Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 September 2020 doi:10.20944/preprints202009.0101.v1

Review

The utility of grimace scales for practical pain assessment in laboratory animals

Daniel Mota-Rojas[1], Adriana Olmos-Hernández[2], Antonio Verduzco-Mendoza[2], Elein Hernández[3], Julio Martínez-Burnes[4] and Alexandra L. Whittaker[5]*

[1] Neurophysiology, behaviour and animal welfare assessment. DPAA. Universidad Autónoma Metropolitana, Xochimilco, Mexico City. Mexico.
[2] Division of Biotechnology - Bioterio and Experimental Surgery, Instituto Nacional de Rehabilitación-Luis Guillermo Ibarra Ibarra (INR-LGII), Mexico City. Mexico.
[3] Department of Clinical Studies and Surgery- Facultad de Estudios Superiores Cuautiltán UNAM, Cuautitlán, Estado de México.
[4] Graduate and Research Department, Facultad de Medicina Veterinaria y Zootecnia, Universidad Autónoma de Tamaulipas, Victoria City, Tamaulipas, Mexico.
[5] School of Animal and Veterinary Sciences, The University of Adelaide, Roseworthy Campus, Australia.

* Correspondence: alexandra.whittaker@adelaide.edu.au

Simple Summary: Grimace scales for laboratory animals were first reported ten years ago. Yet, in spite of their promise as pain assessment tools it appears that they have not been implemented widely in animal research establishments for clinical pain assessment. We discuss potential reasons for this based on the knowledge gained to date on their use and suggest avenues for further research, which might improve uptake of their use in laboratory animal medicine.

Abstract: Animals' facial expressions have been widely used as a readout for emotion. Scientific interest in the facial expressions of laboratory animals has centered primarily on negative experiences, such as pain, experienced as a result of scientific research procedures. Recent attempts to standardize evaluation of facial expressions associated with pain in laboratory animals has culminated in the development of "grimace scales". In the context of laboratory animals, these have been developed and evaluated for mice, rats, rabbits, sheep, and ferrets. The prevention or relief of pain in laboratory animals is a fundamental requirement for *in vivo* research to satisfy community expectations. However, to date it appears that the grimace scales have not seen widespread implementation as clinical pain assessment techniques in biomedical research. In this review, we discuss some of the barriers to implementation of the scales in clinical laboratory animal medicine, progress made in automation of collection, and suggest avenues for future research.

Keywords: facial expressions, pain, grimace scales, mice, rat, rabbit

1. Introduction

Animal welfare is an important societal concern[1,2]. The use of animals in biomedical scientific research is widespread, and globally significant, with approximately 115 million animals used per year[3]. Incontrovertibly, there is an ethical obligation to safeguard welfare of these animals through employing strategies to minimize pain, fear and distress[4-6], in addition to the promotion of positive welfare states. However, to achieve this, validated methods for identification of animal emotional state are required. In spite of significant research attention, ascertaining nature and strength of animal emotion remains a challenging task[7-11]. This is especially so for prey species, such as rats and mice, that are naturally more likely to mask emotional responses to increase their chances of survivals[12].

© 2020 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 September 2020 doi:10.20944/preprints202009.0101.v1

The study of emotion in laboratory animals has typically focused on aversive states such as pain. This area of study has been driven from two perspectives: a scientific and welfare standpoint. The scientific viewpoint, based on the extrinsic value of the animal, relates to the robustness of results acquired from animal models. There is an abundance of data on the impact of pain on a wide range of metabolic, immunologic and other processes in the body. These alterations introduce variability or confound interpretation of results[13-15]. The welfare viewpoint, considering the intrinsic value of the animal, assumes that pain occurs frequently in animal models and should therefore be avoided or minimized for the benefit of the animal. Notwithstanding, differences between these viewpoints in terms of underlying motivation for study, the requirement for a reliable, practical method for assessment of pain is shared by both.

Recently, evaluation of complex motor responses, such as facial and corporal expression has been proposed as a neurobiological readout of mammalian brain neuro-circuitry associated with emotional experience[11,16-18] (**Figure 1)**. The former has received significant research attention, especially in rodents, as a potential assessment method for both positive and negative emotional states[9]. There remains controversy as to the communicative function of facial expressions in rodents, since these species tend to prioritize other senses such as olfaction and touch in communication[8]. However, the finding that in mice, lesions of the insular cortex, modulate facial pain expressions supports the use of facial expression assessment. The insular cortex is associated with human pain perception; hence it is assumed by analogy that facial grimace may represent a negative emotional experience[20]. Furthermore, studies on empathy tends to suggest that rodents are communicating the presence of a painful state to others, to elicit an empathic response[21]. Although not specifically demonstrated, it is feasible that this may be occurring through interpretation of facial expression[8]. Additionally, it has been recently shown through the use of machine learning methods that facial expressions in mice may not only indicate direction of effect or valence of emotion (positive or negative), but intensity and persistence[22].

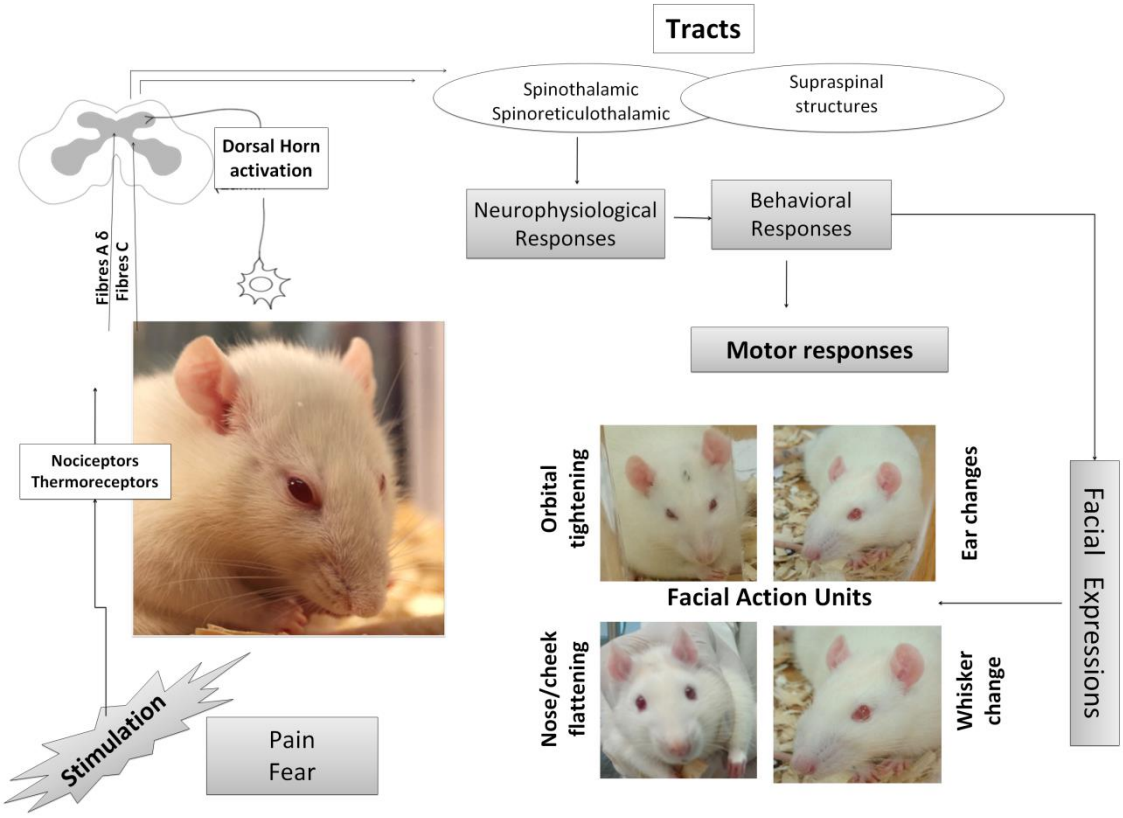

Figure 1. The neurophysiological process of pain begins with the activation and sensitization of the peripheral nociceptors that, through Aδ and C fibres, activate the dorsal horn of the spinal cord to allow modulation by the spinal interneurons, channeling information through the ascending

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 September 2020 doi:10.20944/preprints202009.0101.v1

pathways (spinothalamic and spinoreticulothalamic). The seventh cranial nerve regulates the spontaneous facial movements that produce facial expressions. The spinothalamic tract is considered the principle nociceptive pathway responsible for the ascent of the afferent signals of pain from the spinal cord to the cortex in rats[23].

Attempts to standardize evaluation of facial expressions for pain assessment has culminated in the development of the "grimace scales". These were developed originally for mice[20] and have been adapted for use in rats[24], rabbits[25,26], sheep[27], and ferrets[28], amongst other species. Grimace scales are simplified methods for evaluating facial expressions specifically related to pain based on the assessment of action units focusing on the eyes, ears and cheeks. The utility of the scales has been well-established across a range of laboratory animal species and animal model types. However, this evaluation has typically focused on their use via retrospective video recording review, and as a research tool to obtain data relevant to the animal model. There has been less dedicated study into the scales as 'bedside' pain assessment tools for rapid evaluation of pain status in laboratory animals in order to implement humane endpoints, or provide analgesia. Therefore, the focus of this review is to discuss the practical utility of grimace scales in a range of laboratory animal species, identifying barriers to their use and potential confounders. The focus will be on laboratory animal rodents as the most common species used in biomedical research, but research from other species will be drawn upon. It is anticipated that this review will guide biomedical researchers, animal technicians and ethics committees when implementing pain assessment methods as part of research protocols.

2. History of Facial Expression Scoring for Pain in Laboratory Animals

In recognition of the poor translation of outcomes from animal pre-clinical studies on pain physiology and analgesic development to humans[29,30], there has been a recent focus on development of methods for assessment of the affective pain response using non-evoked (spontaneous) responses[31]. Grimace scales are one such response derived from human facial codification scales[32,33]. The Facial Action Codification System (FACS) systematically catalogues all possible movements of the facial muscles, or combinations of them, such as lowering the eyebrows, tightening and closing the eyelids, wrinkling the nose and raising the upper lip. Categorization of changes in these muscle movements or so-called "Units of Facial Action" (UFA) enables facial recognition and categorization of emotions[17,19,34]. The finding that facial codification scales could quantify pain in humans with limited or non-existent verbal communication[35], provided the basis for using UFA in the development of grimace scales (GS) for animals[36].

The mouse grimace scale (MGS) was the first to be developed. Langford et al.[20] in 2010 applied a nociceptive abdominal constriction test through administration of acetic acid, that allowed the elucidation of facial action units that reliably detected pain. Validation was performed using a variety of traditional preclinical pain assays[20]. Five action units were described: 1) orbital tightening, 2) nose bulge, 3) cheek bulge, 4) ear position and 5) whisker change. A year later, Sotocinal et al.[24] in 2011 published the rat grimace scale (RGS) comprising four action units, due to consolidation of nose and cheek flattening into one unit. Utility of the RGS to detect pain was demonstrated in standard pre-clinical nociceptive tests in addition to following a surgical laparotomy procedure. Furthermore, the RGS was shown to be modified after analgesic administration indicating the specificity to pain[24]. Further, development of grimace scales in other common laboratory animal species followed see Table 1 and Figure 2.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 September 2020 doi:10.20944/preprints202009.0101.v1

Species	Validation Method	Action Units	Study
Mouse Grimace Scale (MGS)	Fourteen commonly used preclinical pain assays.	Five Units: 1) orbital tightening, 2) nose bulge, 3) cheek bulge, 4) ear position and 5) whisker change	[20]
Rat Grimace Scale (RGS)	Three pain-eliciting procedures performed. 1) Intraplantar administration of Complete Freund's adjuvant (CFA); 2) intra-articular administration of kaolin/carrageenan; and 3) post-operative pain after laparotomy.	Four Units: 1) orbital tightening,2) nose/cheek flattening, 3)ear changes, 4) whisker change	[24]
Rabbit Grimace Scale (RbtGS)	Pain caused by ear tattooing, a routine procedure used to identify rabbits. Analgesic test applied in the form They of prilocaine/lidocaine (EMLA) local anaesthetic	Five Units: 1) orbital tightening, 2) cheek flattening,3) nose shape, 4) whisker position, 5) ear position.	[25]
Sheep Grimace Scales (SPFES)	Clinical model based on mastitis and footrot	Five Units: 1) orbital tightening,2) cheek tightness,3) ear position, 4) lip and jaw profile,5) and nostril and philtrum position	[27]
Ferret (FGS)	Surgery involving the implantation of an intraperitoneal telemetry catheter	Five Units: 1) orbital tightening, 2) nose bulging, 3) cheek bulging, 4) ear changes, 5)whisker retraction	[28]
Piglets (PGS)	Castration and Tail docking. Validated orbital tightening for tail docking but remarked that further validation needed.	Ten used for development, later study[37] modified to three: 1) Ear Position	[38]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 September 2020 doi:10.20944/preprints202009.0101.v1

Species	Validation Method	Action Units	Study
		2) Cheek Tightening/Nose bulge 3) Orbital Tightening	

Table 1. Original studies in which grimace scales were developed for a range of species commonly used as laboratory animals.

Figure 2. Grimace scales (GS) standardized and validated for pain recognition in laboratory animals.

3. Terminology around pain classification and assessment

A variety of terms are often used to describe pain and the assessment methods applied to it. Pain is usually classified according to the duration of its effect or its originating source within the body[39,40]. Acute pain arises at the time of injury and is often experienced as different in nature to the alternatively described 'chronic' pain. The latter generally referring to pain experienced over a longer

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 September 2020 doi:10.20944/preprints202009.0101.v1

duration, although there appears to be no accepted duration marking the transition from acute to chronic pain[41]. An alternative distinction between the two time-course descriptors has been suggested by scientists; that related to functionality. Acute pain is argued to be adaptive, provoking a learned response by the animal to avoid a similar painful insult in the future[39]. Chronic pain on the other hand is said to be maladaptive[42]. However, this latter point is controversial with a variety of studies (see[43] for review) suggesting that pain-related hypervigilance may influence estimation of risk, subsequent behavior, and thus enhance survival.

Pain scales themselves are often described in terms of their validity, reliability, sensitivity[41]. Validity describes the extent to which the scale measures its intended outcome i.e. pain. There are a number of sub-categories describing validity. The most commonly referred to in the context of grimace scales are face validity and construct validity. Face validity describes what the test appears to be measuring i.e. pain. Construct validity relates to the extent to which the scales measure that specific construct. Therefore, the test needs to be both sensitive and specific to pain[44,45]. In pain studies construct validity is often determined using an applied analgesic test, since this is assumed to reduce pain and thereby reduce grimace scores if the test is truly pain-related[44]. External validity refers to how generalizable the measure is to other settings. In the context of grimace scales this is relevant in taking the scales from research scenarios to the clinical setting. This relates to practicability to perform during the working day, simplicity of the task, as well as the need for equipment and training. To date, this is the area that has received the least attention with regard to grimace scales.

Reliability refers to the scale producing the same result each time it is used both within, and between animals, and time-points[46]. In the context of grimace scales, this is determined by the variability resulting in a single observer's measurements (intra-observer variability), the variation between different observers' measurements (inter-observer), and variability between laboratories or research centers[44]. Sensitivity describes the ability of the scale to accurately identify changes in the degree of pain such that subtle changes are recognized[45]. In the context of pain scales this is often indicated when scale changes that occur correlate in direction, and proportion with other measures[45]. It is common in assessment of pain in veterinary species to achieve measurement accuracy in pain scoring by utilizing a smaller number of broad category groups, such as mild, moderate and severe, rather than expecting sensitivity when small differences in scores are considered. The following will consider how all of these measurement characteristics may influence the clinical applicability of grimace scales for use in biomedical research.

4. Clinical Applicability of Grimace Scales in Biomedical Research

4.1. Development of Real-Time Grimace Scores

There is now an extensive body of literature on the application of grimace scales in a range of animal models used commonly in biomedical research. The majority of this validation work has occurred in rodent models. It is beyond the scope of this review to describe all of the models used but the range includes oncology (see e.g.[47-50]), infectious disease[51], pain models[48,52,53], neurological conditions[33,54,55], genetic conditions[56], and maxillofacial interventions [49,50,57]. However, the vast majority of research to date has performed grimace scoring retrospectively from captured video footage.

Retrospective scoring is likely superior when using grimace scores to inform research outcomes, for example determining efficacy of analgesics or success of model induction. These methods allow for the possibility of replication, by multiple observers where appropriate, with an increased time available for scoring at the researcher's leisure. A cage-side or 'real time' method on the other hand would ideally provide instant assessment allowing interventions to support welfare, for example by implementing humane endpoints or administering analgesics. Development of the latter is clearly of more interest to ethical review committees and animal carers needing to make rapid clinical

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 September 2020 doi:10.20944/preprints202009.0101.v1

decisions. To date there has been substantially less focus on development and validation of real time methods.

Miller and Leach[58] in 2015 performed the first comprehensive evaluation of a real-time method applied in mice. In this study, both retrospective and real-time scoring were compared. Real-time scoring was performed by observing mice three times over a 10 min period, whilst animals were being filmed for the retrospective analysis. Grimace scores were calculated by summation of each action unit as described by Langford et al.[20], and totals were then averaged across the observation points. Live scores were always found to be significantly lower than corresponding retrospective video scoring. The authors posed that this could have resulted from the activity levels and changing nature of the face during live scoring. Blinking for instance, resulting in a score of 0 for orbital tightening, will likely be selected at least some of the time as a result of random chance selection of photographs for scoring. In a real time scenario, the rapid nature of blinking will likely preclude its scoring. Similarly, Chartier et al.[47] in 2020 also found consistently lower scores from live scoring compared to retrospective scoring in a mouse model of colitis-associated colo-rectal cancer. One potential explanation for this trend is that the presence of a human observer influences performance of the facial action units, for example, an increased alertness to the human (predator) could lead to wider eyes and 'pricked' ears, lowering the grimace score. On the contrary, intriguing findings from Sorge et al.[59] demonstrated that not all observers are equal, with no impact of a female observer on scores in rats and mice (obtained retrospectively), but a reduction of scores in the presence of a male[59]. In the first investigation of real-time scoring in rats, Leung et al.[60] in 2016 found that interval observations (15s of observation) were able to discriminate between control and analgesic–treated groups whereas point observations (conducted several times over a period) showed poor group discrimination. In this study substantial variability was seen between single observations of either point or interval. Limits of agreement, with a retrospective scoring system were however fairly large with a 0.5 score range either side of the bias meaning there is was a substantial risk of both over or underestimating the score. Furthermore, point scoring became generally unreliable at discriminating groups when done for less than 2 minutes, assumed to be due to a loss of power due to fewer observations. A later rat study by the same research group[61], investigated the interval method compared to a retrospective method in a colitis model showing the former to be reliable in predicting pain, with scores similar to the standard method.

The implications of these findings for clinical pain assessment are several. Firstly, it needs to be considered that although good discriminant ability was generally found in these studies, results were obtained by statistical combination of multiple scores. In a clinical scenario, an observer is likely to take one score, and not have the means or time to mathematically manipulate the values to arrive at a reliable score. Secondly, the Leung et al.[60] study suggests that variability across the observation period is likely and that at least 2 minutes of observation is needed. It is unlikely to be practical for a caregiver to spend 2 minutes per animal performing pain assessment across a study. In this case, some other more general method of distress measurement is likely to be needed to 'triage' animals for secondary grimace assessment. There has been no investigation of the effect of movement to the clear cages, in isolation, as typically occurs in grimace studies as opposed to scoring occurring in the home cage environment. A number of factors may influence the grimace scoring between these two scenarios. The novelty of the scoring box may trigger a state of alert influencing grimace scores in a similar vein to that suggested for the presence of a human observer. This novelty may indeed contribute to the variability seen between scores over time since habituation will eventually occur. Alternately, if scoring in the home cage, the presence of cage furniture, a potential more relaxed state of the animal in its familiar environment, or even the influence of circadian rhythms (see later) may all variously influence the action units or ability to see them accurately. A further consideration with real-time scoring is that there may be an inherent observer bias as the animal's overall demeanor, or presence of other pain behaviors such as twitching may be noted leading the observer to err on the side of higher action unit scores when unsure. This is not necessarily an issue per se in a clinical scenario since the goal is to recognize sick animals for further evaluation and treatment. However,

these other behaviors may not be unique to pain but represent general sickness behavior that may not be able to be rectified by analgesic administration, and hence inappropriate medication administration may occur. If such biasing were occurring it would be expected that there may be differences in grimace scoring between observers experienced with working with the species in question versus more naïve observers[47].

Notwithstanding, these findings some research groups do appear to have been able to use the MGS or RGS in a point observation, real-time scenario to obtain predicted results. For example, in chemotherapy-induced toxicity models in mice[62], and rats[63] single grimace scores allowed distinguishment between groups and followed the progression of the disease course as expected, after induction of chemotherapy-induced gut toxicity. Alternately, Hsi et al.[64] in 2020 were unable to use point mouse grimace scores to distinguish between groups either supplemented or not with dextrose following bariatric surgery. However, in this experimental design there was no sham group so it is unknown whether the MGS can reliably determine pain in this model[64].

There is clearly a need for further validation of real-time observation methods with a particular focus on one-off observations versus a series of observations, correlation with other established measures of pain assessment, inter-observer variability and home-cage versus novel area.

4.2 Impact of Biology and the Environment

4.2.1. Strain and sex differences

There is some evidence that features of biology, performance of routine procedures, or aspects of the environment may influence grimace scores. This has implications for setting of intervention scores (see later), and should be a consideration in driving further research or recommendations for application to clinical practice.

Aspects of biology have perhaps been the most researched with regard to their impact on grimace scores. The greatest implication of such changes likely relates to any differences between rodents strains or stocks given the wide range typically used in research. In mice, strain differences in MGS scores in animals not exposed to any painful interventions has been demonstrated. Miller and Leach in 2015[58] found that C3H/He mice showed significantly higher scores than CD-1 and C57BL/6 animals, although the order of effect for the latter two strains was different between males and females. In female BALB/c mice the grimace score was even higher than C3H/He (males were not investigated in this study). Cho et al.[65] in 2019 similarly demonstrated a difference in MGS scores post-craniotomy, with C57BL/6 mice having lower scores than CD-1 animals[65]. However, in pairwise comparisons of the CBA and DBA/2 strains in two further studies, no differences were found[66,67]. It has been suggested by some authors that detection of facial features in dark animals may be more difficult[65,68]. Improving the image quality and providing a contrasting background colour when recording appear to mitigate the effects[20], hence this may not be a feature of animal pigmentation per se. It should however be noted that in the Miller and Leach[58] in 2015 study, female C57BL/6 animals were not scored the lowest; that place being taken by the white CD-1 animals. Brown C3H animals also occupied an intermediate position. In a clinical scenario where real time scoring is likely to take place the issue of poor background contrast on videos if not of concern. However, some investigation of the effects of colour on live grimace scoring is warranted since it may be equally as difficult for a human observer to distinguish features such as whiskers against a similar coat colour background, especially when trying to observe at a distance so as not to influence the animal's behavior.

Differences between sexes have also been uncovered in research to date on the MGS, but results are complex and suggest there may be strain interactions. For example, Miller and Leach[58] observed no differences in MGS scores between male and female C57BL/6 mice[58]. However in the same study, both CD-1 and C3H/He males had greater scores than their female counterparts[58]. Similarly, male

BALB/c mice had higher grimace scores than females[69]. Alternately, Cho et al.[65] found no sex differences in CD-1 mice, although differences in response to analgesic were noted with females appearing to respond to carprofen with a reduction in grimace score more readily than males[65]. In rats limited study has been carried out into sex differences but no differences were found in the original validation study[24], or in a later study[70]. Unfortunately, it appears that most grimace studies in rats and mice appear to have been conducted in one sex, with a large proportion using male animals see eg.[52,71-73]. This bias in study design towards males, coupled with the enhanced understanding of the existence of different pathways and immune-cell types for pain processing between male and female rodents[74], renders extrapolation of findings to female rodents problematic.

4.2.2. Impact of routine procedures

It is clear that procedures occurring fairly often as part of vivarium routines may influence responses and should be taken into consideration when considering practical implementation of the grimace scales. For example, a number of studies have evaluated the impact of anesthetics on rodent grimace scales. In general, both inhalational and injectable anesthetics lead to a short-term increase in grimace scores in both rats[73] and mice[66,75,76], although strain differences in the presence of this response have been reported[66,75,76]. Whilst, this response is generally short-lived, repeated exposures lead to enhanced duration of the increase[73,68]. This is a particular consideration since grimace assessment would typically occur post-operatively to allow rescue analgesia administration and there is suggestion that the score increase may persist for up to a few hours post anesthesia[75,76].

There is a growing body of evidence that non-aversive handling of mice leads to reduced anxiety and improved resilience in the face of accompanying pain[77-79]. Cupping or tunnel handling are proposed as alternatives to the traditional method of picking up by the tail[78]. Perhaps somewhat surprisingly given the reported specificity of the MGS for pain there is some evidence that method of routine handling influences MGS with increased scores in mice handled by the tail compared to those that were tunnel handled[69]. This contradicts the findings of a previous study where no differences between the two methods were reported[67]. This is an area that should be a priority for further investigation for a number of reasons. Firstly, since non-aversive methods have not been widely incorporated into laboratory animal practice, especially amongst researchers[80], it is quite likely that mice even within one study will be subject to different handling techniques. Any effect of handling method on grimace score could therefore confound interpretation of grimace scores used to determine research protocol effects on pain. Secondly, whilst there appears to have been no dedicated study on whether tail handling induces pain, there is suggestion that it is non-painful, yet aversive[78]. If the method is actually non-painful this calls into question the specificity of the MGS for pain, and therefore whether it has construct validity.

Ear-tagging or ear-notching are routine handling procedures used to permanently identify laboratory animals[81]. These procedures are known to cause acute pain as reflected by alterations in physiological indices such as heart rate and blood pressure[82]. However, the results obtained by Miller and Leach[81] in mice did not reveal any change to MGS scores as a result of ear notching[81]. In a later mouse study, with a factorial study design evaluating handling method with ear tagging or tattooing, MGS was increased following ear tagging but tattooing or restraint had no impact on scores[69]. Alternately, Keating et al.[25] in 2012 showed that ear-tattooing in rabbits led to increases in rabbit grimace scale scores, that were ameliorated by the application of a local topical anaesthetic (lidocaine/prilocaine)[25]. Corticosterone measures in this study suggest that the pain response was short-lived and had resolved by 1-hour post procedure. Given that only three studies, performed in different species, have evaluated these common procedures, it would be unwise to draw firm conclusions. However, the lack of grimace score increase in the Miller and Leach[81] study does imply that the scale may not be sensitive to pain of a mild and short-lived nature either intrinsically, or as a result of practical features whereby the pain is missed due to the scoring process required. Conversely, this finding provides some evidence that routine procedures may have minimal effect on grimace scales, reducing

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 September 2020 doi:10.20944/preprints202009.0101.v1

the risk of confounding arising when using the scales for humane endpoint implementation in animal models. When reconciling the difference in findings between this[81], and the later study, Roughan and Sevenoaks[69] in 2019 speculated that ear tagging may be perceived as more painful than notching due to the prolonged irritation by the tag[69].

4.2.3. Environmental Impacts

If the grimace scales are to be utilized as a practical tool they need to be repeatable across time and conditions, and not subject to extraneous influences. This requirement also relates to their face validity as reliable indicators of pain. In common with the other factors that may influence the scales, there has been limited research in this sphere.

Miller and Leach in 2015[58] performed a comprehensive evaluation of some of the factors that might be predicted to influence grimace responses. One factor that may have an impact is the circadian cycle and whether differences in score occur across the day. For both live scoring and retrospective scoring, there were largely no differences seen between scores dependent on whether scoring took place in the morning, lunchtime or at the end of the day. There were some exceptions to this with BALB/c mice showing a greater live MGS score at noon compared to am and C57BL/6 mice showing higher retrospective MGS scores in both afternoon time points in comparison to the morning. It should be noted that in this, as in the majority of the studies examining grimace scores, animals were scored during the light phase of the circadian cycle when they would be expected to be inactive. There is some evidence that grimace scores may not be comparable between dark and light conditions with the finding that MGS was higher in the dark than in bright light in CD1 mice treated with a peptide believed to induce pain and migraine symptoms[83]. Analysis of the actions units showed that the transition to light caused a significant decrease in orbital tightening and nose bulge[83]. Given that this finding was also observed in vehicle controls it appears unrelated to the migraine symptoms, and may be an aspect of normal biology needing consideration. Alternately, Matsumiya et al. 2012 found no difference in baseline MGS scores between morning and evening but did find that in operated animals scores were higher in the dark cycle, implying that pain was greater in the mice active phase[84]. However, in consideration of the use of the scales as a practical tool the reality is that most scoring will occur during the working day, in the light cycle, and therefore the findings of Miller and Leach in 2015[58] provide confidence that time of scoring should not influence the score. A further important outcome from Miller and Leach[58] was that there was no effect of repeatedly being placed in the photography boxes on grimace score i.e. a habituation effect over the 3 occasions used[58]. The later study by Jirkof et al.[85] in 2020 supports this finding. This provides assurance that longitudinal monitoring post-procedure could occur throughout the day without the need to account for time of day or habituation to the box. However, as discussed earlier the need to remove animals to a separate box does impede the practical application of the test. Further study should consider time of day effects in the non-stimulated home environment.

There is further evidence of an impact of the external environment on grimace scores. Sorge et al.[59] compared grimace responses of mice and rats recorded after a painful insult, in the presence of a male compared to a female. Significant decreases in grimace response were recorded compared to the situation with no observer in the room. Females did not induce such a change. The findings therefore suggest that olfactory cues from human males lead to a physiological stress response, and associated stress-induced analgesia.

There is also some evidence of inter-laboratory variation in the outputs obtained from behavioural testing, to include MGS scores. In a multicenter study, Jirkof et al.[85] in 2020 demonstrated some quantitative differences in scores, although they were qualitatively comparable (direction of effect). However, variability between research centers in the MGS, especially when presented as a median score, was less pronounced than in burrowing behavior readouts[84]. This inter-lab variability has been recognized across the spectrum of preclinical research pursuits, arising as a result of environmental

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 September 2020 doi:10.20944/preprints202009.0101.v1

variables leading to stress. [86] Whilst this issue may be a concern when considering basic-to –clinical translation and reproducibility, it is less likely to be of concern for clinical application of the grimace scales. As a clinical tool, provided good inter and intra-observer, and thus intra-site agreement is obtained, grimace scores may be relied on locally for welfare determination subject to some of the other caveats discussed in this paper.

4.3 Validity

If grimace scales are to be implemented as a routine clinical assessment tool in biomedical research facilities, there needs to be a clear understanding of whether they are specific to pain, and can reliably measure pain in the models being used. This is important because it influences the animal caretaker's decision as regard to treatment options, for example, whether analgesics will be effective in mitigating clinical signs. It can be seen from the above discussion that there are a range of external factors that affect grimace scores, speaking to their validity as a pain assessment tool; anesthetics are a prime example. Setting aside the lack of study into their application in a real-time scenario, which influences their generalizability, another key concern is whether they are valid for all pain types. Results of the original Langford et al.[20] in 2010 study suggested that the technique was only applicable for acute pain states[20], since changes were not recorded after the application of traditional models of chronic pain, such as chronic constriction injury (CCI). However, there have now been a range of studies, largely performed in mice, which suggest that the grimace scale may be applicable for pain that is chronic or neuropathic in nature, or of a non-surgical origin.

The study findings of Akintola et al.[52] in 2017 contradict the previous results of Langford et al 2010 with both RGS and MGS increasing after application of the CCI model in these species. Pain arising from cancer has also been shown to cause an elevation of the MGS, for example in colo-rectal cancer[47] and in a metastatic breast cancer model[49,50]. The MGS has been successfully used in models expected to produce pain of a neuropathic nature, for example in headache and migraine[55,87] and craniotomy[65]. There is also suggestion that pain of a visceral nature elevates scores based on studies evaluating colonic nociception[88], pelvic pain[89], colitis[61], and alimentary mucositis[62,63]. Hereditary sickle cell disease frequently leads to painful episodes in human patients. Cold treatment of transgenic sickle mice led to increased grimace scores which were alleviated using a known analgesic agent. Furthermore, body changes of decreased length and increased back curvature were also correlated with the change in grimace scores[56]. These findings lend support to the proposition that the grimace scales have good construct validity for non-acute pain.

In spite of these findings, results from other studies implies that further evaluation of the grimace techniques are necessary to ascertain validity. For example, in contradiction to later work[62,63] demonstrating elevations in scores in rats and mice with mucositis, Whittaker et al.[90] in 2015 found no change in grimace scores in a rat model, albeit using retrospective rather than real time scoring. However, this study did find increases in frequency of established behavioral indicators of pain such as back arching and twitching[90]. Alternately, Leung et al.[61] in a rat DSS- colitis model found grimace score increases in the absence of an increase in composite behavioral score.

Other studies also raise questions of whether the grimace scales are truly unique to pain. Caecal ligation and puncture models are commonly used to study sepsis[91]. Whilst sepsis is undoubtedly a painful condition based on human reports[92], there is also an overwhelming cytokine response causing sickness behavior. Studies to date on this model[51,93] have not teased apart the possible contribution of this sickness response to the facial expression changes. There is a study that lends support to this idea; the work of Yamamoto et al. in 2016[94], whilst not employing the published rat grimace scale, provides evidence that nausea influences the eye action unit. Toxin administration, which might also be expected to cause dual symptoms of pain and sickness, similarly elevated the MGS[95]. Furthermore, analgesic administration was not always successful in reducing the scores implying an alternate

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 September 2020 doi:10.20944/preprints202009.0101.v1

cause of the facial action unit response. Finally, head injury may alter the animal's ability to influence the facial action units via neural mechanisms and render grimace scores unreliable[44].

4.4 Automation of Techniques

One of the main current barriers to widespread clinical application of the grimace scales is the lack of understanding as to their validity and reliability when used for live scoring. However, as illustrated, there is now a wealth of literature on the validity and application of retrospective techniques using video or photo footage. In a clinical scenario these methods have limited application due to the time taken to extract the images, perform the scoring and potentially combine scores using statistical methods. However, there has been investigation of a range of technologies which minimize the time taken for various aspects of this process. At the simplest level use of freeware video to JPG converter software can reduce the time associated with manual searching and capture of images from recorded video footage by automating the capture process[48]. However, this still requires manual viewing of the selected images to obtain unobstructed head shots. Sotocinal et al.[24] in 2011 developed Rodent Face Finder® which is able to detect rodent eyes and ears to generate stills of rodent faces[24]. This software has been used in a range of studies measuring grimace scores in both rats and mice see eg. [52,84,44,96]. Recently, another research group generated an algorithm to generate repeatable, non-observer biased, standardised and randomised pictures in one step. The authors suggest that their system offers benefits in scoring animals with dark fur and allowing several animals to be filmed and generate images simultaneously[97]. They further went on to show that the system was robust across a number of facilities potentially minimising issues around inter-laboraotry variability as discussed previously[98].

This process of semi-automation makes grimace scoring somwehat more applicable to a clinical environment but the time taken to manually score images is still likely to be a barrier to implementation. In recent years there has been some progress on further automation of facial expression recognition using machine learning techniques. Deep learning methods allow classification and predictions on the data without previous feature design[99]. Using combined methods, Andersen et al.[99] in 2020 created a software which locates and extracts the mouse face in an image, as well as scoring expression using a deep neural network. Based on assessment of a binary outcome (pain versus no pain) this system achieved an accuracy of 99%. Other groups have similarly demonstrated the promise of deep learning methods for use with the MGS when based on binary outputs[100,101].

These automation methods are in their infancy and no doubt there will be further development of these techniques over the next few years. A key issue at the moment is that they lack sensitivity- being only able to distinguish a painful from a non-painful state. This renders their current use for welfare assessment and endpoint implementation limited. However, given the success and practical implementation of machine learning methods in recognition of human facial expression, it is likely to be only a matter of time before a similar level of sensitivity of scoring will be possible in animal – focused methods[102].

5. Practical Considerations

The above discussion highlights some areas in need of future research particularly in regard to practical usage of the grimace scales in laboratory animal medicine. A key issue is what to do with the data when it is acquired, and what it means for the animal. In research use of the grimace scales, statistically significant differences in grimace scores in comparison with controls are typically reported. However, in a clinical scenario, a mass of data or control animals' results may not be available to make this comparison on-the-spot. Moreover, statistical significance may not always equate with clinical significance. There needs to be ascertainment of the level of grimace score at which pain is actually occurring, since the evidence suggests that grimace scores in healthy animals

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 September 2020 doi:10.20944/preprints202009.0101.v1

are rarely zero[58]. Some attempts have been made to address this issue with the development of an intervention threshold. Scores that are above this level signify that the animal is in pain, and consideration should be given to providing rescue analgesia[103]. These thresholds would need to be derived based on the method of combining individual action unit scores used, for example summation of scores leads to a maximum of 10, whereas averaging leads to a maximum of 2. Oliver et al.[103] in 2014 determined for rats, that 0.67/2 was a suitable intervention threshold. Intervention thresholds have not yet been developed in other species. Since individuals experience pain differently, and there are associated sex differences in both pain experience and response to analgesics, work is needed to tailor intervention thresholds considering these factors.

Given, the lack of established intervention thresholds perhaps the best current advice would be to use a holistic approach in pain assessment and consider grimace scores alongside other measures of well-being such as standard clinical scoring, and where possible look for trends in score progression within the same animal to guide decision-making. Animal carers also need to consider the potential impacts of inter and intra-observer variability on scoring which may be significant when statistical methods on group data are not used to smooth out variability. A prudent approach, where possible, would be to use the same scorer in a clinical case. This concern also brings up the issue of training of scorers which has received minimal research attention. Some studies have implied that minimal training, such as the provision of online instructions, is all that is necessary to achieve consistent results between expert and novice scorers[69,104]. However, another study has shown that more in-depth training, utilizing practice scoring associated with structured opportunities for discussion, enhanced scoring ability[105].

6. Conclusions and Future Directions

In spite of 10 years of investigation, widespread uptake of grimace scoring in biomedical research has not occurred. The grimace scales offer enormous potential for clinical use in biomedical research. They are simple, require no equipment and have been shown through research study to have good construct validity for most conditions. However, the methodology used in research on grimace scales is unlikely to lend to practical implementation due to its time intensive and retrospective nature. To date, few studies have investigated the validity of grimace scales in scenarios requiring on the spot pain assessment and clinical decision-making. Key areas for focus are on grimace score validity in animals housed in home cages, the reliability of using a limited number of real-time observation points, the impact of observers on scores, and the need for observer training. This is an area in urgent need of future research to realise the potential value of grimace scales.

One area that has received attention is the automation of scales using machine learning and algorithmic methods. This is a welcome development and will enhance the practical potential of grimace scales. It is hoped that in future years, grimace scale scoring may just be one of a number of outcome measures acquired routinely through facility-automated systems. This scenario is most likely to address the practical issues inherent when dealing with large numbers of animals, going some way towards addressing public concern around ethical decision-making in biomedical research.

Author Contributions:

Funding: This research received no external funding. A.W. is supported by an Australian Government, NHMRC Peter Doherty Biomedical Research Fellowship (APP1140072).

Conflicts of Interest: The authors declare that they have no conflict of interest.

References

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 September 2020 doi:10.20944/preprints202009.0101.v1

1 Mota-Rojas, D., Velarde, A., Maris-Huertas, S., Cajiao, M. N. Editors. In: Animal welfare, a global vision in Ibero-America. 3rd ed. Barcelona, Spain. 1-516. (Elsevier, 2016).
2 Lewejohann, L., Schwabe, K., Häger, C. & Jirkof, P. Impulse for animal welfare outside the experiment. *Lab. Anim.* 54, 150-158 (2020).
3 Taylor, K., Gordon, N., Langley, G. & Higgins, W. Estimates for worldwide laboratory animal use in 2005. *Altern. Lab. Anim*. 36, 327-342 (2008).
4 Mota-Rojas, D. et al. Neurological modulation of facial expressions in pigs and implications for production. *J. Anim. Behav. Biometeorol*. 8, 232–243 (2020).
5 Mota-Rojas D. et al. Infrared thermal imaging associated with pain in laboratory animals. *Exp. Anim.* 70, Accepted (2021).
6 Baumans, V. Science-based assessment of animal welfare: laboratory animals. *Rev. sci. tech. Off. int. Epiz*. 24 (2005)
7 Lezama-García, K. et al. Facial expressions and emotions in domestic animals. *CAB Rev. Perspect. Agric. Vet. Sci. Nutr. Nat. Resour.* 14, (2019).
8 Finlayson, K., Lampe, J. F., Hintze, S., Würbel, H. & Melotti, L. Facial indicators of positive emotions in rats. *PloS One* 11 (2016).
9 Whittaker, A. L. & Marsh, L. E. The role of behavioural assessment in determining 'positive' affective states in animals. CAB Rev. *Perspect. Agric. Vet. Sci. Nutr. Nat. Resour.* 14, 1-13 (2019).
10 Boissy, A. et al. Assessment of positive emotions in animals to improve their welfare. *Physiol. Behav*. 92, 375–397 (2007)
11 Panksepp, J. Affective consciousness: Core emotional feelings in animals and humans. *Conscious. Cogn.* 14, 30-80 (2005).
12 Stasiak, K. L., Maul, D., French, E., Hellyer, P. W. & VandeWoude, S. Species-specific assessment of pain in laboratory animals. *Contemp. Top Lab. Anim. Sci.* 42, 13-20 (2003).
13 Carbone, L. & Austin, J. Pain and Laboratory Animals: Publication Practices for Better Data Reproducibility and Better Animal Welfare. *PLoS One* 11, e0155001 (2016).
14 Peterson, N. C., Nunamaker, E. A. & Turner, P. V. To Treat or Not to Treat: The Effects of Pain on Experimental Parameters. *Comp. Med.* 67, 469-482 (2017).
15 Zurlo, J. & Hutchinson, E. Refinement. *ALTEX* 31, 4-10 (2014).
16 Bennett, V., Gourkow, N. & Mills, D. S. Facial correlates of emotional behaviour in the domestic cat (Felis catus). *Behav. Processes* 141, 342-350 (2017).
17 Ekman, P. Are there basic emotions?. *Psychol. Rev.* 99, 550-553 (1992).
18 Mota-Rojas, D. *et al.* Teaching animal welfare in veterinary schools in Latin America. *Int. J. Vet. Sci. Med.* 6, 131-140 (2018).
19 Descovich, K. A. *et al.* Facial Expression: An Under-Utilized Tool for the Assessent of Welfare in Mammals. *ALTEX* 34, 409-429 (2017).
20 Langford, D. J. *et al.* Coding of facial expressions of pain in the laboratory mouse. *Nat. Methods* 7, 447 (2010).
21 Langford, D. J. *et al.* Social modulation of pain as evidence for empathy in mice. *Science* 312, 1967-1970 (2006).
22 Dolensek, N., Gehrlach, D. A., Klein, A. S. & Gogolla, N. Facial expressions of emotion states and their neuronal correlates in mice. *Science* 368, 89-94 (2020).
23 Palecek, J., Paleckova, V. & Willis, W. D. Postsynaptic dorsal column neurons express NK1 receptors following colon inflammation. *Neuroscience* 116, 565-572 (2003).
24 Sotocinal, S. G. *et al.* The Rat Grimace Scale: A partially automated method for quantifying pain in the laboratory rat via facial expressions. *Mol. Pain* 7 (2011).
25 Keating, S. C., Thomas, A. A., Flecknell, P. A. & Leach, M. C. Evaluation of EMLA cream for preventing pain during tattooing of rabbits: changes in physiological, behavioural and facial expression responses. *PLoS One* 7, e44437 (2012).
26 Hampshire, V. & Robertson, S. Using the facial grimace scale to evaluate rabbit wellness in post-procedural monitoring. *Lab. Anim. (NY)* 44, 259-260 (2015).
27 McLennan, K. M. *et al.* Development of a facial expression scale using footrot and mastitis as models of pain in sheep. *Appl. Anim. Behav. Sci*. 176, 19-26 2016.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 September 2020 doi:10.20944/preprints202009.0101.v1

28 Reijgwart, M. L. *et al.* The composition and initial evaluation of a grimace scale in ferrets after surgical implantation of a telemetry probe. *PLoS One* 12, e0187986 (2017).
29 Apkarian, A. V., Hashmi, J. A. & Baliki, M. N. Pain and the brain: specificity and plasticity of the brain in clinical chronic pain. *Pain* 152, S49-64 (2011).
30 Blackburn-Munro, G. Pain-like behaviours in animals - how human are they? *Trends Pharmacol. Sci.* 25, 299-305 (2004).
31 Hernandez-Avalos, I. et al. Review of different methods used for clinical recognition and assessment of pain in dogs and cats. *Int. J. Vet. Sci. Med.* 7, 43-54 (2019).
32 Nagakura, Y. *et al.* Spontaneous pain-associated facial expression and efficacy of clinically used drugs in the reserpine-induced rat model of fibromyalgia. *Eur. J. Pharmacol.* 864 (2019).
33 Serizawa, K., Tomizawa-Shinohara, H., Yasuno, H., Yogo, K. & Matsumoto, Y. Anti-IL-6 Receptor Antibody Inhibits Spontaneous Pain at the Pre-onset of Experimental Autoimmune Encephalomyelitis in Mice. *Front. Neurol.* 10, 341 (2019).
34 LeResche, L. Facial expression in pain: A study of candid photographs. *J. Nonverbal Behav.* 7, 46-56 (1982).
35 Williams, A. Facial expression of pain: An evolutionary account. *Behav. Brain Sci.* 25, 439-455 (2002).
36 Schanz, L., Krueger, K. & Hintze, S. Sex and Age Don't Matter, but Breed Type Does-Factors Influencing Eye Wrinkle Expression in Horses. *Front. Vet. Sci.* 6, 154 (2019).
37 Viscardi, A. V., Hunniford, M., Lawlis, P., Leach, M. & Turner, P. V. Development of a Piglet Grimace Scale to Evaluate Piglet Pain Using Facial Expressions Following Castration and Tail Docking: A Pilot Study. *Front. Vet. Sci.* 4, 51 (2017).
38 Di Giminiani, P. *et al.* The assessment of facial expressions in piglets undergoing tail docking and castration: toward the development of the piglet grimace scale. *Front. Vet. Sci.* 3, 100 (2016).
39 Bateson, P. Assessment of pain in animals. *Anim. Behav.* 42, 827-839 (1991).
40 Whittaker, A. L. & Howarth, G. S. Use of spontaneous behaviour measures to assess pain in laboratory rats and mice: How are we progressing? *Appl. Anim. Behav. Sci.* 151, 1-12 (2014).
41 Rutherford, K. Assessing pain in animals. *Anim. Welfare* 11, 31-53 (2002).
42 Williams, A. C. d. C. Persistence of pain in humans and other mammals. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 374, 20190276 (2019).
43 Walters, E. T. & Williams, A. C. De C. Evolution of mechanisms and behaviour important for pain. *Philos. Trans. R. Soc. Lond. B Biol. Sci* 374, 20190275 (2019).
44 McLennan, K. M. *et al.* Conceptual and methodological issues relating to pain assessment in mammals: The development and utilisation of pain facial expression scales. *Appl. Anim. Behav* 217, 1-15 (2019).
45 Bendinger, T. & Plunkett, N. Measurement in pain medicine. *BJA Educ.* 16, 310-315 (2016).
46 Good, M. *et al.* Sensation and distress of pain scales: reliability, validity, and sensitivity. *J. Nurs. Meas.* 9, 219-238 (2001).
47 Chartier, L. C., Hebart, M. L., Howarth, G. S., Whittaker, A. L. & Mashtoub, S. Affective state determination in a mouse model of colitis-associated colorectal cancer. *PLoS One* 15, e0228413 (2020).
48 George, R. P., Howarth, G. S. & Whittaker, A. L. Use of the Rat Grimace Scale to Evaluate Visceral Pain in a Model of Chemotherapy-Induced Mucositis. *Animals* 9, 678 (2019).
49 de Almeida, A. S. *et al.* Characterization of Cancer-Induced Nociception in a Murine Model of Breast Carcinoma. *Cell. Mol. Neurobiol.* 39, 605–617 (2019).
50 de Almeida, A. S. *et al.* Role of transient receptor potential ankyrin 1 (TRPA1) on nociception caused by a murine model of breast carcinoma. *Pharmacol. Res.* 152, 104576 (2020).
51 Mai, S. H. C. *et al.* Body temperature and mouse scoring systems as surrogate markers of death in cecal ligation and puncture sepsis. *Intensive Care Med. Exp.* 6, 20 (2018).
52 Akintola, T. *et al.* The grimace scale reliably assesses chronic pain in a rodent model of trigeminal neuropathic pain. *Neurobiol. Pain* 2, 13-17 (2017).
53 Akintola, T. *et al.* In search of a rodent model of placebo analgesia in chronic orofacial neuropathic pain. *Neurobiol. Pain* 6, 100033 (2019).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 September 2020 doi:10.20944/preprints202009.0101.v1

54 Duffy, S. S. *et al.* Peripheral and Central Neuroinflammatory Changes and Pain Behaviors in an Animal Model of Multiple Sclerosis. *Front. Immuno.l* 7, (2016).
55 Hassler, S. N. *et al.* Protease activated receptor 2 (PAR2) activation causes migraine-like pain behaviors in mice. *Cephalalgia* 39, 111-122 (2019).
56 Mittal, A., Gupta, M., Lamarre, Y., Jahagirdar, B. & Gupta, K. Quantification of pain in sickle mice using facial expressions and body measurements. *Blood Cells, Mol. Dis.* 57, 58-66 (2016).
57 Gao, M. *et al.* The role of periodontal ASIC3 in orofacial pain induced by experimental tooth movement in rats. *Eur. J. Orthod.* 38, 577-583 (2016).
58 Miller, A. L. & Leach, M. C. The Mouse Grimace Scale: A Clinically Useful Tool? *PLoS One* 10, e0136000 (2015).
59 Sorge, R. E. *et al.* Olfactory exposure to males, including men, causes stress and related analgesia in rodents. *Nat. Methods* 11, 629-632 (2014).
60 Leung, V., Zhang, E. & Pang, D. S. J. Real-time application of the Rat Grimace Scale as a welfare refinement in laboratory rats. *Sci. Rep.* 6, 31667 (2016).
61 Leung, V. S. Y., Benoit-Biancamano, M. O. & Pang, D. S. J. Performance of behavioral assays: the Rat Grimace Scale, burrowing activity and a composite behavior score to identify visceral pain in an acute and chronic colitis model. *PAIN Rep.* 4, e718 (2019).
62 Wardill, H. R. *et al.* Irinotecan-Induced Gastrointestinal Dysfunction and Pain Are Mediated by Common TLR4-Dependent Mechanisms. *Mol. Cancer Ther.* 15, 1376-1386 (2016).
63 Gibson, R. J. *et al.* Chemotherapy-induced gut toxicity and pain: involvement of TLRs. *Support. Care Cancer* 24, 2251-2258 (2016).
64 Hsi, Z. Y., Stewart, L. A., Lloyd, K. C. K. & Grimsrud, K. N. Hypoglycemia after Bariatric Surgery in Mice and Optimal Dosage and Efficacy of Glucose Supplementation. *Comp. Med.* 70, 111-118 (2020).
65 Cho, C. *et al.* Evaluating analgesic efficacy and administration route following craniotomy in mice using the grimace scale. *Sci. Rep.* 9, 359 (2019).
66 Miller, A., Kitson, G., Skalkoyannis, B. & Leach, M. The effect of isoflurane anaesthesia and buprenorphine on the mouse grimace scale and behaviour in CBA and DBA/2 mice. *Appl. Anim. Behav. Sci.* 172, 58-62 (2015).
67 Miller, A. L. & Leach, M. C. The effect of handling method on the mouse grimace scale in two strains of laboratory mice. *Lab. Anim.* 50, 305-307 (2016).
68 Dalla Costa, E. *et al.* Can grimace scales estimate the pain status in horses and mice? A statistical approach to identify a classifier. *PLoS One* 13, e0200339 (2018).
69 Roughan, J. V. & Sevenoaks, T. Welfare and Scientific Considerations of Tattooing and Ear Tagging for Mouse Identification. *J. Am. Assoc. Lab. Anim. Sci.*58, 142-153 (2019).
70 Waite, M. E. *et al.* Efficacy of Common Analgesics for Postsurgical Pain in Rats. *J. Am. Assoc. Lab. Anim. Sci.* 54, 420-425 (2015).
71 Wang, S. *et al.* TRPV1 and TRPV1-Expressing Nociceptors Mediate Orofacial Pain Behaviors in a Mouse Model of Orthodontic Tooth Movement. *Front. Physiol.* 10 (2019).
72 Zhu, Y. *et al.* Effect of static magnetic field on pain level and expression of P2X3 receptors in the trigeminal ganglion in mice following experimental tooth movement. *Bioelectromagnetics* 38, 22-30 (2017).
73 Miller, A. L., Golledge, H. D. R. & Leach, M. C. The Influence of Isoflurane Anaesthesia on the Rat Grimace Scale. *Plos One* 11, e0166652 (2016).
74 Sorge, R. *et al.* & Mogil, JS (2015). Different immune cells mediate mechanical pain hypersensitivity in male and female mice. *Nat. Neurosci*. 18,1081-1083 (2015)
75 Hohlbaum, K. *et al.* Severity classification of repeated isoflurane anesthesia in C57BL/6JRj mice-Assessing the degree of distress. *PLoS One* 12, e0179588 (2017).
76 Hohlbaum, K. *et al.* Impact of repeated anesthesia with ketamine and xylazine on the well-being of C57BL/6JRj mice. *PLoS One* 13, e0203559 (2018).
77 Gouveia, K. & Hurst, J. L. Optimising reliability of mouse performance in behavioural testing: the major role of non-aversive handling. *Sci. Rep.* 7, 44999 (2017).
78 Hurst, J. L. & West, R. S. Taming anxiety in laboratory mice. *Nat. Methods* 7, 825-826 (2010).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 September 2020 doi:10.20944/preprints202009.0101.v1

79 Gouveia, K. & Hurst, J. L. Reducing Mouse Anxiety during Handling: Effect of Experience with Handling Tunnels. *PLoS One* 8, e66401 (2013).
80 Henderson, L. J., Smulders, T. V. & Roughan, J. V. Identifying obstacles preventing the uptake of tunnel handling methods for laboratory mice: An international thematic survey. *PLoS One* 15, e0231454 (2020).
81 Miller, A. & Leach, M. Using the mouse grimace scale to assess pain associated with routine ear notching and the effect of analgesia in laboratory mice. *Lab. Anim.* 49, 117-120 (2015).
82 Kasanen, I. H. E., Voipio, H. M., Leskinen, H., Luodonpää, M. & Nevalainen, T. O. Comparison of ear tattoo, ear notching and microtattoo in rats undergoing cardiovascular telemetry. *Lab. Anim.* 45, 154-159 (2011).
83 Rea, B. J. *et al.* Peripherally administered calcitonin gene-related peptide induces spontaneous pain in mice: implications for migraine. *Pain* 159, 2306-2317 (2018).
84 Matsumiya, L. C. *et al.* Using the Mouse Grimace Scale to reevaluate the efficacy of postoperative analgesics in laboratory mice. *J. Am. Assoc. Lab. Anim. Sci.* 51, 42-49 (2012).
85 Jirkof, P. *et al.* A safe bet? Inter-laboratory variability in behaviour-based severity assessment. *Lab. Anim.* 54, 73-82 (2020).
86 Mogil, J. S. Laboratory environmental factors and pain behavior: the relevance of unknown unknowns to reproducibility and translation. *Lab. Anim. (NY).* 46, 136-141 (2017).
87 Burgos-Vega, C. C. *et al.* Non-invasive dural stimulation in mice: A novel preclinical model of migraine. *Cephalalgia* 39, 123-134 (2019).
88 Hassan, A. M. *et al.* Visceral hyperalgesia caused by peptide YY deletion and Y2 receptor antagonism. *Sci. Re.p* 7, 40968 (2017).
89 Bu, X., Liu, Y., Lu, Q. & Jin, Z. Effects of "Danzhi Decoction" on Chronic Pelvic Pain, Hemodynamics, and Proinflammatory Factors in the Murine Model of Sequelae of Pelvic Inflammatory Disease. *Evid. Based Complement. Alternat. Med.* 2015, 547251, doi:10.1155/2015/547251 (2015).
90 Whittaker, A. L., Leach, M. C., Preston, F. L., Lymn, K. A. & Howarth, G. S. Effects of acute chemotherapy-induced mucositis on spontaneous behaviour and the grimace scale in laboratory rats. *Lab. Anim.* 50, 108-118 (2015).
91 Toscano, M. G., Ganea, D. & Gamero, A. M. Cecal ligation puncture procedure. *J. Vis. Exp.* doi:10.3791/2860 (2011).
92 Nguyen, H. B. *et al.* Severe Sepsis and Septic Shock: Review of the Literature and Emergency Department Management Guidelines. *Ann. Emerg. Med.* 48, 54.e51 (2006).
93 Dwivedi, D. J. *et al.* Differential expression of PCSK9 modulates infection, inflammation, and coagulation in a murine model of sepsis. *Shock* 46, 672-680, (2016).
94 Yamamoto, K., Tatsutani, S. & Ishida, T. Detection of Nausea-Like Response in Rats by Monitoring Facial Expression. *Front. Pharmacol.* 7, 534 (2017).
95 Herrera, C., Bolton, F., Arias, A. S., Harrison, R. A. & Gutierrez, J. M. Analgesic effect of morphine and tramadol in standard toxicity assays in mice injected with venom of the snake Bothrops asper. *Toxicon* 154, 35-41 (2018).
96 Wong S.M.*et al. The Rat Face Finder and improved assessment of visceral pain*. 9th SALAS Annual Regional Conference -Neuroscience: A New Frontier (2013).
97 Ernst, L. *et al.* Improvement of the Mouse Grimace Scale set-up for implementing a semi-automated Mouse Grimace Scale scoring (Part 1). *Lab Anim* 54, 83-91, doi:10.1177/0023677219881655 (2020).
98 Ernst, L. *et al.* Semi-automated generation of pictures for the Mouse Grimace Scale: A multi-laboratory analysis (Part 2). *Lab. Anim.* 54, 92-98 (2020).
99 Andresen, N. *et al.* Towards a fully automated surveillance of well-being status in laboratory mice using deep learning: Starting with facial expression analysis. *PLoS One* 15, e0228059 (2020).
100 Tuttle, A. H. *et al.* A deep neural network to assess spontaneous pain from mouse facial expressions. *Mol. Pain* 14, 1744806918763658 (2018).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 September 2020 doi:10.20944/preprints202009.0101.v1

101 Eral, M., Aktas, C. C., Kocak, E. E., Dalkara, T. & Halici, U. Assessment of pain in mouse facial images. in 2016 *20th National Biomedical Engineering Meeting (BIYOMUT)* 1–4 (IEEE, 2016). doi:10.1109/BIYOMUT.2016.7849416.
102 Littlewort, G., Frank, M., Lainscsek, C., Fasel, I. & Movellan, J. Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior. *2012 IEEE Conference on Computer Vision and Pattern Recognition* 2, 568-573, doi:10.1109/CVPR.2005.297 (2005).
103 Oliver, V. *et al.* Psychometric Assessment of the Rat Grimace Scale and Development of an Analgesic Intervention Score. *PLOS ONE* 9, e97882 (2014).
104 Roughan, J. V., Bertrand, H. & Isles, H. M. Meloxicam prevents COX-2-mediated post-surgical inflammation but not pain following laparotomy in mice. *Eur. J. Pain* 20, 231-240(2016).
105 Zhang, E. Q., Leung, V. S. & Pang, D. S. Influence of Rater Training on Inter- and Intrarater Reliability When Using the Rat Grimace Scale. *J. Am. Assoc. Lab. Anim. Sci.* 58, 178-183, (2019).