way to quantify spontaneous pain and pain relief in mice.

Keywords

Spontaneous pain, animal models, machine learning, facial expressions

Date Received: 13 December 2017; revised 30 January 2018; accepted: 5 February 2018

Introduction

Despite a persistent demand for new analgesics with fewer adverse side effects, successes in translating basic discoveries into clinical treatments have been limited. Standard tests of animal hypersensitivity and allodynia do not measure spontaneous pain—the main symptom associated with chronic pain in humans.[1–3] In an attempt to address this mismatch, assays that monitor ongoing spontaneous pain in rodents were developed, although many of these assays have uncertain predictive value (see Mogil[4]). Moreover, a significant proportion of spontaneous pain assays are prohibitively labor-intensive to score and require highly trained personnel.

The Mouse and Rat Grimace Scales (MGS and RGS) measure characteristic changes in facial expressions that associate with pain in rodents.[5,6] To make use of these scales, close-up video is obtained from rodents, individual images containing the face are manually extracted, and the images are scored for the presence or absence of grimacing, as defined by changes in facial musculature. This rodent scale is based on human facial coding

[1]Department of Cell Biology and Physiology, UNC Neuroscience Center, The University of North Carolina, Chapel Hill, NC, USA
[2]Department of Psychology, Alan Edwards Centre for Research on Pain, McGill University, Montreal, QC, Canada
[3]Department of Psychiatry, Carolina Institute for Developmental Disabilities, The University of North Carolina, Chapel Hill, NC, USA

Corresponding Author:
Mark J Zylka, Department of Cell Biology and Physiology, UNC Neuroscience Center, The University of North Carolina, Chapel Hill, NC 27599, USA.
Email: zylka@med.unc.edu

scales.[7,8] Facial grimace scales were subsequently adapted to detect spontaneous pain in additional species including sheep,[9,10] horses,[11,12] rabbits,[13] cattle,[14] pigs,[15] and cats.[16] Researchers used this scoring system to evaluate analgesic efficacy in mice[17] and rats,[18,19,20,21] as well as evaluate post-surgical spontaneous pain in rodents.[22,23] To eliminate the need for manual image extraction, the Rodent Face Finder® software was developed. This software identifies and extracts video frames when mice or rats are facing the camera.[6,17] However, this software does not eliminate two of the most time-consuming and subjective aspects of using grimace scales—(1) training lab personnel to score images and (2) scoring the large numbers of images that are generated as part of each experiment. Widespread adoption of grimace scales has been limited as a result, despite the fact these scales are reproducible between labs and accurately predict analgesic efficacy.

We reasoned that a machine learning model could be used to eliminate the need for humans to score images. To test this possibility, we trained and optimized a convolutional neural network based on Google's InceptionV3 model[24] to analyze and classify a large number of "pain" and "no pain" face images from outbred CD-1 mice. Similar machine-learning models were used to predict human self-reported pain[25,26] and to classify changes in mouse behavior at the sub-second level.[27] Our *a*utomated *M*ouse *G*rimace *S*cale (aMGS) was validated on novel image sets generated in two different labs (Zylka and Mogil) and accuracy was compared to human scoring. In contrast to previous efforts,[28] we show a consistently high accuracy across large data sets that include high-definition color images procured from multiple pain tests. Finally, we evaluate the predictive validity of our model using an assay of post-operative pain (laparotomy) and relief of pain with carprofen, a non-steroidal anti-inflammatory drug. This analgesic was previously found to reduce pain-induced grimacing in mice and rats when images were scored by humans.[17,21]

Methods

Animals

All animal experiments were approved by the Institutional Animal Care and Use Committee of the University of North Carolina at Chapel Hill and in accordance with NIH guidelines. Archived images from the Mogil lab were of mice previously tested in compliance with the McGill Downtown Animal Care and Use committee and were consistent with Canadian Council on Animal Care guidelines. Images of mice were obtained from video or still image archives kept at the Mogil lab at McGill University (Quebec, Canada) or recorded and processed at the University of North Carolina (Chapel Hill, NC, USA). All animals used in this study were CD-1® (ICR:Crl) mice (6–12 weeks old) and were purchased from Charles River Laboratories (Albany, NY; St. Constant, QC). Mice at UNC were housed in standard 7.5 in. × 11.5 in. × 5 in. polycarbonate cages with 1/4-in. corncob bedding (Bed-o'Cobs, Maumee, OH) in groups of five with same-sex littermates under a 12:12-h light:dark cycle (lights on at 07:00), in a temperature-controlled environment (20 ± 1°C), and with ad libitum access to food (Envigo Teklad 2920, Harlan Teklad) and tap water. Each animal underwent one surgery (and/or drug or anesthesia exposure). Mice received isoflurane during surgery and were sacrificed immediately after post-surgical observation. In cases of surgical complications, mice were immediately sacrificed without further testing. Roughly equal numbers of male and female mice were tested in each cohort. Neither main effects of sex nor interactions with sex were noted, so collapsed data are reported.

Compounds

Zymosan was obtained from Sigma-Aldrich (St. Louis, MO) and dissolved in physiological saline (5.0 mg/ml), and unilaterally injected (20 µl volume) into the intraarticular space around the ankle joint. Carprofen (Rimadyl®) was purchased from Zoetis, Inc. (Kalamazoo, MI) and administered to a subset of animals (50 mg/kg, *s.c.*) as described below.

Video capture

Data sets provided from the Mogil lab were generated by placing mice individually on a tabletop in cubicles (9 × 5 × 5 cm high) with two walls of transparent Plexiglas and two side walls of removable stainless steel. Two high-resolution (1920 × 1080) digital video cameras (High-Definition Handycam Camcorder, model HDR-CX100, Sony, San Jose, CA) were placed immediately outside both walls to maximize the opportunity for clear head shots (see Figure 1(a)). Video was taken for 30 min immediately before zymosan ankle injection (baseline) and for 30-min periods at various time points after injection.

To generate additional images for our training set, as well as validate the classification accuracy of the aMGS, additional mice were filmed using a second experimental assay. Mice were placed individually on the edge of a tabletop in a custom-built four-cubicle array (each cubicle measuring 5 × 12 × 6 cm high) with one back wall of transparent acrylic glass and two side walls of stainless steel (see Figure 1(b)). The fourth wall was open and positioned directly above the floor (1.1-m drop).

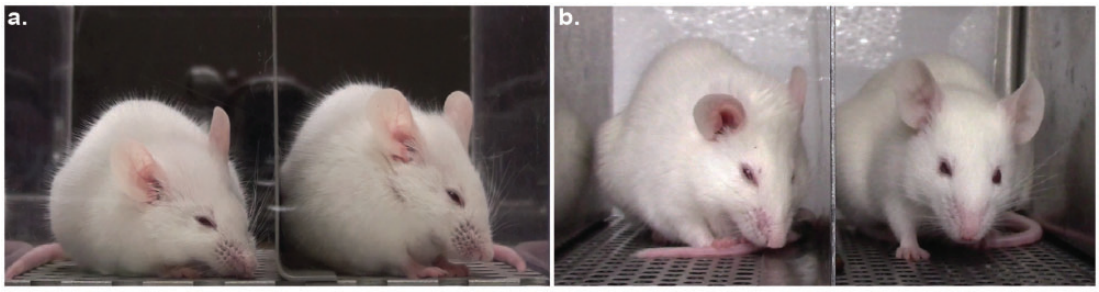

Figure 1. Setups used to capture continuous video footage of mice. Examples of mice grimacing in (a) the traditional recording setup with mice enclosed in Plexiglas boxes and (b) the new elevated cliff recording setup. Mice face towards the visual cliff for most of the recording session in the new setup, allowing the use of a single camera instead of two cameras. The camera also captures clearer images (without reflections) through air than through the Plexiglas partition.

We found that this arrangement encourages mice to look towards the visual cliff and hence face the camera. One camera was positioned 0.25 m away from the open side to capture video footage mice in the cubicle array. In this configuration, four mice can be recorded simultaneously using two high-resolution (1920×1080) digital video cameras (High-Definition Camcorder, model HFR700, Canon, Tokyo, Japan). After a 30-min habituation to the cubicle, mice were filmed for a 30-min baseline session immediately before surgery or inter-articular zymosan injection.

Nociceptive assays

We obtained scored mouse grimace image sets from the Mogil lab to create our initial training set. These images were generated as part of a previous study,[5] and represent data from multiple pain tests. We also used the intrarticular zymosan inflammation assay (ZYM) to supplement the initial training image set as well as generate a separate set of images to serve as a novel validation set (not used in the initial training). We used this second validation set to compare automated and human image classification. ZYM consisted of a single 20 μl injection of 5 mg/ml zymosan into the right ankle joint. Mice were allowed to recover for 3 h in their home cage after injection and before being placed back into our video setup for a 30-min habituation prior to filming. Mice were filmed for 1 h.

A post-surgical pain assay (laparotomy, LAP) is designed to mimic a ventral ovariectomy. We used the LAP test to further validate the aMGS as a second pain assay, as well as evaluate our system's ability to detect analgesia-related reduction in mouse grimacing. LAP surgeries were performed on isoflurane–oxygen-anesthetized mice by a single surgeon (AHT). After shaving and disinfection of the surgical site, a 1-cm midline incision was made by using a scalpel. Muscle layers and skin edges were closed with 6-0 non-absorbable braided silk suture and skin edges apposed by using tissue glue. Once recovered from anesthesia, mice were reintroduced to video cubicles to obtain post-surgical video recordings. Mice ($n = 64$) were placed into one of three groups: (1) surgery alone (laparotomy); (2) surgery plus carprofen (50 mg/kg, *s.c.*) during recovery (carprofen); (3) isoflurane anesthesia and surgery preparation only (sham surgery). All surgeries were performed at $13{:}00 \pm 1$ h. Cases of excessive bleeding or poor surgical outcome were excluded from testing ($n = 7$). Mice were allowed to recover for 15–20 min following surgery before being filmed for 1 h in the video rig.

Image generation and human scoring

Individual frames from video files were "grabbed" automatically by the Rodent Face Finder software, developed previously to assist in capturing mouse and rat face images from video.[6,17] This software detects frames, in an unbiased fashion, where eyes and ears are visible and image quality is not compromised by motion blurring. The Rodent Face Finder was set to grab one image for every 10 s of video in order to avoid oversampling extremely short observation periods. All images were output from the software and placed into Microsoft PowerPoint, with one image per slide. A PowerPoint macro (http://www.tusharmehta.com/powerpoint/randomslideshow/index.htm) was used to randomize the slide order. Identifications were removed to ensure that subsequent coding was performed in a blinded fashion by experienced human scorers.

Randomized and unlabeled photos were presented sequentially on a large, high-resolution computer monitor. For each photo, the scorer assigned a value of 0, 1, or 2 for each of the five MGS action units: (1) orbital tightening, (2) nose bulge, (3) cheek bulge, (4) ear position, and (5) whisker change. In each case, a score of 0 indicated high confidence by the scorer that the action unit was absent. A score of 1 indicated either high confidence of a moderate appearance of the action unit or equivocation over its presence or absence. A score of

2 indicated high confidence of a marked appearance of the action unit. The final MGS score was the summed score across the five action units (resulting in a maximum hypothetical score of 10 and minimum score of 0). All human scores were input into our initial training set for the automated scoring system.

The automated Mouse Grimace Scale

To systematically classify an individual mouse facial expression as "pain" or "no pain," the InceptionV3 convolutional neural net—one of the current standard convolutional neural net's for generalized image classification—was retrained on our own data set. We began by building the initial training set from a combination of previously published pain images (provided by the Mogil lab[5]) and supplemented by additional ZYM images generated in the Zylka lab. Altogether, our training set comprised 5,771 unique images: 2,444 "pain" and 3,327 "no pain." We then validated our model by combining unpublished ZYM images from the Mogil lab (scored by SGS) and adding annotated laparotomy images generated by the Zylka lab (scored by JJ). To determine inter-rater reliability, we calculated a concordance value based on MGS scores produced by our human raters on a subset of training images. We found our two MGS scorers to be in close agreement with one another on the same set of ZYM images (Cronbach's alpha = 0.89).

Initial training images were selected equally from published data sets as well as new experiments in order to maximize image heterogeneity. This was to ensure that our model would be able to classify pain faces across a variety of different testing parameters (including image background, image quality, rodent facial differences, and pain assays). We confirmed the absence of large systematic differences between "pain" and "no pain" images using a pixel-by-pixel structural similarity analysis on a subset of our training images (n = 12–16). Analysis revealed that within-group and between-group differences were statistically indistinguishable from one another.

To further anticipate future variations in image quality, a random selection of "pain" images from the training set were duplicated, programmatically altered by cropping (± up to 10% on each side), horizontal flipping, or brightness scaling by a factor of $N(1,0.2)$, and then added back into the "pain" image subset. By altering some of the pain images, we were also able to create an overall balanced training set (3,536 "pain" and 3,326 "no pain" images). Both the training and classification scripts utilized Google's TensorFlow (v1.0) library and were written in Python (v3.5). To tailor the pre-trained model for our specifications, we exploited the power of transfer learning to retrain only the final fully connected and softmax layers of the inception model on our own training image set.

To train the model, we split our balanced training set (comprising 6,862 total images) into training (80%), testing (10%), and validation (10%) subsets. After empirically testing various training hyper-parameters, we found a learning rate of 0.01 to be ideal. The training batch size was set to 100 images, and the batch size for both testing and validation sets remained unbounded in order to balance the stability of results between training runs. The model trained to converge after 7,500 iterations of 100-image training batches processed by our neural network.

The aMGS was designed to output a classification ("pain" or "no pain") and respective confidence rating (from 0.5 to 1.0). High-confidence images determined by a confidence rating above a specified threshold were retained and classified, while low-confidence images were discarded. This increased the model's accuracy while simultaneously serving as a quality control step to further identify and remove low-quality images. Model training and image classification was completed on standard notebook computers (Macbook Pro mid-2015, Palo Alto, CA, and Dell Inspiron 15 7000 series mid-2016, Round Rock, TX).

Laparotomy experiment

To test the predictive validity of our new model, we processed novel images obtained from animals undergoing a laparotomy (with or without analgesia) or sham surgery (control) using the aMGS. Our new model classified an image as either "pain" or "no pain" with an accompanying confidence score. High-confidence images were then selected (0.75 confidence or greater) and classified by our automated system. After identifying and selecting high-confidence images, videos that did not contain a minimum number of high-confidence images (mean image number = 1 image/minute) were excluded from analysis (n = 16–19). Spontaneous mouse pain was quantified by taking a percentage: % "pain" = ("pain" images/total images) × 100. To compare degree of spontaneous mouse pain following surgery versus baseline, a simple difference score was calculated: % pain(post-surgery) − % pain(baseline).

Statistics

All statistical analyses were performed using Systat v.13 (SPSS, Chicago, IL), with a criterion of $\alpha = 0.05$. To compare human and machine grimace scores, a linear regression analysis was performed between images that were rated by two coders (SGS or JJ) on the 11-point MGS compared with the confidence rating of the aMGS. To measure human inter-rater

concordance, Cronbach's α was calculated on coder scores from the same set of 275 sample mouse images. Human versus machine scores were compared using a simple linear regression; machine confidence intervals were compared to chance levels using one-sample *t*-tests followed by Bonferroni correction. For the laparotomy experiment, normality and homoscedasticity were confirmed using the Anderson-Darling and Levene tests, respectively, and parametric statistics were used in all cases. Baseline behavioral outliers were identified (Studentized residual > 3.0) and removed prior to final analysis (n = 2). Group data were analyzed by one-way analysis of variance (ANOVA), followed by Tukey's Test to determine group differences.

Results

The aMGS shows a high degree of internal accuracy

After training the aMGS, we reran the original dataset through the trained model and checked preliminary classification results. We found that when every training image was included for analysis, the model's calculated sensitivity (79%), specificity (87.2%), and accuracy (83%) were suboptimal as compared to experienced human coders. To increase the model's accuracy, we used the confidence interval output generated automatically by the aMGS to eliminate "ambiguous" images. Specifically, we restricted subsequent analysis to high-confidence (>0.75 confidence interval) images. After recalculation, the same metrics were significantly improved: sensitivity (90.5%), specificity (96.1%), and accuracy (93.2%) (Table 1).

We next validated the aMGS using a reserved subset (10%) of our initial image pool that was not used to train the model ("validation set"). After analyzing the entirety of the validation set, the aMGS achieved an accuracy of 84%. To increase the model's classification accuracy, we once again restricted the aMGS to analyze only high-confidence images from the validation set (>0.75 confidence interval). This increased the model's reported accuracy to 94%, which is comparable to what highly experienced human coders achieve.[5,17]

Table 1. Legend- Images assessed were from our initial training set. We found that restricting analysis to high confidence images (⩾0.75) yielded the highest degree of accuracy while maintaining a high number of quantifiable images (67% of total images from the training set). Human prediction values denote images determined to be "in pain" or "not in pain" by human assessment. Machine predictions denote images assessed by the aMGS.

	Human Prediction (Truth)		
Machine Prediction	Pain (images)	No pain	***Total***
Pain	2,159	85	***2,224***
No pain	226	2,107	***2,333***
Total	***2,385***	***2,192***	***4,577***

Sensitivity-90.5%
Specificity-96.1%
Accuracy-93.2%

Classification accuracy of the aMGS and human coders is comparable when presented with novel images

To further validate the aMGS, we acquired unpublished images of mice at baseline or after receiving ankle zymosan injections scored by the Mogil lab (SGS, using the standard cubicle system; Figure 1(a)) and combined them with additional baseline and ankle zymosan images generated and scored by the Zylka lab using our new cubicle system (JJ, Figure 1(b)). The collective image set (total n = 433) was analyzed by the aMGS and assigned a classification ("pain" or "no pain") as well as a confidence rating (0.5 to 1.0). We then compared these confidence ratings against human MGS scores and found a positive linear relationship between machine and human scores (Pearson's $r = 0.75$). One-sample *t*-tests reveal that aMGS confidence ratings were significantly different from chance for every human MGS score except "3," although it is apparent that the aMGS, as might be expected, had more difficulty classifying intermediate MGS scores (reflected by a lower confidence rating) than images that garnered very high or very low MGS scores (Figure 2). This finding is in line with our initial validation results—the aMGS is significantly more accurate when classifying

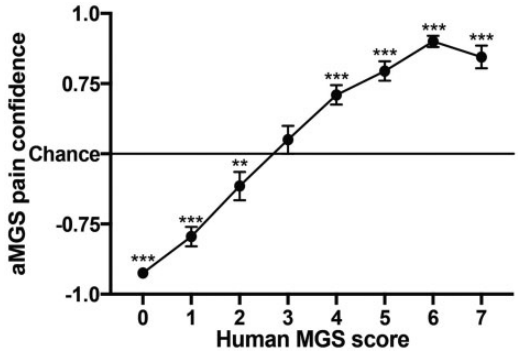

Figure 2. Direct correlation between human and machine grimace scores using novel grimace images. aMGS confidence scores were placed on a continuous scale, with −1.0 being equal to 100% confidence that the mouse image was not showing pain and 1.0 being equal to 100% confidence that the mouse image was showing pain. Resulting aMGS confidence scores were grouped by corresponding human MGS score. Bars represent mean ± SEM of transformed aMGS scores. Values significantly different from chance (**$p < 0.01$; ***$p < 0.0001$) according to one-sample *t* test. aMGS: automated Mouse Grimace Scale; MGS: Mouse Grimace Scale. "Chance"= 0.5

high-confidence "pain" and "no pain" images than ambiguous images.

The aMGS accurately detects pain and relief of pain in a post-operative model

We next tested the predictive validity of the aMGS in the laparotomy assay of post-operative pain. We used the aMGS to quantify the relative proportion of high-confidence images scored from mice at baseline with images taken 60 min after laparotomy or sham surgery. We compared these groups with a third cohort of animals that were given carprofen immediately after surgery. Mice filmed after laparotomy surgery produced significantly more "pain" images than sham surgery or mice given carprofen during recovery (one-way ANOVA found a significant main effect of group ($F(2,36) = 9.7$, $p < 0.0001$) (Figure 3(a)).

Despite the marked differences in the relative amount of "pain" images observed between laparotomy and control groups, we noticed that mice in the sham group also produced more "pain" images during the post-operative period. To minimize the possibility of mice falling asleep, a known confound for mouse grimace scoring, we re-ran the aMGS on a subset of videos taken from the first 30 min following surgery (n = 12 mice/condition). A one-way ANOVA of the results found a similar main effect by surgery group ($F(2,33) = 4.0$, $p = 0.03$) with laparotomized mice producing significantly more "pain" images than shams (Figure 3(b)). Mice given carprofen following laparotomy showed a decreased analgesic effect during this abbreviated observation period after surgery. This finding suggested that the analgesic was not as effective during the first 30 min observation period ($p = 0.07$).

Discussion

Using a large number of human-scored images of mouse faces, obtained from different experimental settings and across two different labs as a training set, we found that a machine learning model can accurately classify the presence of a mouse pain face in a rapid and unbiased manner when compared to human scoring. Internal validation of the initial (human scored) image training set revealed that the aMGS was comparable to trained human coders.[5] Furthermore, we observed a high degree of concordance between the aMGS and human MGS scores. Although images assigned intermediate human MGS scores (2 to 4) correspond with relatively low confidence ratings produced by our model, it is our experience that these images are of little diagnostic value. Intermediate MGS scores are assigned when a mouse fails to clearly demonstrate facial pain in a majority of action units. Furthermore, by excluding low-confidence images, the aMGS has a built-in quality control mechanism whereby low-quality images (i.e., where the face is not fully captured, is blurred, or includes the face of the adjacent mouse) are excluded from analysis. By restricting image classification to high-confidence ratings, the aMGS was comparable (94% accuracy) to the most highly trained human coder in previous studies.[5]

The aforementioned accuracy values were based on comparing aMGS output to human MGS scores across multiple pain assays. We also compared human scores to aMGS output in a second cohort of mice undergoing ankle zymosan injections, an assay of inflammatory pain. We found that our computer model's predictions correlated strongly with human scores across the entire MGS range, yielding a Pearson's $r = 0.75$. We further tested the utility of the aMGS with a post-surgical pain assay. We found that

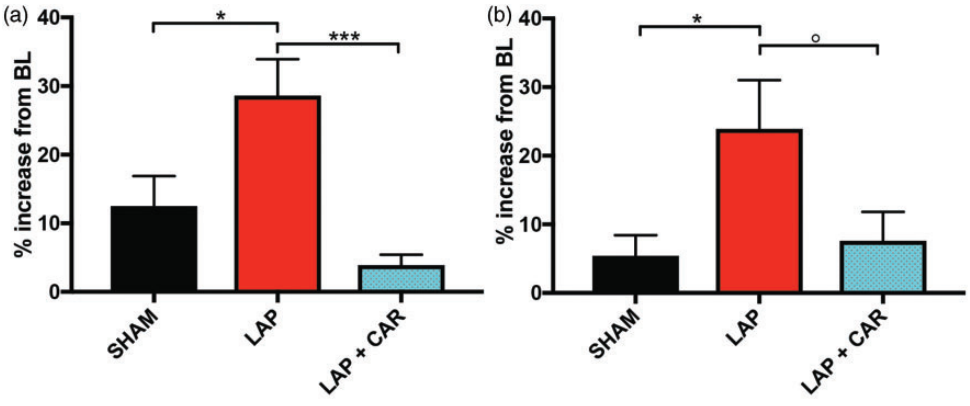

Figure 3. The aMGS correctly predicted analgesic efficacy in a post-operative pain assay. High-confidence images collected 60 min (a) or 30 min (b) following surgery. Bars represent mean ± SEM of difference scores (number of pain images after surgery – number of pain images at baseline). BL: baseline; SHAM: sham surgery; LAP: laparotomized animal; LAP + CAR: laparotomized animal given 50 mg/kg of the NSAID carprofen immediately following surgery. n = 12 to 14 for all conditions. Values $^{*}p < 0.05$; $^{***}p < 0.001$; $^{\circ}p = 0.074$ as determined by Tukey Test following one-way ANOVA.

the aMGS classified a significantly larger proportion of mouse images as being in pain following a laparotomy procedure than a comparable sham surgical procedure. Likewise, the aMGS detected significantly less mouse pain images after mice were administered the analgesic carprofen immediately following a laparotomy surgery, suggesting that our aMGS can detect pain-specific changes in the mouse face following the relief of pain.

We noticed that an increasing number of images from mice in the sham surgical cohort were scored as "pain" during the 1-h observation period when compared to baseline, albeit significantly less than laparotomized animals (Figure 3(a)). After checking video footage, we noted that the prolonged, 120-min testing period resulted in some of our sham animals falling asleep. An analysis of the first 30 min of behavior following surgery abrogated this effect, although it also decreased the effect of the analgesic, which was injected immediately after surgery and thus may not have had enough time to reach full efficacy (Figure 3(b)). Based on this analysis, we determined that our model has difficulty distinguishing between images of sleeping and grimacing mice, much like human observers.[5] To reduce chances of animals falling asleep, future experiments would benefit from shorter observation periods. One additional caveat of our model is that the current version of the aMGS is optimized to detect facial grimacing in albino mice. Further training sets containing non-albino mouse images will be needed to adapt the aMGS to classify pain in other strains of mice with agouti or black fur. Finally, the current version of the aMGS is only able to provide a binary pain assessment ("pain" or "no pain") and a semi-quantitative confidence rating, in contrast to human scores that can detect subtler changes in mouse facial expressions.

The successful development of machine learning applications presents an enormous opportunity for biomedical research. Researchers used large data sets (from previously established diagnostic testing) to train convolutional neural networks to detect human health issues with a high degree of accuracy, including skin cancer,[29,30] breast cancer,[31–34] and thoracic abnormalities.[35,36] Like these other automated systems, the aMGS can be deployed quickly and cheaply as a reliable replacement for manual MGS scoring, which is relatively labor intensive, low-throughput, and prone to human bias. Human efforts to evaluate rodent grimacing in real-time have so far produced mixed results.[37,38] The aMGS has the potential to be useful for a large number of applications, providing an objective metric for pain research, as well as serving as a front-line diagnostic for animal pain monitoring. Finally, the throughput of the aMGS makes possible long-term monitoring of pain, quickly and accurately assessing the presence of facial grimace in much larger data sets than previously described[5] and in models of chronic pain.[39,40,41]

Authors' Note

A copy of the aMGS is available for download. Please contact Dr Mark J Zylka for details regarding the model.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by an NICHD training grant (T32HD040127; AHT) and an NIH clinical translation research center grant for neurodevelopmental disorders (U5HD079124; JCP, MAS) administered by the Carolina Institute for Developmental Disabilities. Further funding was provided by The National Institutes of Health (DP1ES024088 & R01NS081127; MJZ) and an unrestricted grant from the Louise and Alan Edwards Foundation (JSM).

References

1. Borsook D, Hargreaves R, Bountra C, and Porreca F. Lost but making progress – where will new analgesic drugs come from? *Sci Transl Med* 2014; 6: 249sr3.
2. Mogil JS and Crager SE. What should we be measuring in behavioral studies of chronic pain in animals? *Pain* 2004; 112: 12–15.
3. Rice ASC, Cimino-Brown D, Eisenach JC, Kontinen VK, Lacroix-Fralish ML, Machin I, Mogil JS, and Stöhr T. Animal models and the prediction of efficacy in clinical trials of analgesic drugs: a critical appraisal and call for uniform reporting standards. *Pain* 2008; 139: 243–247.
4. Mogil JS. Animal models of pain: progress and challenges. *Nat Rev Neurosci* 2009; 10: 283–294.
5. Langford DJ, Bailey AL, Chanda ML, Clarke SE, Drummond TE, Echols S, Glick S, Ingrao J, Klassen-Ross T, Lacroix-Fralish ML, Lynn Matsumiya L, Sorge RE, Sotocinal SG, Tabaka JM, Wong D, van den Maagdenberg AMJM, Ferrari MD, Craig KD, and Mogil JS. Coding of facial expressions of pain in the laboratory mouse. *Nat Methods* 2010; 7: 447–449.
6. Sotocinal SG, Sorge RE, Zaloum A, Tuttle AH, Martin LJ, Wieskopf JS, Mapplebeck JCS, Wei P, Zhan S, Zhang S, McDougall JJ, King OD, and Mogil JS. The Rat Grimace Scale: a partially automated method for quantifying pain in the laboratory rat via facial expressions. *Mol Pain* 2011; 7: 55.
7. Ekman P and Friesen W. *Facial action coding system*. Palo Alto: Consulting Psychologists Press, 1978.
8. Grunau RVE and Craig KD. Pain expression in neonates: facial action and cry. *Pain* 1987; 28: 395–410.

9. Guesgen MJ, Beausoleil NJ, Leach M, Minot EO, Stewart M, and Stafford, KJ. Coding and quantification of a facial expression for pain in lambs. *Behav Processes* 2016; 132: 49–56.
10. Lu Y, Mahmoud M and Robinson P. *Estimating sheep pain level using facial action unit detection*. Washington, DC: IEEE, 2017, pp.394–399.
11. Dalla Costa E, Minero M, Lebelt D, Stucke D, Canali E, and Leach MC. Development of the Horse Grimace Scale (HGS) as a pain assessment tool in horses undergoing routine castration. *PLoS One* 2014; 9: e92281.
12. Gleerup KB, Forkman B and Lindegaard C. An equine pain face. *Vet Anaesth Analg* 2015; 42: 103–114.
13. Keating SCJ, Thomas AA, Flecknell PA, and Leach MC. Evaluation of EMLA cream for preventing pain during tattooing of rabbits: changes in physiological, behavioural and facial expression responses. *PLoS One* 2012; 7: e44437.
14. Rialland P, Otis C, de Courval M-L, Mulon P-Y, Harvey D, Bichot S, Gauvin D, Livingston A, Beaudry F, Hélie P, Frank D, del Castillo JRE, and Troncy E. Assessing experimental visceral pain in dairy cattle: a pilot, prospective, blinded, randomized, and controlled study focusing on spinal pain proteomics. *J Dairy Sci* 2014; 97: 2118–2134.
15. Viscardi AV, Hunniford M, Lawlis P, Leach M, and Turner PV. et al. Development of a Piglet Grimace Scale to evaluate piglet pain using facial expressions following castration and tail docking: a pilot study. *Front Vet Sci* 2017; 4: 230.
16. Holden E, Calvo G, Collins M, Bell A, Reid J, Scott EM, and Nolan AM. Evaluation of facial expression in acute pain in cats. *J Small Anim Pract* 2014; 55: 615–621.
17. Matsumiya LC, Sorge RE, Sotocinal SG, Tabaka JM, Wieskopf JS, Zaloum A, King OD, and Mogil JS. Using the Mouse Grimace Scale to reevaluate the efficacy of postoperative analgesics in laboratory mice. *J Am Assoc Lab Anim Sci* 2012; 51: 42–49.
18. Jeger V, Arrigo M, Hildenbrand FF, Müller D, Jirkof P, Hauffe T, Seifert B, Arras M, Spahn DR, Bettex D, and Rudiger A. Improving animal welfare using continuous nalbuphine infusion in a long-term rat model of sepsis. *ICMx* 2017; 5: 23.
19. Miller AL, Golledge HDR and Leach MC. The influence of isoflurane anaesthesia on the Rat Grimace Scale. *PLoS One* 2016; 11: e0166652.
20. Thomas A, Miller A, Roughan J, Malik A, Haylor K, Sandersen C, Flecknell P, and Leach M. Efficacy of intrathecal morphine in a model of surgical pain in rats. *PLoS One* 2016; 11: e0163909.
21. Waite ME, Tomkovich A, Quinn TL, Schumann AP, Dewberry LS, Totsch SK, and Sorge RE. Efficacy of common analgesics for postsurgical pain in rats. *J Am Assoc Lab Anim Sci* 2015; 54: 420–425.
22. Faller KME, McAndrew DJ, Schneider JE, and Lygate, C. A. Refinement of analgesia following thoracotomy and experimental myocardial infarction using the Mouse Grimace Scale. *Exp Physiol* 2015; 100: 164–172.
23. Leach MC, Klaus K, Miller AL, Scotto di Perrotolo M, Sotocinal SG, and Flecknell PA. The assessment of postvasectomy pain in mice using behaviour and the Mouse Grimace Scale. *PLoS One* 2012; 7: e35656.
24. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, and Wojna Z. Rethinking the inception architecture for computer vision. In: *Comp Vis Found* 2015; 1-10
25. Liu D, Peng F, Shea A, Rudovic O, and Picard R. DeepFaceLIFT: interpretable personalized models for automatic estimation of self-reported pain. *J Mach Learn Res* 2017; 66: 1–16.
26. Martinez DL, Rudovic O and Picard R. *Personalized automatic estimation of self-reported pain intensity from facial expressions*. Washington, DC: IEEE, 2017, pp.2318–2327.
27. Wiltschko AB, Johnson MJ, Iurilli G, Peterson RE, Katon JM, Pashkovski SL, Abraira VE, Adams RP, and Datta SR. Mapping sub-second structure in mouse behavior. *Neuron* 2015; 88: 1121–1135.
28. Eral M, Aktas CC, Kocak EE, and Dalkara T. *Assessment of pain in mouse facial images*. Washington, DC: IEEE, 2016, pp.1–4.
29. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, and Thrun S. Corrigendum: Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; 546: 686.
30. Zhang J, Song Y, Xia F, Zhu C, Zhang Y, Song W, Xu J, and Ma X. Rapid and accurate intraoperative pathological diagnosis by artificial intelligence with deep learning technology. *Med Hypotheses* 2017; 107: 98–99.
31. Antropova N, Huynh BQ and Giger ML. A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. *Med Phys* 2017; 15: 327.
32. Becker AS, Marcon M, Ghafoor S, Wurnig MC, Frauenfelder T, and Boss A. Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. *Invest Radiol* 2017; 52: 434–440.
33. Han Z, Wei B, Zheng Y, Yin Y, Li K, and Li S. Breast cancer multi-classification from histopathological images with structured deep learning model. *Sci Rep* 2017; 7: 4172.
34. Li H, Giger ML, Huynh BQ, and Antropova NO. Deep learning in breast cancer risk assessment: evaluation of convolutional neural networks on a clinical dataset of full-field digital mammograms. *J Med Imag* 2017; 4: 041304.
35. Bar Y, Diamant I, Wolf L, , and Greenspan H. Deep learning with non-medical training used for chest pathology identification. In: Proceedings SPIE 2015, vol. 9414, p. 94140
36. Rajkomar A, Lingam S, Taylor AG, Blum M, and Mongan J. High-throughput classification of radiographs using deep convolutional neural networks. *J Digit Imaging* 2017; 30: 95–101.
37. Leung V, Zhang E and Pang DS. Real-time application of the Rat Grimace Scale as a welfare refinement in laboratory rats. *Sci Rep* 2016; 6: 31667.
38. Miller AL and Leach MC. The Mouse Grimace Scale: a clinically useful tool? *PLoS One* 2015; 10: e0136000.

39. Duffy SS, Perera CJ, Makker PGS, Lees JG, Carrive P, and Moalem-Taylor G. Peripheral and central neuroinflammatory changes and pain behaviors in an animal model of multiple sclerosis. *Front Immunol* 2016; 7: 938.
40. Schneider LE, Henley KY, Turner OA, Pat B, Niedzielko TL, and Floyd CL. Application of the Rat Grimace Scale as a marker of supraspinal pain sensation after cervical spinal cord injury. *J Neurotrauma* 2017; 34: 2982–2993.
41. Wu J, Zhao Z, Zhu X, Renn CL, Dorsey SG, and Faden AI. Cell cycle inhibition limits development and maintenance of neuropathic pain following spinal cord injury. *Pain* 2016; 157: 488–503.