Methods used and Application of the Mouse Grimace Scale in Biomedical Research 10 Years On: A Systematic Scoping Review

Alexandra L Whittaker [1]*,Yifan Liu [1], Timothy H Barker [2]

[1] School of Animal and Veterinary Sciences, The University of Adelaide, Roseworthy Campus, Roseworthy, South Australia 5371, Australia

[2] JBI, Faculty of Health and Medical Sciences, The University of Adelaide, South Australia 5005, Australia; timothy.barker@adelaide.edu.au

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

Abstract

The Mouse Grimace Scale (MGS) was developed 10 years ago to assess pain through characterisation of changes in five facial features or action units. The strength of the technique is that it is proposed to be a measure of spontaneous or non-evoked pain. A comprehensive scoping review of the academic literature was performed. The MGS has been employed mainly in evaluation of acute pain, particularly in the pain and neuroscience research fields. There has however been use of the technique in a wide range of fields, and based on limited study it does appear to have utility for pain assessment across a spectrum of animal models. Use of the method does allow detection of pain of a longer duration, up to a month post-initial insult. There has been less use of the technique using real-time methods and this is an area in need of further research.

Keywords: Mouse Grimace Scale, Pain, Validity, Methods, Reliability

1.Introduction

Mice are commonly used as models for a range of conditions in biomedical research. This use is globally significant with approximately 5.7M used in Europe alone. (European Commission, 2019) Many of these models may result in pain or sickness arising either directly, or from other pathological processes. Furthermore, a range of husbandry or routine procedures also undertaken in vivaria may also cause pain or distress. Assessment of affective states in research animals is important to enable implementation of humane endpoints, thus meeting ethical and legal responsibilities, as well as enhancing the translational validity of animal research. However, the assessment of animal emotion is challenging, tending to combine behavioural and physiological measures to provide a holistic assessment. (Boissy et al., 2007; Finlayson et al., 2016; Panksepp, 2005; Whittaker and Marsh, 2019) There has been comparatively more research focus on negative states such as pain, and as such available methods have undergone more extensive testing and validation across a wide range of study types.

Pain is an important issue in animal research for several reasons. Firstly, in recent years there has been growing concern about translational failures from animal studies to the clinic. (Leenaars et al., 2019) Numerous authors have levelled criticism at animal models of pain for being poorly predictive of the clinical scenario, (see e.g. Blackburn-Munro, 2004; Nagakura, 2017; Whiteside et al., 2013) based on issues such as variability between animals and relevance of the assay outcomes to the human pain experience. (Mogil, 2009) This concern is not unique to pain research; pain commonly arises in other disease conditions and may be a target for novel therapeutics. Secondly, pain and its sequelae may influence the results obtained from animal model studies, affecting a range of physiological and immunological processes occurring. This further impacts on the reliability and translatability of the results obtained from these studies. (Carbone and Austin, 2016; González-Cano et al., 2020; Peterson et al., 2017) Finally, pain presents a significant cost to animal welfare through the impact on individual animals. Therefore, the assessment of pain and application of methods to mitigate its effects, are needed to safeguard animal welfare and to conform to ethical requirements in biomedical research, for instance the refinement aspect of the 3Rs. (Russell and Burch, 1959) This assists in addressing societal concerns around the use of animals in research.

One of the more commonly used assessment methods, suggested to be specific to pain, is the use of facial expression scoring or the so-called 'grimace scales'. (Mogil et al., 2020; Whittaker and Howarth, 2014) The idea behind using facial expressions as a readout for pain neurobiology came from human facial codification scales. (Nagakura et al., 2019; Serizawa et al., 2019) The Facial Action Codification System (FACS) allows categorization of movements of the facial muscles. Specific combinations of movements leads to changes in discrete facial regions or "facial action units (FAU)", for instance the closing of the eyelids. Recognition of changes in these FAUs has been proposed to allow determination of emotional state. (Descovich et al., 2017; Ekman, 1992; LeResche, 1982) Grimace scales were developed for non-human animals, with the goal of standardizing methods for different species. The original grimace scale was developed for mice by Langford and colleagues in 2010, and validated through application of a variety of preclinical pain assays. In this scale, changes in 5 facial action units are assessed to determine level of pain: (1) orbital tightening, (2) nose bulge, (3) cheek bulge, (4) ear position and (5) whisker change. Grimace scale development in other species followed (see Mogil et al., 2020 for full history), as did further examination of the mouse grimace scale (MGS) in a range of animal models and conditions.

There have been a number of reviews on grimace scales in a variety of species, (Descovich et al., 2017; McLennan et al., 2019; Mogil et al., 2020; Mota-Rojas et al., 2020) but none to our knowledge that have focussed solely on mice, and used systematic methods to identify all studies where the MGS was utilised. Now 10 years on from the publication of the original study, a comprehensive systematic assimilation of the evidence on the MGS is warranted; mice being the most commonly used mammal in biomedical research. (Homberg et al., 2017) In contrast to a systematic review and meta-analysis, scoping reviews are broader in scope and bring together all current evidence, regardless of quality (Colquhoun et al., 2014). They may also pave the way for future systematic reviews on a clearly defined question identified in the scoping review. Therefore the aim of this systematic scoping review was to identify all published studies on the MGS and assimilate the evidence based on features of the scale use, with a particular focus on the application of the technique across a range of animal models, the methods used, and the impact of external variables on validity and reliability. This review will provide increased strength of evidence to guide researchers, ethics committees, and policy makers on the use and application of the MGS in biomedical research.

2.Methods

JBI guidelines on conducting systematic scoping reviews were used to guide review performance and reporting. (Peters et al., 2015) A protocol for this review was not registered since common protocol databases (e.g. PROSPERO) do not accept scoping review protocols.

2.1 Search Strategy

The search strategy aimed to locate published studies in English. An initial limited search of Medline was undertaken to identify articles on the topic. The text words contained in the titles and abstracts of relevant articles, and the index terms used to describe the articles were used to develop a full search strategy for Medline via Pubmed using MeSH and free text terms. The search strategy was adapted for Scopus and Web of Science (including CAB abstracts) database searches. The three databases were searched in May 2020 using the developed search strategies (see Appendix A). The search was updated in October 2020. Key concepts used for searching were "mice" and "grimace scale". Hand searching of reference lists was performed to identify additional studies. Studies published from database inception were eligible for inclusion. Publications were excluded electronically if they were conference abstracts with full study detail and results not available, or review articles.

2.2 Eligibility Criteria

Studies were included if they investigated the Mouse Grimace Scale in mice irrespective of age, sex or strain. Studies that looked at a change in any number of facial action units but that did not report this as use of a 'grimace scale' were excluded. Studies that used the MGS and reported it as such but modified the method slightly were however eligible for inclusion. Only studies that investigated the MGS based on an understanding that this was a measure of pain were eligible, for example a study using the MGS to assess positive emotion would have been ineligible for inclusion. All study designs were eligible for inclusion. Studies investigating new ways of collecting MGS data, for example by automation techniques were excluded. However, studies evaluating the objective nature of the test, for example those studies examining reliability between observers or institutions were eligible for inclusion.

2.3 Study Selection

Following the search, all identified citations were collated and uploaded into EndNote X8.0.1 and duplicates removed. Potentially relevant studies were retrieved in full and their citation details imported into Covidence (Veritas Health Innovation, Melbourne, Australia). Titles were screened by one reviewer (AW) for assessment against the inclusion criteria for the review. Abstract and full text screening were performed by all authors (AW, YL, THB) with two independent reviewers being required to certify the inclusion of each study. Disagreements that arose between the reviewers at each stage of the study selection process were resolved through discussion with the third reviewer.

2.4 Data Extraction

Data were extracted from the included studies by three independent reviewers (AW, YL, THB) using an electronic form developed by the authors (Appendix 2). All reviewers initially performed independent review of the same 3 studies (Chartier et al., 2020; Dwivedi et al., 2016; Hassan et al., 2017) to pilot the extraction tool and check for data consistency. Following this the remaining studies were allocated between the 3 reviewers; each study being extracted by one reviewer. Distribution of the papers between the data extractors was done randomly. Only data directly relevant to the research question were extracted. All data extracted were reviewed by the authorship team to ensure completeness of extraction. Contact with study authors was undertaken where necessary to clarify findings or seek further information. In accordance with guidelines on systematic scoping reviews, (Peters et al., 2015) the goal of the review was to provide an overview of evidence on the MGS regardless of quality. Hence methodological quality assessment of included studies was not undertaken.

3. Results

3.1 Study Characteristics

A total of 240 articles were retrieved. Six studies were retrieved through hand searching of the reference lists of included studies or forward citation searching. Following title and abstract screening 59 articles were assigned for full-text retrieval with 48 articles being included at full-text review (Figure 1). The reason for the majority (n=7) of the exclusions after full text review, was due to the studies evaluating MGS automation methods, rather than pain in mice. The characteristics of the included studies are presented in Table 1. Observational studies were eligible for inclusion. However, the majority (92%) of the studies (n=43) adopted an experimental study design, using the typical randomized controlled trial (RCT) design or pseudo-RCT design (where allocation to groups is systematic and not random). The remaining studies adopted a quasi-experimental design, such as using a pre-test, post-test repeated measures design with no group running in parallel. Since the first report of the MGS by Langford and colleagues in 2010, the number of publications investigating the method has grown considerably to a current approximately steady state rate of around 6-9 publications per year, sustained over the last 5 years (Figure 2).

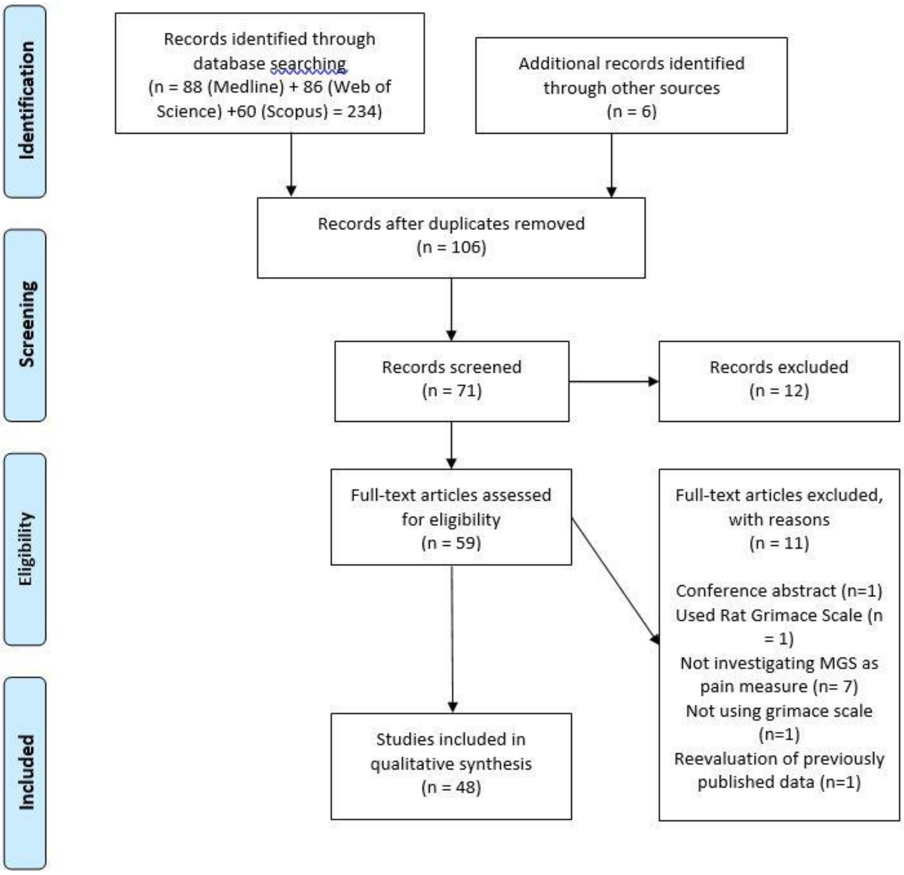

Identification
Screening
Eligibility
Included
Records identified through database searching (n = 88 (Medline) + 86 (Web of Science) +60 (Scopus) = 234)
Additional records identified through other sources (n = 6)
Records after duplicates removed (n = 106)
Records screened (n = 71)
Records excluded (n = 12)
Full-text articles assessed for eligibility (n = 59)
Full-text articles excluded, with reasons (n = 11)
Conference abstract (n=1)
Used Rat Grimace Scale (n = 1)
Not investigating MGS as pain measure (n= 7)
Not using grimace scale (n=1)
Reevaluation of previously published data (n=1)
Studies included in qualitative synthesis (n = 48)

Figure 1. The PRISMA (Moher et al., 2009) flow diagram for the systematic review detailing the database searches, the number of abstracts screened and the full texts retrieved.

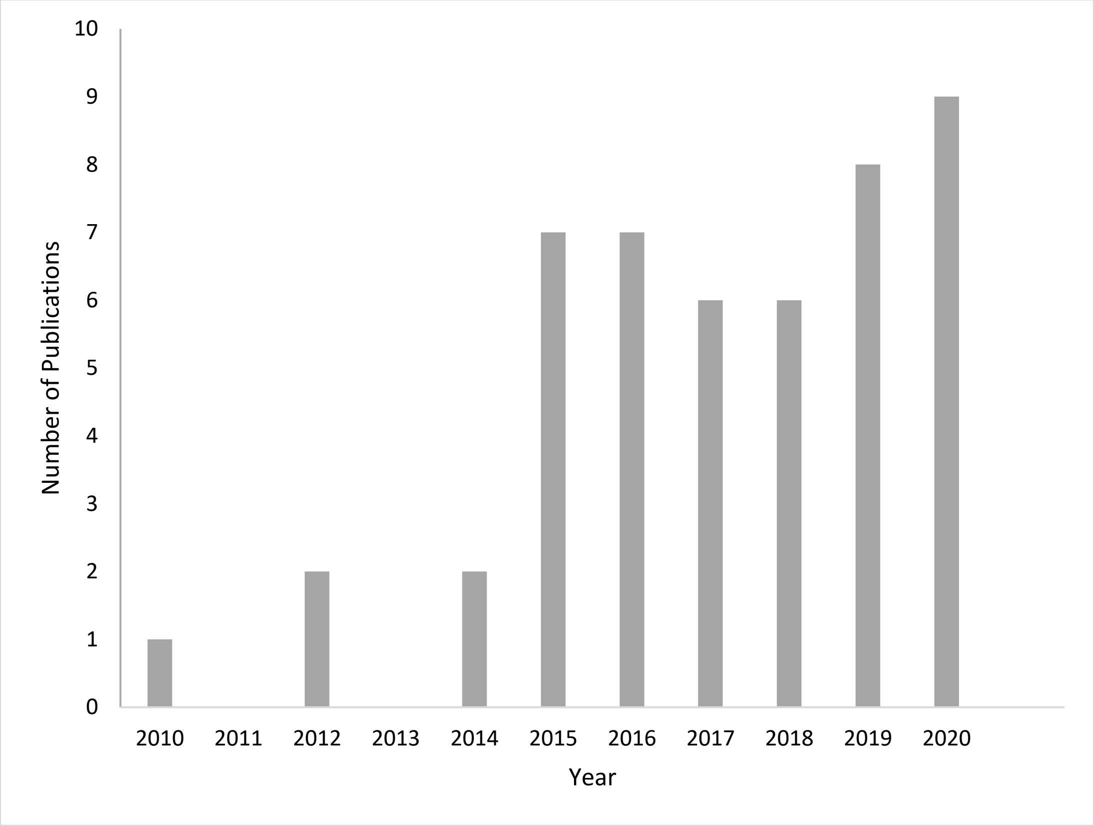


Figure 2. Number of publications per year investigating the MGS in relation to pain

3.2 Animal Model Characteristics

Studies were allocated into three categories based on the types of interventions applied to the mice for subsequent grimace score measurement. The categories considered were 1) animal model, 2) husbandry/procedural and 3) biological. Studies were categorised as utilising animal models if they used an animal model of a human condition likely to cause pain. Husbandry/procedural grouping was applied if the study investigated procedures commonly performed as part of laboratory routines, breeding procedures or veterinary treatments including anaesthesia and analgesia provision. The biological classification was reserved for those studies that investigated grimace scores resulting from inherent biological variation such as between sexes and strains or as a result of difficult to control environmental variables such as circadian rhythms. Based on our classification 65% (n=31) of studies used animal models, 31% (n=15) looked at husbandry/procedural interventions and 4% (n=2) investigated biological variation in grimace scores. It was considered that the interventions applied would lead to pain arising of substantially different natures. We utilised a published pain classification system (Melnikova, 2010) for assignment of studies based on pain type (Figure 3). Figure 4 presents a sub-classification of the type of animal models or procedures used in the included studies, with expected pain type resulting. The animal model groupings are based on that presented by Hau and Shapiro, 2010. It should be noted that whilst some studies may have had a primary focus on evaluating response to one intervention, they may have reported on impact of other factors, for example sex differences. In reporting, we have considered evidence from all studies irrespective of the classification assigned.

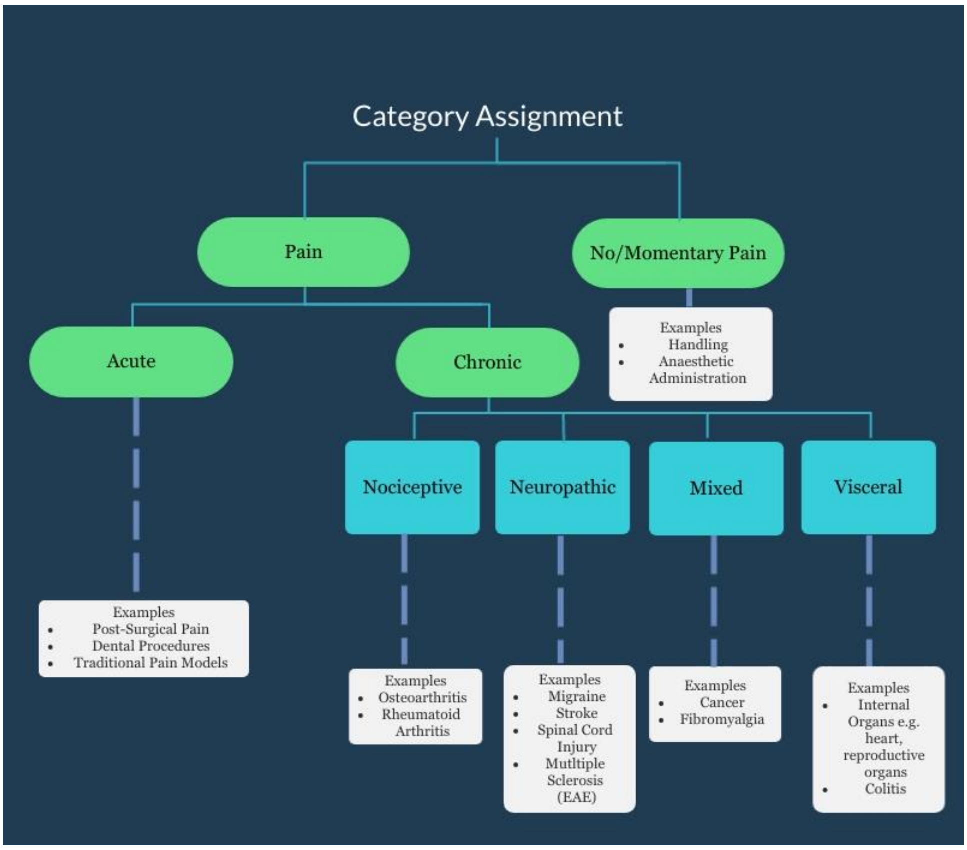
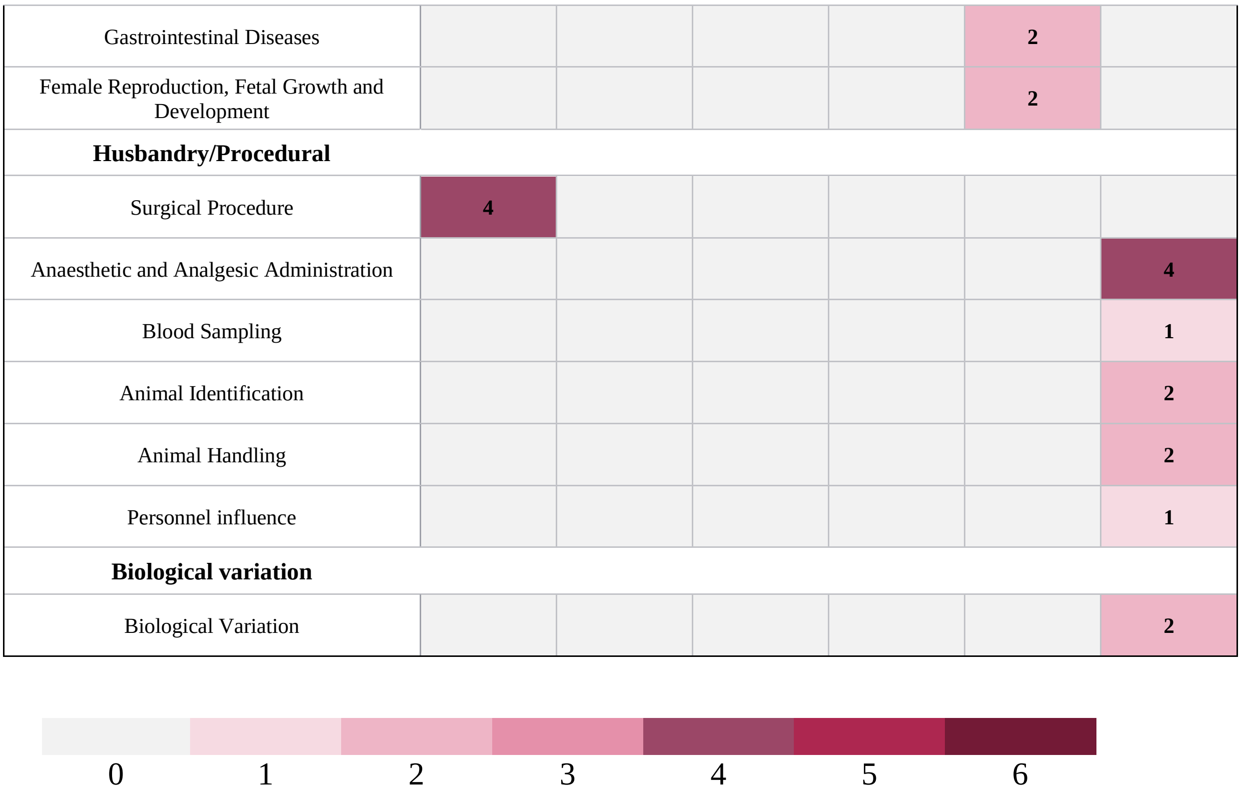


Figure 3. Pain classifications used to guide assignment of studies to categories. Adapted by permission from Springer Nature and Copyright Clearance Center: Springer Nature, Nature Reviews Drug Discovery, Pain Market, Melnikova, I, COPYRIGHT 2010. For specific category assignment for included studies refer Table 1.

	Acute	**Chronic**				
		Nociceptive	Neuropathic	Mixed	Visceral	None expected or momentary
Animal Models						
Pain Research, includes Migraine/Headache	**5**		**5**		**1**	
Cancer Biology				**2**	**1**	
Cardiovascular Diseases					**1**	
Obesity and Diabetes Mellitus						**1***
Sepsis					**2**	
Hematopoietic Diseases	**1‡**					
Oral Health Sciences	**5**					
Neuroscience, includes EAE			**3**			
Pharmacology and Toxicology					**1**	



Gastrointestinal Diseases					2	
Female Reproduction, Fetal Growth and Development					2	
Husbandry/Procedural						
Surgical Procedure	4					
Anaesthetic and Analgesic Administration						4
Blood Sampling						1
Animal Identification						2
Animal Handling						2
Personnel influence						1
Biological variation						
Biological Variation						2

Figure 4. Heat map contrasting interventions used by the type of pain expected to be elicited. Colouration/number in box represents number of studies. Whilst some studies could arguably have been included in multiple categories to simplify reporting one category has been assigned. ǂ Pain in this study, Mittal et al., 2016, was assigned as acute since induced by cold stress, although sickle cell pain can be neuropathic in origin. * Model of Hsi et al., 2020 did not relate to a neuropathy.

3.3 Mouse Characteristics

The included studies used a wide range of inbred strains and outbred stocks of mice. The C57BL/6 strain was used in the majority of studies (38% of uses), followed by the outbred ICR/CD-1 (24%). Transgenic or knockout/in strains of specific relevance to the research questions investigated in the publications were commonly used (14%). Figure 5 illustrates the relative uses of the various strains. Excluding the mutant, transgenic and other categories 45% of the mice used were black- coloured, 44% white-coloured and 11% brown/agouti. Considering standard inbred or outbred strains/stocks only, eight studies used more than one strain. (Cho et al., 2019; Miller et al., 2015; Miller and Leach, 2015a, 2016; Rea et al., 2018; Rosen et al., 2017; Sorge et al., 2014; Tillu et al., 2015) Only 3 of these studies directly contrasted grimace scores between the strains. (Cho et al., 2019; Miller et al., 2015; Miller and Leach, 2015a) The direction of effect for grimace scores in these comparisons are presented in Table 2. There are some differences in strain effects on grimace scores between the sexes.

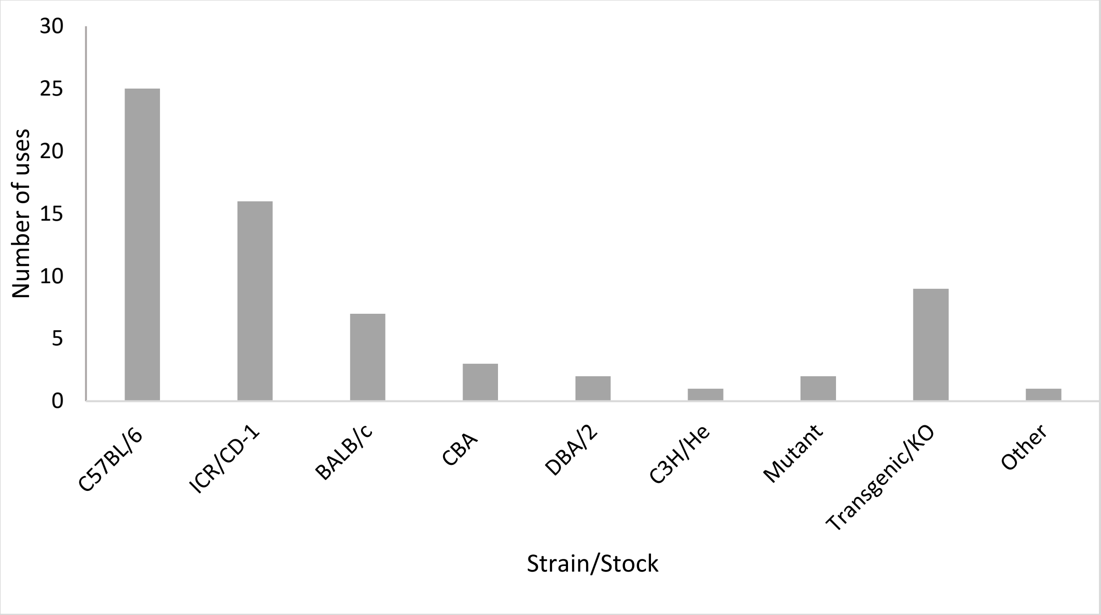


Figure 5. Representation of the different mouse strains and stocks used in the published papers. Note that a number of papers used more than one strain. The ICR and CD-1 nomenclature has been considered to represent the same stock. Other includes hybrid or recombinant strains.

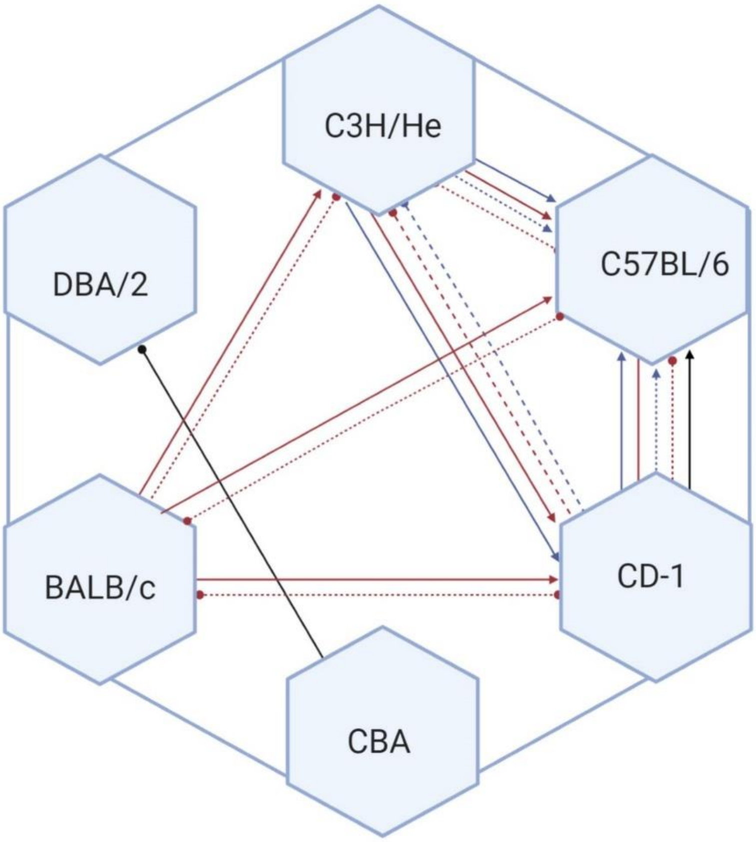


Figure 6. Network map comparing MGS scores between strains. Each line represents a study effect. The direction of the arrow represents that the strain at the arrowhead responded with a lower MGS score. Red lines indicate a comparison between female mice, blue lines indicate comparison between male mice and black lines indicate comparisons where sex was not separated. A solid line indicates that a live score was used, a dashed line indicates that a retrospective score was used.

Male mice only were investigated in 40% (n= 19) of the studies, females in 21% (n=10) of the studies, with 36% (n=17) of the studies investigating both sexes. Sex of mice was unreported in one study (Table 3).

Reference	Intervention	Strain	Direction of effect (F v M)
Burgos-Vega et al., 2019	Migraine	ICR	Not directly compared
Cho et al., 2019	Craniotomy with different analgesics	CD-1 C57BL/6N	= =
Dwivedi et al., 2016	Caecal Ligation and Puncture model of sepsis	Transgenic (BL/6 background) PCSK9 KO mice and PCSK9 overexpression	Not reported
Guo et al., 2019	Orofacial pain	C57BL/6	Not reported
Hohlbaum et al., 2017	Isoflurane anaesthesia	C57BL/6JRj	=
Hohlbaum et al., 2018	Repeated ketamine anaesthesia	C57BL/6JRj	=
Kim et al., 2015	Hyperalgesic priming via IL-6 and Carrageenan injection	ICR	Not reported
Langford et al., 2010	14 models of pain	CD-1 (ICR:Crl)	=
Matsumiya et al., 2012	Ventral ovariectomy and response to analgesics	CD1 (ICR:Crl)	=
Miller and Leach, 2015a	Biological	 C57BL/6 C3H/He CD-1 C57BL/6 C3H/He CD-1	**Live Scoring** = F< M F < M (at 1 time point) **Retrospective Scoring** F > M = =
Mittal et al., 2016	Sickle cell disease and effects of cold	Transgenic HbSS-BERK (with relevant controls)	F > M

Rea et al., 2018	CGRP- induced migraine	C57BL/6J CD1	=
Rosen et al., 2017	CD-1 (Crl:ICR) , Nude (Crl:CD1-Foxn1 nu), C57BL/6J, C57BL/6-Rag1 tm1Mom, mutant mice lacking expression of the Oprd1 (-opioid receptor) gene	Pregnancy analgesia	Not directly compared for grimace outcome
Rossi et al., 2020	Tooth pulp injury	Mixed CD1 and C57BL6/J background	=
Roughan and Sevenoaks, 2019	Ear tattooing and tagging, with tail handling method or tunnel	BALB/cAnNCrl	F < M
Sorge et al., 2014	Effect of gender/gender-specific and other animal pheromones on response to nociceptive assays	CD-1 (ICR:Crl) C57BL/6J	F > M (baseline values) Females displayed greater 'male observer' effect e.g. increased reduction in grimace scores.
Tuttle et al., 2018	Ventral ovariectomy and response to analgesics, xymogen assay (validation of automated scoring)	CD-1 (ICR:Crl)	=

Table 3: Comparison of MGS scores between sexes.

3.4 MGS Measurement Methods

The majority (88%) of studies evaluated MGS by retrospective scoring via photographs obtained directly via camera use, or extracted as stills from video footage, as reported in the original study. (Langford et al., 2010) To date only 5 studies have used real time methods, (Bu et al., 2015; Chartier et al., 2020; Gallo et al., 2020; Hsi et al., 2020; Miller and Leach, 2015a)'with 3 of these studies directly contrasting these results with those obtained from retrospective scoring. (Chartier et al., 2020; Gallo et al., 2020; Miller and Leach, 2015a). One study, did not state the method of MGS scoring. (Kim et al., 2015) The breakdown of collection method and timing is detailed in Figure 7.

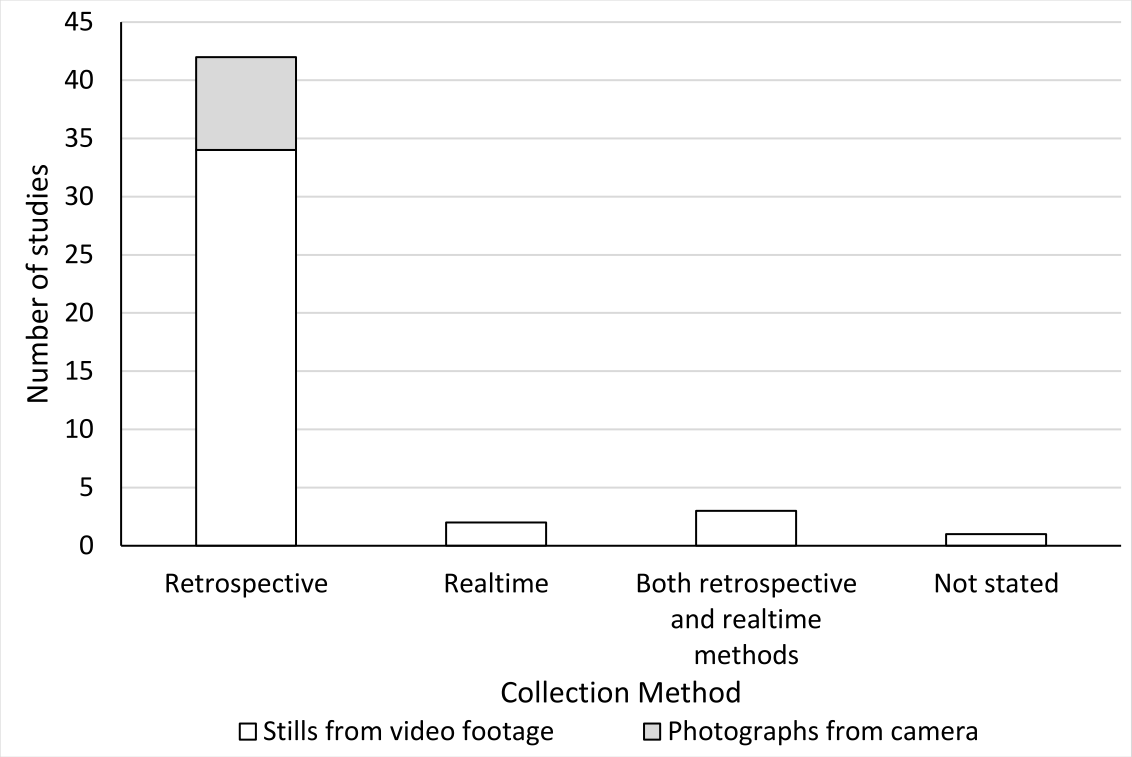


Figure 7: Collection methods used in included studies.

Real time methods used varied between the studies. In Miller and Leach 2015a, the mice were observed 3 times during a 10 minute period and scored on each action unit as per the original method to arrive at 3 scores for the period. Gallo et al., 2020 assigned a single score to each mouse after a 30s observation period, and Bu et al., 2015 assigned a single score on provoking abdominal pain in a pelvic pain model. The methods used in Hsi et al., 2020 are unclear but real time methods appear to have been used. However, Chartier et al., 2020 assigned a score every 15 s for 15 minutes and then calculated final scores based on averaging of scores across three 90 s periods to account for any effect of novelty of the box.

In the studies performing direct comparison, live scores were found to be significantly lower than corresponding retrospective scoring in two of the studies. (Chartier et al., 2020; Miller and Leach, 2015a) In the final study, (Gallo et al., 2020) a PCA produced a component where real time MGS and image scoring were highly intercorrelated (with nesting behaviour as a third factor).

The original study described the MGS in terms of 5 FAU's. However, in 18 (38%) of the studies scoring was modified by excluding specific action units, or in one case combining the cheek and nose bulge action unit into one. (Mai et al., 2018) In the studies that used 4 action units for scoring, whisker position was the action unit excluded in the majority (60%) of cases (Figure 8). The method of combining the scores to arrive at a final score for the photograph or time point (real time scoring) was in the majority of studies (36/48) by averaging of individual action unit scores (yielding a maximum score of 2). In 10 studies, summation of the individual action units scores was performed to arrive at the final score (maximum score of 10 for 5FAUs). The method of achieving the final score was unclear in the remaining 2 studies. (Hassan et al., 2017; Mitchell et al., 2020) A number of studies accounted for individual responses to pain by using mean difference scores in data presentation and analysis to correct for baseline grimace scores. For studies where the whisker position FAU was excluded, 50% of the studies used mice (6/12) that were black coloured, 33% (4/12) white, and 17% (2/12) brown coloured (X^2(2, N = 12) = 3, p=0.22).

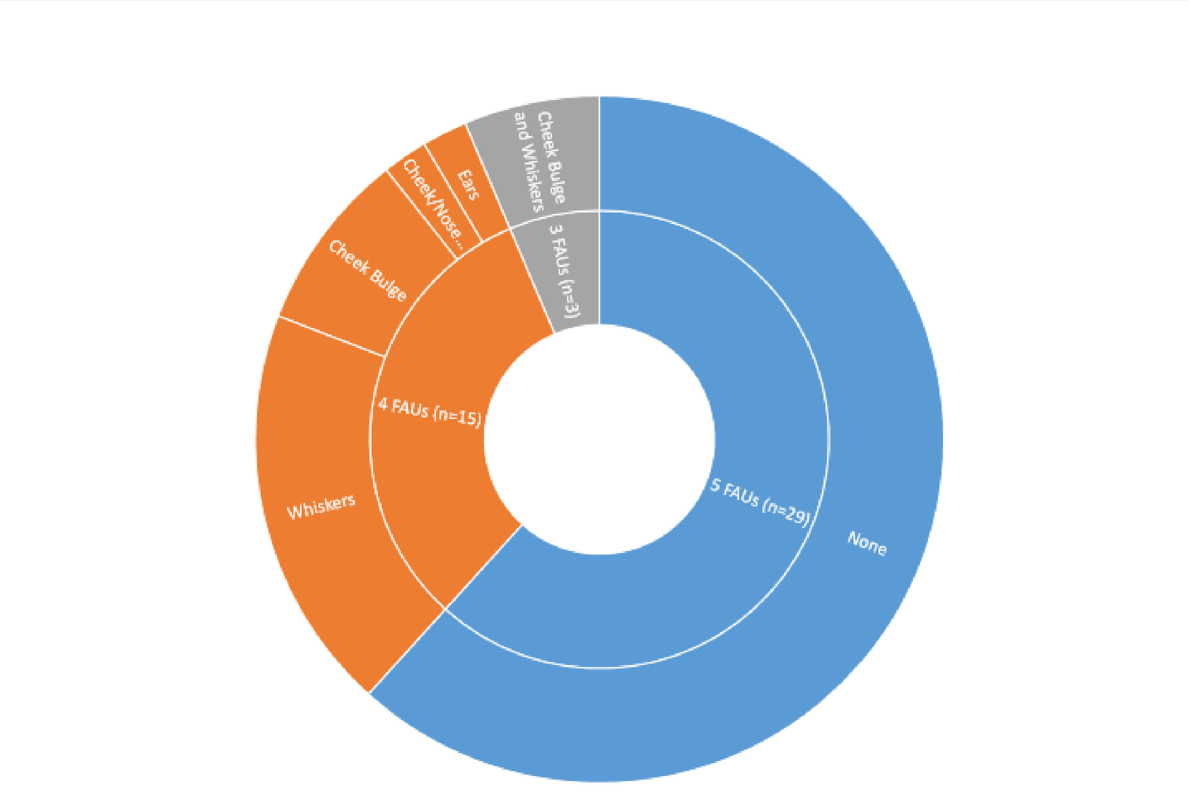


Figure 8: Facial Action Units (FAU's) utilised for scoring in the included studies. n represents study number. Specific action units were generally excluded as described here, although in one study two of the action units were combined.

A range of study durations were used in included studies, often with multiple time points being assessed within a single study. Duration of MGS assessment ranged from directly after the intervention to over a month following. This is illustrated in Figure 9, categorised by expected pain type. Refer to Table 1 for detail of interventions applied in the studies.

	Timepoint				
	≤ 24hrs	≤ 72hrs	≤ 1 wk	≤ 1 mth	> 1 mth
Pain Classifications					
Acute	100% (15/15)	100% (7/7)	50% (3/6)	0 (0/1)	
Visceral	88% (7/8)		0 (0/1)	100% (1/1)	0 (0/1)
Neuropathic	100% (5/5)	0 (0/2)	66% (2/3)	100% (4/4)	0 (0/1)
Mixed	100% (1/1)		100% (1/1)	100% (2/2)	
None/momentary*	45% (5/11)	100% (3/3)	100% (1/1)		

Not examined	<20%	20-40%	40-60%	60-80%	80-100%

Figure 9. Heat map contrasting type of pain expected to arise from the interventions with the time points after the intervention investigated. Colouration gradation represents percentage of studies where grimace scores moved in the expected direction of effect, with increased shading indicating greater number of investigations, for example 100% of studies evaluating procedures likely to cause acute pain showed increased MGS scores within the 24 hours after the intervention. * Consider that no change in MGS score is expected.

3.5 Corroborating Methods of Affective State Assessment Used

A range of alternate methods for assessing animal affective state were utilised in 37/47 (79%) of the included studies (Figure 10). These methods were largely behavioural in nature but did include measures of physiology, such as corticosterone analyses or bodyweight (being an expression of feeding behaviour). The most common measures used in rank order were: use of Von Frey filaments for assessment of mechanical allodynia, bodyweight, general clinical/disease scoring which may have been tailored to the model used e.g. EAE scoring scheme, burrowing behaviour, pain-related behaviour scoring such as the use of composite pain measures, and open field tests for activity and locomotion. In the majority of cases (31 studies), data from these tests corroborated MGS scoring. In the remaining studies, either no association was seen with the chosen measures, (Chartier et al., 2020; Miller et al., 2015; Mitchell et al., 2020; Zhu et al., 2017) or there was unclear reporting or lack of direct comparison in the same animals. (Guo et al., 2019; Hsi et al., 2020).

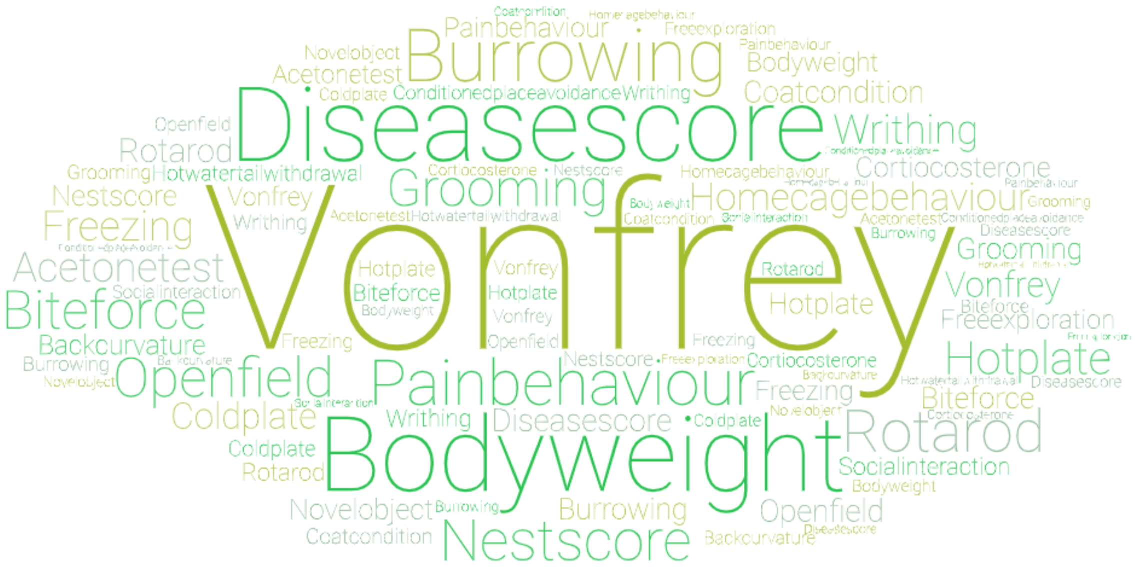


Figure 10: Word cloud illustrating corroborating methods of affective state assessment used in the included studies. Size of the word illustrates their relative frequency of use.

3.6 Impact of external factors on MGS

3.6.1 Circadian Rhythm

In the majority of the studies there was no specific reporting of light cycle stage for recording of MGS data. It was assumed that given the lack of reporting these were performed during the light stage. Five (11%) of studies either reported conducting recording during the dark stage or timelines of measurement suggested that both stages would be crossed. (Dwivedi et al., 2016; Jurik et al., 2014; Matsumiya et al., 2012; Miller and Leach, 2015a; Rea et al., 2018) However, only three of these studies performed an examination of circadian rhythm effects. (Matsumiya et al., 2012; Miller and Leach, 2015a; Rea et al., 2018) studies and their impact on the MGS are reported in Table 4.

Reference	Reporting Detail	Intervention	Compared with measures in light (Y/N)	Direction of effect for comparison between light and dark
Matsumiya et al., 2012	Conducted circadian study comparing light and dark recordings	Ventral ovariectomy and response to analgesics	Y	Compared mice which had surgery in the morning versus the evening with measurement timepoints of baseline, and every 6 hours past surgery for 48 hrs. There was no circadian effect on baseline MGS scores. However, mice operated on in the morning displayed larger MGS increases12 h after surgery compared to 24 h, whilst mice operated on in the evening showed smaller increases at these time points. This suggests that mice experience higher levels of postoperative pain at night (dark phase)
Miller and Leach, 2015a	No dark cycle recording but did compare MGS across the light phase	Biological	N	**Live scoring** There was no difference in MGS score between three time points (9am, 12.30pm, 4pm) for C57BL/6, CD-1 or C3H/He mice. BALB/c mice showed a greater score at Noon compared to AM. **Retrospective Scoring**

				There was no significant difference in MGS scores between the three time points for CD-1, C3H/He or BALB/c mice. C57BL/6 mice showed a greater MGS scores at both Noon and PM time points compared to the AM time point
Rea et al., 2018	Performed a restrained grimace technique in dark (manipulated dark condition- not part of cycle)	CGRP-induced migraine	Y (compared bright light)	Grimace scores were higher in the dark than in bright light for the CD1 mice. Light transition led to decreased orbital tightening and nose bulge. C57BL/6J mice showed no significant difference between the CGRP-induced grimace in light and dark. Responses to CGRP were generally similar in direction as those recorded in the light.

Table 4: Studies that utilised grimace scoring in the dark stage of the circadian cycle and the impact on scores where reported.

3.6.2 Variability Arising From Observers

A number of studies (20/48) utilised more than one observer for ascertaining grimace scores. Ten of these studies (Table 5) specifically reported the metrics associated with agreement between the observers, that allowed them to combine the results with an assurance of external reliability.

Reference	**Number of Observers**	**Consistency Metrics**
Inter-Observer Variability		
Faller et al., 2015	2	There was an excellent correlation between the two observers for MGS measurement (r = 0.98) assessed using Type II regression analysis. However, Bland–Altman analysis showed that the slope differed from unity with a bias towards higher MGS scores in one observer.
Hohlbaum et al., 2020	4 (2 Novice, 2 Expert Scorers)	Good agreement between all observers was observed (ICC = 0.851) when all three time points were examined. However, interrater reliability differed across timepoints. The best agreement was achieved for orbital tightening, and the poorest agreement for nose and cheek bulge, and this depended on the observers' experience levels. In general, experienced observers produced scores of higher consistency when compared to inexperienced.
Langford et al., 2010	7	Inter-rater reliability was high as assessed by intra-class correlation coefficient (ICC average = 0.90). When high-definition video cameras were used, over 97% of pain versus no-pain images were categorised correctly.
Mittal et al., 2016	6	ICC and Cronbach's alpha values were low (ICC average <0.7, α < 0.8). This resulted from large intra-coder variability for three of the coders. Therefore, only the results of the coders with low variability were used in data presentation (updated metrics not reported).
Rea et al., 2018	2	Correlation coefficients ranged between 0.89 and 0.92.
Roughan et al., 2016	4	There was high inter-observer consistency, with ICC values ranging from 0.75-0.84.
Roughan and Sevenoaks, 2019	6 Novice and 6 Expert Scorers	The α values for experts and novices were high (0.88 to 0.94; 0.78 to 0.87 respectively). Agreement between novices and experts was generally good (ICC ranging from 0.7 to 0.84 across the timepoints).
Sorge et al., 2014	2	Moderate to high inter-rater correlation (r = 0.64, P < 0.001). Group data from one rater compared to the other were almost identical.
Tuttle et al., 2018	2	High inter-rater consistency with Cronbach's alpha of 0.89.
Inter- Laboratory Variability		
Jirkof et al., 2020	3	Median MGS scores were significantly different at a number of timepoints between the 3 laboratories. They were however qualitatively similar i.e. direction of effect.

Table 5. Consistency metrics reported in the included studies. Inter-observer and inter-laboratory analyses were reported.

4. Discussion

In this paper we have presented the first comprehensive overview of all studies investigating the MGS, assimilating information on the types of animal models/conditions where the MGS has been applied, methods applied, and external factors affecting validity of the technique. It is hoped that this assimilation will guide future validation, and use of the MGS by researchers and thus promote wider scale implementation of the method. Key findings of our assimilation are discussed below.

4.1 Methods Used

To date the majority (88%) of uses of the MGS in biomedical research settings have used retrospective recording through collection of video footage, and subsequent still extraction, or primary collection of photographic images. Retrospective scoring brings some key advantages when using the MGS as a research outcome measure. These methods provide a greater degree of certainty in the findings by allowing for the possibility of re-confirming scores and thus replicating the data, utilising multiple observers for cross-checking, and allowing scoring to occur at a time that suits the researcher. (Mota-Rojas et al., 2020) This can all occur without the potential modulating influence on the scores of a human observer. (Sorge et al., 2014) Whilst, not discussed in the included studies an assumed challenge in using cameras to secure facial images is the need to achieve a face-on shot. This might be achieved by using a 'burst' mode to take photos in rapid succession, or by manual performance by an observer. However, this does raise concerns about the effect of observer presence on grimace scores and the impact of any noise produced by the camera when photographs are taken.

A real time method has advantages for clinical pain assessment, since scores can be attained quickly, to allow immediate action such as applying a humane endpoint or providing analgesics. The method may also provide some advantages in a research scenario by limiting the need for post-processing of images, which is invariably time consuming. (Mogil et al., 2020) To date there has been limited evaluation of real time scoring in mice, and of the five studies that have utilised this, only three directly contrasted this with validated retrospective scoring methods. Two studies found live scores to be lower than corresponding retrospective scoring. (Chartier et al. 2020; Miller and Leach, 2015a) A reason proposed for the lower scores resulting from live scoring is that the nature of the face changes rapidly during live scoring whereas in images, for example, random selection will lead to capture of blinking which is assigned a high score, contributing to relatively higher scores. (Miller and Leach, 2015a) Alternately, as proposed by Chartier et al. 2020 the presence of a human observer in real time scoring may influence mouse performance of the facial action units; increased alertness could lower the grimace scores through eye widening and 'pricking' of ears. It should be noted that there are considerable differences in the technique used for collection of real time data with some studies basing a score on a single observation point, (Bu et al., 2015; Gallo et al., 2020) as opposed to mathematical integration of several scores taken across a period. (Chartier et al. 2020; Miller and Leach, 2015a) The former would be simpler in a clinical context but may be associated with loss of sensitivity and validity. In spite of this, point grimace scores were determined to move in the expected direction of effect in these studies, implying validity. In rats there has been dedicated study into methods of real time scoring and their relationship with retrospective scoring, (Leung et al., 2016; Leung et al., 2019) and this is clearly needed in mice.

Whilst 62% of studies did use all of the five original described action units for scoring, in a significant proportion of studies (37%) scoring was modified by excluding specific action units, or combining units. Most of these adaptations involved excluding whisker scoring, which seems to be regarded as hard to visualise/score. (Mogil et al., 2020) It has been suggested by some authors that this difficulty in scoring whiskers is related to black coat colour. (Cho et al., 2019; Mai et al., 2018) However, this proposition is not supported by our synthesis which implies that whiskers are excluded from scoring at similar rates independent of coat colour (although study numbers are low). There may also be an impact of inexperience in scoring on ability to accurately identify action units, for instance Hohlbaum et al., 2020 demonstrated that cheek and nose bulge scoring had reduced inter-observer agreement compared with orbital tightening, with inexperienced scorers having even reduced accuracy.

4.2 Validity of the MGS across a range of pain types

The MGS is described as a measurement of pain i.e. it has face validity for pain. There is clear evidence from the included studies that the MGS changes in response to painful events and is modified by analgesics, further supporting this proposition see eg. (Faller et al., 2015; Leach et al., 2012; Matsumiya et al., 2012) However, another important aspect of the validity of a pain measure is the extent to which the technique measures pain, and is not influenced by other conditions such as sickness behaviour, in other words whether it has construct validity. This review assists in evaluating these concepts in a number of ways.

It is clear that whilst the majority of the studies examining the MGS are conducted by researchers in the pain field, there has now been use of the technique across a range of non-pain focussed animal models. The technique being especially utilised in the oral health science and neuroscience fields. There has also been significant focus on the technique in husbandry and welfare investigations in mice, with a focus on the effects of surgery and analgesic administration on the score, and by inference pain. In the majority of these models, especially over an acute timeframe the MGS has good utility. However, even though use of the technique has increased over the past decade 48 studies is a small fraction of all the studies being conducted in laboratory mice. It is surprising that more researchers have not taken the opportunity to include the technique in their study. This may be due to a lack of awareness by researchers outside the pain and veterinary research fields of both the technique, and its validity. It is hoped that this review will promote awareness to these researchers, but there is probably a significant role for animal ethics committees in this dissemination effort.

In the original study by Langford et al, 2010 it was considered that the MGS was only suitable for measuring acute pain, based on the lack of grimace response when models of chronic pain were applied. This would make sense from an evolutionary perspective since, as prey animals, mice may learn to control a facial pain response to avoid predation. (Matsumiya et al., 2012) However, later studies question this assumption. Figure 8 provides clear evidence that across a range different expected pain types, grimace scores are detected up to a month post-initial insult in situations where pain might be expected. This evidence is particularly strong for neuropathic pain which might be expected to be longer lasting and has been investigated in a reasonable number of studies. In visceral or mixed pain the MGS also appears to be able to detect an effect but there have been limited studies, and it should be noted that the studies into mixed pain both come from the same laboratory looking at pain in one model of breast carcinoma. (de Almeida et al., 2019; de Almeida et al., 2020) There is clearly a need for future study in models where these types of pain are expected. To date, no studies have shown the existence of changed grimace scores at timepoints greater than a month after the assumed painful treatment. However, only two studies specifically looked at these timepoints and there is the possibility that pain was not actually present at these times, especially in one of the studies, which utilised a relapsing-remitting colitis model induced by DSS. (Chartier et al., 2020) Interpreting findings at these later timepoints is made more challenging given the lack of other validated measures of pain against which to corroborate MGS findings.

A range of physiological and behavioural outputs were measured in the included studies which lend support to the proposition that the MGS has good construct validity. These included the use of assessment of mechanical allodynia, general clinical scoring, pain behaviour scoring or indicators of luxury behaviour such as nest building or burrowing. In the main, outcomes from these tests moved in the same direction as mouse grimace scores, suggesting convergent validity. However, out of all the measures assessed, arguably only a couple are specific to pain and are plagued by the same issue that surrounds MGS validation; that of establishing incontrovertibly what they are measuring. For example, burrowing and nest building behaviour are largely taken to be generalised indicators of well-being or affective state, (Jirkof, 2014) and are modified not just in response to pain, but sickness behaviours see e.g.(Cunningham et al., 2007; Gaskill and Pritchett-Corning, 2016; Jirkof et al., 2013; Whittaker et al., 2015). Composite pain behaviour scoring and use of Von Frey testing are specific to pain and therefore more reliable corroborating measures. However, the debate around the differences between nociception and pain needs to be borne in mind (see Deuis et al., 2017 for full discussion). The former is a physiological function, but a reaction to a stimulus does not necessarily signify the *experience* of pain. Therefore the widespread historical use of stimulus-evoked tests, such as the Von Frey filaments, may be a contributing factor to the poor translation rates in pain research. (Deuis et al., 2017) One of the key cited advantages of the MGS is that it measures spontaneous pain. (Mogil, 2009) Based on this discussion perhaps the most reliable corroborating measure against which to assess the MGS is another readout of spontaneous pain, with composite pain behaviour scoring being the only measure to completely fulfil this description. In the studies that compared these two readouts, the direction of effect was aligned but the studies are few in number (5). (Hassan et al., 2017; Jurik et al., 2014; Leach et al., 2012; Miller et al., 2015; Miller et al., 2016)

Another finding of this review that questions the construct validity of the MGS is the change in grimace scores in response to techniques that would not be expected to elicit pain. Out of the eleven studies that examined the MGS over the 24 hour period after an intervention, that were expected to elicit none or momentary pain, six found grimace score elevations. A further examination of these studies shows that three of the studies were examining the effect of anaesthesia/analgesia on grimace scores. (Hohlbaum et al., 2017, 2018; Miller et al., 2015) In general both inhalational, (Hohlbaum et al., 2017; Miller et al., 2015) and injectable anaesthetics (Hohlbaum et al. 2018) increase scores, in the absence of a presumed painful event. However, whilst analgesia might similarly be expected to elevate scores, in two studies both tramadol, (Jirkof et al., 2020) and buprenorphine, (Miller et al., 2015) were not determined to have any impact. The impact of the anaesthetics is short-lived, having resolved by 24 hours. It is postulated that this could be related to a ‘hangover’ or sedative effect remaining after the procedure, which could be envisaged to lead to eye closure as in sleep. However, perhaps a lingering muscle relaxant effect could similarly affect the other action units. The evidence on an elevation with inhalational anaesthetics is also not clear with a strain effect being identified in the Miller et al. 2015 study. The study by Sorge et al., 2014 is mechanistically different to the other studies within this group since exposure to a painful insult was applied, with differences in grimace response shown to result from a form of male pheromone –induced stress analgesia. The remaining two studies found increased grimace scores as a result of blood sampling (Meyer et al., 2020) and handling and identification. (Roughan and Sevenoaks, 2019) In the former, (Meyer et al., 2020) facial vein and retrobulbar bleeding increased scores in the immediate post-procedural period. This study also provides further evidence for the effects of isoflurane on the MGS with increased scores seen in anaesthetised compared to sham handled groups. In the study of Roughan and Sevenoaks, 2019 increased scores were seen as a result of tail handling and ear tagging. There are several points of relevance here in relation to MGS construct validity. Firstly, the blood sampling interventions applied are likely to produce momentary pain as opposed to no pain, (Whittaker and Barker, 2020) so evidence of a change actually supports construct validity. Secondly, tail handling has been suggested to be aversive rather than painful (Hurst and West, 2010) so an effect does call into question the specificity of the scale for pain (although noting that a previous study found no effect of handling (Miller and Leach, 2016). Thirdly, whilst blood sampling only caused immediate post-procedural changes in MGS (later time points were not examined), differences between groups for handling and identification often persisted for 24 hours, when it might be assumed that any pain would have resolved, although as demonstrated in this study, at a time point when inflammation remained. (Roughan and Sevenoaks, 2019) Interestingly, there was also non-convergence of findings relating to inflammatory response and MGS with tunnel handled mice demonstrating a greater response than tail-handled animals.

4.3 Reliability

Pain scales should be reliable, that is produce similar results whenever they are used. (Good et al.,2001) This requires that between animal, intra-animal and temporal variations are minimised unless they result from differences in pain experience. Reliability impacts on validity since if errors in measurement are significant, the scale no longer performs well at assessing pain. (Mogil et al., 2020) The included papers assessed a number of measures of reliability including within observer variability (intra-observer), between observer (inter- observer) and across site variability.

Whilst a fair proportion of the studies investigating grimace scales utilised more than one observer for scoring, only 50% of these analysed and reported on between observer metrics. This represents a significant loss of data on the reliability of these scales. This raises the question of whether these data were not analysed, or not reported, perhaps because of low agreement. If there was more ability, and uptake of protocol registration in pre-clinical studies this question may not have arisen. Moreover, in encouraging the use of these scales for practical welfare assessment as clinical tools, this question is important; few institutions will be able to rely on the same, single observer to perform all scoring.

Based on the limited evidence available, inter-rater agreement generally ranges from good to excellent. However, a recent study (Hohlbaum et al., 2020) does suggest that this may change across time, with differences potentially being obscured by assimilation of all data. This is a factor that should be considered in future studies. Related to this, there may also be differences in scores for similar treatments when taken across laboratories. (Jirkof et al., 2020) It is not clear whether this relates to inter-observer differences, or differences in housing/test conditions but does call into question the external validity of MGS results. (Jirkof et al., 2020) However, importantly this study did find that whilst values across research centres were numerically different, direction of effect was similar so general validity was maintained. Fewer studies have reported on intra-rater variability. Although the study by Mittal et al 2016 which used a large number of coders (6) did report significant within coder variability in 3 individuals. All of these findings raise the question of whether training and experience in the use of the scales impacts on reliability. Few studies have specifically examined this, and detailed information on training was rarely provided in the included studies. Evidence for a training effect is currently conflicting with one study (Hohlbaum et al., 2020) suggesting greater consistency if scorers were experienced, whilst another study (Roughan and Sevenoaks, 2019) finding good correlation between novice and expert scorers. The impact of, and type/frequency of training, needed to produce reliable grimace scores is an area that needs further research especially if the technique is going to gain more widespread acceptance as a pain assessment tool. This is also a particular consideration for real time scoring which needs to be performed quickly and does not offer the opportunity for re-review of collected images.

4.4 The Impact of Biological Variation and the External Environment on the MGS

The synthesis demonstrates that there are a number of features of biology and the external environment that influence grimace scores. These include the influences of strain, sex, the circadian cycle and observers. These differences should be considered in future investigations of grimace scores, especially in the development of intervention scores.

A limited number of studies have directly contrasted more than one strain (Cho et al., 2019; Miller et al., 2015; Miller and Leach, 2015a). It is difficult to draw any conclusions on impact of strain on grimace scores since there appears, at least on the basis of one study, (Miller and Leach, 2015a) to be interactions between sex and strain on grimace scores. In general, with some exceptions due to sex differences, it appears that the strain order from propensity to score low to high is C57BL/6, CD1, C3H/H, BALB/c. However, it is worth noting that much of this information on strain differences comes from one study, (Miller and Leach, 2015a) where a painful insult was not applied. This may be of relevance, particulary on consideration of the interaction between sex and strain, since it is well established that there are differences in pain thresholds between male and female rodents, with females having a lower pain threshold in response to a variety of nociceptive inputs. (Hurley and Adams, 2008) It is interesting to note that this strain ranking shows no obvious trend based on coat colouration, implying that inability to score individual action units due to this may have minimal impact on scores obtained.

Evidence as to the presence, or nature of any differences in scores as a result of sex is far from settled. The majority of studies that compared sex differences within the same strain found no differences in scores. In regard to the minority of studies that did find sex differences, there is a fairly even split between those that found scores were lower in females and vice versa. This is perhaps surprising given the finding using traditional pain assays that female rodents have a lower pain threshold in the face of hot thermal, (Sternberg et al., 2004) chemical, (Gaumond et al., 2002) inflammatory, (Dina et al., 2001) and mechanical nociceptive insults. (Barrett et al., 2002) However, varied findings in relation to sex differences are not uncommon in these other models, and probably arise due to differences in study design as well as genotype.(Mogil et al., 2000) The absence of a sex effect in the majority of the studies that evaluated both sexes may also speak to a lack of sensitivity of the scoring, whereby differences are present, but cannot be discriminated. Another more general finding arising from the assimilation is that in spite of increased promotion of the use of both sexes in preclinical research due to concerns about translation, (Clayton and Collins, 2014; Whittaker and Hickman, 2020) the majority of studies used one sex (predominantly males). Even when two sexes were used in the included studies, an opportunity was often missed by failing to make direct comparisons between them.

Circadian rhythms commonly apply to biological and physiological processes in animals. (Konecka and Sroczynska, 1998) Mice as nocturnal animals are active mainly during the dark phase. (Ripperger et al., 2011) The strength of this circadian clock is such that even in constant darkness this pattern of activity will persist despite the absence of external cues. (Ripperger et al., 2011) There is also evidence of a circadian rhythm in pain sensitivity across a range of animal species, (Frederickson et al., 1977; Hamra et al., 1993; Konecka and Sroczynska, 1998) potentially brought about by a rhythm associated with opioid peptide production. (Naber et al., 1981; Oliverio et al., 1982) Considering that general levels of activity are likely to confound behavioural measurements particularly (although not exclusively), it follows that experimental protocols would control for this, and report on time of testing. This also raises the question of whether performing behavioural tests in the light phase is a major methodological error. (Yang et al., 2008) Given this, it is surprising that many of the included studies failed to report on timing of MGS measurements; this being an item in the updated ARRIVE guidelines recommended set. (Percie du Sert et al., 2020) Given the lack of dedicated study and reporting deficiencies, there is limited evidence to support or refute an effect of circadian rhythm on the MGS. However, two studies hint at potential differences (Matsumiya et al., 2012; Rea et al., 2018) with a suggestion of higher scores or pain in the dark phase. Nevertheless, Rea et al. 2018 did discuss that light transition appeared to cause decreases in orbital tightening and nose bulge, and it is not clear whether this effect would have persisted once acclimatised to the light.

Observer effects on the scale has been little investigated. This is unsurprising given that the majority of studies using the MGS have utilised retrospective analysis for scoring. However, as previously discussed observer effects may be relevant when photography is used, and are of clear importance in real time scoring since it is well established from animal behaviour research that a human observer may influence animal behaviour. (Martin and Bateson, 2007) There is some suggestion from other species of minimal impacts on grimace scores by human observers, see eg. Leung et al., 2016. However, this needs dedicated investigation in the context of mice. Furthermore, the nature of the observer may be important in determining their impact on scores. For example, Sorge et al., 2014 demonstrated that the presence of human males led to a stress-induced analgesia and reduced grimace scores, and familiarity with the observer may also be a factor in response. (Mogil et al., 2020)

4.5 Conclusions and Recommendations for Future Research

This review has assimilated all primary literature to date on the MGS. It is concluded that the MGS has utility across a range of animal models, and expected pain types. There do however appear to be some differences arising as a result of biological variation such as sex or strain of mouse. These variables need consideration in study design or analysis to account for them appropriately. There is also some limited evidence that the MGS may not be wholly specific to pain. However, this evidence mainly comes from studies into husbandry or drug interventions, the latter generally only having a short-term effect, which can likely be explained by the pharmacological effects. It would be interesting to delve further into any potentially non-pain related grimace effects in animal models where other symptoms might be assumed to co-occur with pain, for example sickness behaviour. This could potentially be achieved by using analgesics to eliminate the pain response, although of course the risk of drug confounding would need to be considered.

Further research is needed on the use of the MGS as a real time method, and how this can be done to maintain validity of the method, whilst being practically feasible. Related to this is the question of how reliable scoring between observers is, and what type of training (if any) is needed to maximise between observer agreement. Finally, whilst there is suggestion from studies in this synthesis, (Sorge et al., 2014) and others, (Langford et al., 2006) that there is a social modulation of pain by conspecifics and the presence of other species, there has been little investigation of this fascinating area in the context of grimace responses.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. AW was supported by a Peter Doherty Biomedical Fellowship (APP1140072).

Competing Interests

None of the authors have any competing interests that may influence the content of this review or the conclusions made.

Appendix 1

Search Strategy

Search conducted on 4th May 2020

Database	Search Strategy
Medline	(Mouse [tiab]) OR (Mice [tiab])) OR (Murine [tiab])) OR ((Murin*) [tiab])) OR (Mus [tiab])) OR (Musculus [tiab])) OR (Transgenic Animal [tiab])) OR (Mice [mh])) AND (Grimace Scale)) OR (Grimace Score[tiab])) OR (Facial grimace[tiab]))
Scopus	TITLE-ABS-KEY("Mouse" OR "Mice" OR "Murine" OR "Murin* " OR "Mus " OR "Musculus" OR "Transgenic Animal")AND TITLE-ABS-KEY("Grimace Scale" OR "Grimace Score" OR "Facial grimace")
Web of Science	TS=(Mouse OR Mice OR Murine OR Murin*OR Mus OR Musculus OR Transgenic Animal OR Mice) AND TS= (Grimace Scale OR Grimace Score OR Facial grimace)

Appendix 2

Data extraction template (see excel supporting file)

References

Akintola, T., Raver, C., Studlack, P., Uddin, O., Masri, R., Keller, A., 2017. The grimace scale reliably assesses chronic pain in a rodent model of trigeminal neuropathic pain. Neurobiol Pain 2, 13-17.

Barrett, A.C., Smith, E.S., Picker, M.J., 2002. Sex-related differences in mechanical nociception and antinociception produced by μ-and κ-opioid receptor agonists in rats. Eur. J. Pharmacol. 452, 163-173.

Blackburn-Munro, G., 2004. Pain-like behaviours in animals–how human are they? Trends Pharmacol. Sci. 25, 299-305.

Boissy, A., Manteuffel, G., Jensen, M.B., Moe, R.O., Spruijt, B., Keeling, L.J., Winckler, C., Forkman, B., Dimitrov, I., Langbein, J., Bakken, M., Veissier, I., Aubert, A., 2007. Assessment of positive emotions in animals to improve their welfare. Phys. Behav. 92, 375-397.

Bu, X., Liu, Y., Lu, Q., Jin, Z., 2015. 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.

Burgos-Vega, C.C., Quigley, L.D., Trevisan Dos Santos, G., Yan, F., Asiedu, M., Jacobs, B., Motina, M., Safdar, N., Yousuf, H., Avona, A., Price, T.J., Dussor, G., 2019. Non-invasive dural stimulation in mice: A novel preclinical model of migraine. Cephalalgia 39, 123-134.

Carbone, L., Austin, J., 2016. Pain and laboratory animals: publication practices for better data reproducibility and better animal welfare. PloS one 11, e0155001.

Chartier, L.C., Hebart, M.L., Howarth, G.S., Whittaker, A.L., Mashtoub, S., 2020. Affective state determination in a mouse model of colitis-associated colorectal cancer. PloS one 15, e0228413.

Cho, C., Michailidis, V., Lecker, I., Collymore, C., Hanwell, D., Loka, M., Danesh, M., Pham, C., Urban, P., Bonin, R.P., Martin, L.J., 2019. Evaluating analgesic efficacy and administration route following craniotomy in mice using the grimace scale. Sci. Rep. 9, 359.

Clayton, J.A., Collins, F.S., 2014. Policy: NIH to balance sex in cell and animal studies. Nature News 509, 282.

Colquhoun, H.L., Levac, D., O'Brien, K.K., Straus, S., Tricco, A.C., Perrier, L., Kastner, M., Moher, D., 2014. Scoping reviews: time for clarity in definition, methods, and reporting. J. Clin. Epidemiol. 67, 1291-1294.

European Commission, 2019. Report on the statistics on the use of animals for scientific purposes in the Member States of the European Union in 2015-2017. Available online: https://op.europa.eu/en/publication-detail/-/publication/04a890d4-47ff-11ea-b81b-01aa75ed71a1. Accessed on 5th December 2020.

Cunningham, C., Campion, S., Teeling, J., Felton, L., Perry, V.H., 2007. The sickness behaviour and CNS inflammatory mediator profile induced by systemic challenge of mice with synthetic double-stranded RNA (poly I:C). Brain. Behav. Immun. 21, 490-502.

de Almeida, A.S., Rigo, F.K., De Prá, S.D., Milioli, A.M., Dalenogare, D.P., Pereira, G.C., Ritter, C.D.S., Peres, D.S., Antoniazzi, C.T.D., Stein, C., Moresco, R.N., Oliveira, S.M., Trevisan, G., 2019. Characterization of Cancer-Induced Nociception in a Murine Model of Breast Carcinoma. Cell. Mol. Neurobiol. 39, 605-617.

de Almeida, A.S., Rigo, F.K., De Prá, S.D., Milioli, A.M., Pereira, G.C., Lückemeyer, D.D., Antoniazzi, C.T., Kudsi, S.Q., Araújo, D., Oliveira, S.M., Ferreira, J., Trevisan, G., 2020. Role of transient receptor potential ankyrin 1 (TRPA1) on nociception caused by a murine model of breast carcinoma. Pharmacol. Res. 152, 104576.

Descovich, K., Wathan, J., Leach, M.C., Buchanan-Smith, H.M., Flecknell, P., Farningham, D., Vick, S.-J., 2017. Facial expression: an under-utilised tool for the assessment of welfare in mammals. ALTEX 34(3):409-429

Deuis, J.R., Dvorakova, L.S., Vetter, I., 2017. Methods Used to Evaluate Pain Behaviors in Rodents. Front. Mol. Neurosci. 10, 284.

Dina, O.A., Aley, K., Isenberg, W., Messing, R.O., Levine, J.D., 2001. Sex hormones regulate the contribution of PKCε and PKA signalling in inflammatory pain in the rat. Eur. J. Neurosci. 13, 2227-2233.

Duffy, S.S., Perera, C.J., Makker, P.G., Lees, J.G., Carrive, P., Moalem-Taylor, G., 2016. Peripheral and Central Neuroinflammatory Changes and Pain Behaviors in an Animal Model of Multiple Sclerosis. Front. Immunol. 7, 369.

Dwivedi, D.J., Grin, P.M., Khan, M., Prat, A., Zhou, J., Fox-Robichaud, A.E., Seidah, N.G., Liaw, P.C., 2016. Differential expression of PCSK9 modulates infection, inflammation, and coagulation in a murine model of sepsis. Shock 46, 672-680.

Ekman, P., 1992. Are there basic emotions? Psychol. Rev. 99, 550-553.

Faller, K.M., McAndrew, D.J., Schneider, J.E., Lygate, C.A., 2015. Refinement of analgesia following thoracotomy and experimental myocardial infarction using the Mouse Grimace Scale. Exp. Physiol. 100, 164-172.

Finlayson, K., Lampe, J.F., Hintze, S., Würbel, H., Melotti, L., 2016. Facial indicators of positive emotions in rats. PloS one 11.

Frederickson, R.C., Burgis, V., Edwards, J.D., 1977. Hyperalgesia induced by naloxone follows diurnal rhythm in responsivity to painful stimuli. Science 198, 756-758.

Gallo, M.S., Karas, A.Z., Pritchett-Corning, K., Garner Guy Mulder, J.P., Gaskill, B.N., 2020. Tell-tale TINT: Does the Time to Incorporate into Nest Test Evaluate Postsurgical Pain or Welfare in Mice? J. Am. Assoc. Lab. Anim. Sci. 59, 37-45.

Gaskill, B.N., Pritchett-Corning, K.R., 2016. Nest building as an indicator of illness in laboratory mice. Appl. Anim. Behav. Sci. 180, 140-146.

Gaumond, I., Arsenault, P., Marchand, S., 2002. The role of sex hormones on formalin-induced nociceptive responses. Brain Res. 958, 139-145.

González-Cano, R., Montilla-García, Á., Ruiz-Cantero, M., Bravo-Caparrós, I., Tejada, M., Nieto, F., Cobos, E., 2020. The search for translational pain outcomes to refine analgesic development: Where did we come from and where are we going? Neurosci. Biobehav. Rev. 113.

Good, M., Stiller, C., Zauszniewski, J.A., Anderson, G.C., Stanton-Hicks, M., Grass, J.A.,2001. Sensation and Distress of Pain Scales: Reliability, Validity, and Sensitivity. J Nurs. Meas., 219-238.

Guo, W., Zou, S., Mohammad, Z., Wang, S., Yang, J., Li, H., Dubner, R., Wei, F., Chung, M.K., Ro, J.Y., Ren, K., 2019. Voluntary biting behavior as a functional measure of orofacial pain in mice. Physiol. Behav. 204, 129-139.

Hamra, J.G., Kamerling, S., Wolfsheimer, K., Bagwell, C., 1993. Diurnal variation in plasma ir-beta-endorphin levels and experimental pain thresholds in the horse. Life Sci. 53, 121-129.

Hassan, A.M., Jain, P., Mayerhofer, R., Fröhlich, E.E., Farzi, A., Reichmann, F., Herzog, H., Holzer, P., 2017. Visceral hyperalgesia caused by peptide YY deletion and Y2 receptor antagonism. Sci. Rep. 7, 40968.

Hassler, S.N., Ahmad, F.B., Burgos-Vega, C.C., Boitano, S., Vagner, J., Price, T.J., Dussor, G., 2019. Protease activated receptor 2 (PAR2) activation causes migraine-like pain behaviors in mice. Cephalalgia 39, 111-122.

Hau, J., Schapiro, S., 2010. Handbook of Laboratory Animal Science, Volume I, in: Hau, J.E., Schapiro, S. (Ed.). (Ed.). CRC Press, Boca Raton.

Herrera, C., Bolton, F., Arias, A.S., Harrison, R.A., Gutiérrez, J.M., 2018. Analgesic effect of morphine and tramadol in standard toxicity assays in mice injected with venom of the snake Bothrops asper. Toxicon. 154, 35-41.

Hohlbaum, K., Bert, B., Dietze, S., Palme, R., Fink, H., Thöne-Reineke, C., 2017. Severity classification of repeated isoflurane anesthesia in C57BL/6JRj mice-Assessing the degree of distress. PLoS One 12, e0179588.

Hohlbaum, K., Bert, B., Dietze, S., Palme, R., Fink, H., Thöne-Reineke, C., 2018. Impact of repeated anesthesia with ketamine and xylazine on the well-being of C57BL/6JRj mice. PLoS One 13, e0203559.

Hohlbaum, K., Corte, G.M., Humpenöder, M., Merle, R., Thöne-Reineke, C., 2020. Reliability of the Mouse Grimace Scale in C57BL/6JRj Mice. Animals 10.

Homberg, J.R., Wöhr, M., Alenina, N., 2017. Comeback of the Rat in Biomedical Research. ACS Chem. Neurosci. 8, 900-903.

Hsi, Z.Y., Stewart, L.A., Lloyd, K.C.K., Grimsrud, K.N., 2020. Hypoglycemia after Bariatric Surgery in Mice and Optimal Dosage and Efficacy of Glucose Supplementation. Comp. Med. 70, 111-118.

Hurley, R.W., Adams, M.C.B., 2008. Sex, gender, and pain: an overview of a complex field. Anesth. Analg. 107, 309-317.

Hurst, J.L., West, R.S., 2010. Taming anxiety in laboratory mice. Nat. Meth. 7, 825-826.

Jirkof, P., 2014. Burrowing and nest building behavior as indicators of well-being in mice. J. Neurosci. Meth. 234, 139-146.

Jirkof, P., Abdelrahman, A., Bleich, A., Durst, M., Keubler, L., Potschka, H., Struve, B., Talbot, S.R., Vollmar, B., Zechner, D., Häger, C., 2020. A safe bet? Inter-laboratory variability in behaviour-based severity assessment. Lab. Anim. 54, 73-82.

Jirkof, P., Leucht, K., Cesarovic, N., Caj, M., Nicholls, F., Rogler, G., Arras, M., Hausmann, M., 2013. Burrowing is a sensitive behavioural assay for monitoring general wellbeing during dextran sulfate sodium colitis in laboratory mice. Lab. Anim. 47, 274-283.

Jurik, A., Ressle, A., Schmid, R.M., Wotjak, C.T., Thoeringer, C.K., 2014. Supraspinal TRPV1 modulates the emotional expression of abdominal pain. Pain 155, 2153-2160.

Kim, J.Y., Tillu, D.V., Quinn, T.L., Mejia, G.L., Shy, A., Asiedu, M.N., Murad, E., Schumann, A.P., Totsch, S.K., Sorge, R.E., Mantyh, P.W., Dussor, G., Price, T.J., 2015. Spinal dopaminergic projections control the transition to pathological pain plasticity via a D1/D5-mediated mechanism. J. Neurosci. 35, 6307-6317.

Konecka, A.M., Sroczynska, I., 1998. Circadian rhythm of pain in male mice. Gen. Pharmacol. 31, 809-810.

Langford, D.J., Bailey, A.L., Chanda, M.L., Clarke, S.E., Drummond, T.E., Echols, S., Glick, S., Ingrao, J., Klassen-Ross, T., Lacroix-Fralish, M.L., Matsumiya, L., Sorge, R.E., Sotocinal, S.G., Tabaka, J.M., Wong, D., van den Maagdenberg, A.M., Ferrari, M.D., Craig, K.D., Mogil, J.S., 2010. Coding of facial expressions of pain in the laboratory mouse. Nat. Meth. 7, 447-449.

Langford, D.J., Crager, S.E., Shehzad, Z., Smith, S.B., Sotocinal, S.G., Levenstadt, J.S., Chanda, M.L., Levitin, D.J., Mogil, J.S., 2006. Social modulation of pain as evidence for empathy in mice. Science 312, 1967-1970.

Leach, M.C., Klaus, K., Miller, A.L., Scotto di Perrotolo, M., Sotocinal, S.G., Flecknell, P.A., 2012. The assessment of post-vasectomy pain in mice using behaviour and the Mouse Grimace Scale. PLoS One 7, e35656.

Leenaars, C.H.C., Kouwenaar, C., Stafleu, F.R., Bleich, A., Ritskes-Hoitinga, M., De Vries, R.B.M., Meijboom, F.L.B., 2019. Animal to human translation: a systematic scoping review of reported concordance rates. J. Transl. Med. 17, 223.

LeResche, L., 1982. Facial expression in pain: A study of candid photographs. J. Nonverbal Behav. 7, 46-56.

Leung, V., Zhang, E., Pang, D.S., 2016. Real-time application of the Rat Grimace Scale as a welfare refinement in laboratory rats. Sci. Rep. 6, 31667.

Leung, V.S.Y., Benoit-Biancamano, M.-O., Pang, D.S.J., 2019. 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.

Mai, S.H.C., Sharma, N., Kwong, A.C., Dwivedi, D.J., Khan, M., Grin, P.M., Fox-Robichaud, A.E., Liaw, P.C., 2018. Body temperature and mouse scoring systems as surrogate markers of death in cecal ligation and puncture sepsis. Intensive Care Med. Exp. 6, 20.

Martin, P.R., & Bateson, P. P. G. (2007). Measuring behaviour: An introductory guide. Cambridge, United Kingdom: Cambridge University Press.

Matsumiya, L.C., Sorge, R.E., Sotocinal, S.G., Tabaka, J.M., Wieskopf, J.S., Zaloum, A., King, O.D., Mogil, J.S., 2012. Using the Mouse Grimace Scale to reevaluate the efficacy of postoperative analgesics in laboratory mice. J. Am. Assoc. Lab. Anim. Sci. 51, 42-49.

McLennan, K.M., Miller, A.L., Dalla Costa, E., Stucke, D., Corke, M.J., Broom, D.M., Leach, M.C., 2019. Conceptual and methodological issues relating to pain assessment in mammals: The development and utilisation of pain facial expression scales. Appl. Anim. Behav. Sci. 217, 1-15.

Melnikova, I., 2010. Pain market. Nat. Rev. Drug Discov. 9, 589-590.

Meyer, N., Kröger, M., Thümmler, J., Tietze, L., Palme, R., Touma, C., 2020. Impact of three commonly used blood sampling techniques on the welfare of laboratory mice: Taking the animal's perspective. PLoS One 15, e0238895.

Miller, A., Kitson, G., Skalkoyannis, B., Leach, M., 2015. 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.

Miller, A.L., Kitson, G.L., Skalkoyannis, B., Flecknell, P.A., Leach, M.C., 2016. Using the mouse grimace scale and behaviour to assess pain in CBA mice following vasectomy. Appl. Anim. Behav. Sci. 181, 160-165.

Miller, A.L., Leach, M.C., 2015a. The Mouse Grimace Scale: A Clinically Useful Tool? PLoS One 10, e0136000.

Miller, A.L., Leach, M.C., 2015b. 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.

Miller, A.L., Leach, M.C., 2016. The effect of handling method on the mouse grimace scale in two strains of laboratory mice. Lab. Anim. 50, 305-307.

Mitchell, C.J., Howarth, G.S., Chartier, L.C., Trinder, D., Lawrance, I.C., Huang, L.S., Mashtoub, S., 2020. Orally administered emu oil attenuates disease in a mouse model of Crohn's-like colitis. Exp. Biol. Med., 1535370220951105.

Mittal, A., Gupta, M., Lamarre, Y., Jahagirdar, B., Gupta, K., 2016. Quantification of pain in sickle mice using facial expressions and body measurements. Blood Cells Mol. Dis. 57, 58-66.

Mogil, J.S., 2009. Animal models of pain: progress and challenges. Nat. Rev. Neurosci. 10, 283-294.

Mogil, J.S., Chesler, E., Wilson, S., Juraska, J.M., Sternberg, W., 2000. Sex differences in thermal nociception and morphine antinociception in rodents depend on genotype. Neurosci. Biobehav. Rev. 24, 375-389.

Mogil, J.S., Pang, D.S., Dutra, G.G.S., Chambers, C.T., 2020. The development and use of facial grimace scales for pain measurement in animals. Neurosci. Biobehav. Rev.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., The, P.G., 2009. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLOS Med. 6, e1000097.

Mota-Rojas, D., Olmos-Hernández, A., Verduzco-Mendoza, A., Hernández, E., Martínez-Burnes, J., Whittaker, A.L., 2020. The Utility of Grimace Scales for Practical Pain Assessment in Laboratory Animals. Animals 10, 1838.

Naber, D., Cohen, R.M., Pickar, D., Kalin, N.H., Davis, G., Pert, C.B., Bunney Jr, W.E., 1981. Episodic secretion of opioid activity in human plasma and monkey CSF: evidence for a diurnal rhythm. Life Sci. 28, 931-935.

Nagakura, Y., 2017. The need for fundamental reforms in the pain research field to develop innovative drugs. Expert Opin. Drug Discov. 12, 39-46.

Nagakura, Y., Miwa, M., Yoshida, M., Miura, R., Tanei, S., Tsuji, M., Takeda, H., 2019. Spontaneous pain-associated facial expression and efficacy of clinically used drugs in the reserpine-induced rat model of fibromyalgia. Eur. J. Pharmacol. 864.

Oliverio, A., Castellano, C., Puglisi-Allegra, S., 1982. Opiate analgesia: evidence for circadian rhythms in mice. Brain Res. 249, 265-270.

Panksepp, J., 2005. Affective consciousness: Core emotional feelings in animals and humans. Conscious Cogn. 14, 30-80.

Percie du Sert, N., Hurst, V., Ahluwalia, A., Alam, S., Avey, M.T., Baker, M., Browne, W.J., Clark, A., Cuthill, I.C., Dirnagl, U., Emerson, M., Garner, P., Holgate, S.T., Howells, D.W., Karp, N.A., Lazic, S.E., Lidster, K., MacCallum, C.J., Macleod, M., Pearl, E.J., Petersen, O.H., Rawle, F., Reynolds, P., Rooney, K., Sena, E.S., Silberberg, S.D., Steckler, T., Würbel, H., 2020. The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. PLOS Biol. 18, e3000410.

Peters, M.D.J., Godfrey, C.M., Khalil, H., McInerney, P., Parker, D., Soares, C.B., 2015. Guidance for conducting systematic scoping reviews. JBI Evidence Implementation 13.

Peterson, N.C., Nunamaker, E.A., Turner, P.V., 2017. To treat or not to treat: the effects of pain on experimental parameters. Comp. Med. 67, 469-482.

Rea, B.J., Wattiez, A.S., Waite, J.S., Castonguay, W.C., Schmidt, C.M., Fairbanks, A.M., Robertson, B.R., Brown, C.J., Mason, B.N., Moldovan-Loomis, M.C., Garcia-Martinez, L.F., Poolman, P., Ledolter, J., Kardon, R.H., Sowers, L.P., Russo, A.F., 2018. Peripherally administered calcitonin gene-related peptide induces spontaneous pain in mice: implications for migraine. Pain 159, 2306-2317.

Ripperger, J.A., Jud, C., Albrecht, U., 2011. The daily rhythm of mice. FEBS Letters 585, 1384-1392.

Rosen, S.F., Ham, B., Drouin, S., Boachie, N., Chabot-Dore, A.-J., Austin, J.-S., Diatchenko, L., Mogil, J.S., 2017. T-Cell Mediation of Pregnancy Analgesia Affecting Chronic Pain in Mice. J. Neurosci. 37, 9819-9827.

Rossi, H.L., See, L.P., Foster, W., Pitake, S., Gibbs, J., Schmidt, B., Mitchell, C.H., Abdus-Saboor, I., 2020. Evoked and spontaneous pain assessment during tooth pulp injury. Sci. Rep. 10, 2759.

Roughan, J.V., Bertrand, H.G., Isles, H.M., 2016. Meloxicam prevents COX-2-mediated post-surgical inflammation but not pain following laparotomy in mice. Eur. J. Pain 20, 231-240.

Roughan, J.V., Sevenoaks, T., 2019. Welfare and Scientific Considerations of Tattooing and Ear Tagging for Mouse Identification. J. Am. Assoc. Lab. Anim. Sci. 58, 142-153.

Russell, W.M.S. and Burch, R.L., (1959). *The Principles of Humane Experimental Technique*, Methuen, London.

Serizawa, K., Tomizawa-Shinohara, H., Yasuno, H., Yogo, K., Matsumoto, Y., 2019. Anti-IL-6 Receptor Antibody Inhibits Spontaneous Pain at the Pre-onset of Experimental Autoimmune Encephalomyelitis in Mice. Front. Neurol. 10, 341.

Sorge, R.E., Martin, L.J., Isbester, K.A., Sotocinal, S.G., Rosen, S., Tuttle, A.H., Wieskopf, J.S., Acland, E.L., Dokova, A., Kadoura, B., Leger, P., Mapplebeck, J.C., McPhail, M., Delaney, A., Wigerblad, G., Schumann, A.P., Quinn, T., Frasnelli, J., Svensson, C.I., Sternberg, W.F., Mogil, J.S., 2014. Olfactory exposure to males, including men, causes stress and related analgesia in rodents. Nat. Methods 11, 629-632.

Sternberg, W.F., Smith, L., Scorr, L., 2004. Nociception and antinociception during the first week of life in mice: sex differences and test dependence. J. Pain 5, 420-426.

Tillu, D.V., Hassler, S.N., Burgos-Vega, C.C., Quinn, T.L., Sorge, R.E., Dussor, G., Boitano, S., Vagner, J., Price, T.J., 2015. Protease-activated receptor 2 activation is sufficient to induce the transition to a chronic pain state. Pain 156, 859-867.

Tuttle, A.H., Molinaro, M.J., Jethwa, J.F., Sotocinal, S.G., Prieto, J.C., Styner, M.A., Mogil, J.S., Zylka, M.J., 2018. A deep neural network to assess spontaneous pain from mouse facial expressions. Mol. Pain 14, 1744806918763658.

Wang, S., Brigoli, B., Lim, J., Karley, A., Chung, M.K., 2018. Roles of TRPV1 and TRPA1 in Spontaneous Pain from Inflamed Masseter Muscle. Neurosci. 384, 290-299.

Wang, S., Kim, M., Ali, Z., Ong, K., Pae, E.K., Chung, M.K., 2019. TRPV1 and TRPV1-Expressing Nociceptors Mediate Orofacial Pain Behaviors in a Mouse Model of Orthodontic Tooth Movement. Front. Physiol. 10, 1207.

Wang, S., Lim, J., Joseph, J., Wang, S., Wei, F., Ro, J.Y., Chung, M.K., 2017. Spontaneous and Bite-Evoked Muscle Pain Are Mediated by a Common Nociceptive Pathway With Differential Contribution by TRPV1. J Pain 18, 1333-1345.

Whiteside, G.T., Pomonis, J.D., Kennedy, J.D., 2013. An industry perspective on the role and utility of animal models of pain in drug discovery. Neurosci. Lett. 557, 65-72.

Whittaker, A.L., Barker, T.H., 2020. The Impact of Common Recovery Blood Sampling Methods, in Mice (Mus Musculus), on Well-Being and Sample Quality: A Systematic Review. Animals 10, 989.

Whittaker, A.L., Hickman, D.L., 2020. The Impact of Social and Behavioral Factors on Reproducibility in Terrestrial Vertebrate Models. ILAR J. 60, 252-269.

Whittaker, A.L., Howarth, G.S., 2014. Use of spontaneous behaviour measures to assess pain in laboratory rats and mice: How are we progressing? Appl. Anim. Behav. Sci. 151, 1-12.

Whittaker, A.L., Lymn, K.A., Nicholson, A., Howarth, G.S., 2015. The assessment of general well-being using spontaneous burrowing behaviour in a short-term model of chemotherapy-induced mucositis in the rat. Lab. Anim. 49, 30-39.

Whittaker, A.L., Marsh, L.E., 2019. The role of behavioural assessment in determining 'positive' affective states in animals. CAB Rev. 14, 1-13.

Wu, J., Zhao, Z., Zhu, X., Renn, C.L., Dorsey, S.G., Faden, A.I., 2016. Cell cycle inhibition limits development and maintenance of neuropathic pain following spinal cord injury. Pain 157, 488-503.

Yang, M., Weber, M., Crawley, J., 2008. Light phase testing of social behaviors: not a problem. Front. Neurosci. 2.

Zhu, Y., Wang, S., Long, H., Zhu, J., Jian, F., Ye, N., Lai, W., 2017. 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.

Figure Captions

Figure 1. The PRISMA (Moher et al., 2009) flow diagram for the systematic review detailing the database searches, the number of abstracts screened and the full texts retrieved.

Figure 2. Number of publications per year investigating the MGS in relation to pain

Figure 3. Pain classifications used to guide assignment of studies to categories. Adapted by permission from Springer Nature and Copyright Clearance Center: Springer Nature, Nature Reviews Drug Discovery, Pain Market, Melnikova, I, COPYRIGHT 2010. For specific category assignment for included studies refer Table 1.

Figure 4. Heat map contrasting interventions used by the type of pain expected to be elicited. Colouration/number in box represents number of studies. Whilst some studies could arguably have been included in multiple categories to simplify reporting one category has been assigned. ‡ Pain in this study, Mittal et al., 2016, was assigned as acute since induced by cold stress, although sickle cell pain can be neuropathic in origin. * Model of Hsi et al., 2020 did not relate to a neuropathy.

Figure 5. Representation of the different mouse strains and stocks used in the published papers. Note that a number of papers used more than one strain. The ICR and CD-1 nomenclature has been considered to represent the same stock. Other includes hybrid or recombinant strains.

Figure 6. Network map comparing MGS scores between strains. Each line represents a study effect. The direction of the arrow represents that the strain at the arrowhead responded with a lower MGS score. Red lines indicate a comparison between female mice, blue lines indicate comparison between male mice and black lines indicate comparisons where sex was not separated. A solid line indicates that a live score was used, a dashed line indicates that a retrospective score was used.

Figure 7. Collection methods used in included studies.

Figure 8: Facial Action Units (FAU's) utilised for scoring in the included studies. n represents study number. Specific action units were generally excluded as described here, although in one study two of the action units were combined

Figure 9. Heat map contrasting type of pain expected to arise from the interventions with the time points after the intervention investigated. Colouration gradation represents percentage of studies where grimace scores moved in the expected direction of effect, with increased shading indicating greater number of investigations, for example 100% of studies evaluating procedures likely to cause acute pain showed increased MGS scores within the 24 hours after the intervention. * Consider that no change in MGS score is expected.

Figure 10. Word cloud illustrating corroborating methods of affective state assessment used in the included studies. Size of the word illustrates their relative frequency of use.

Table 1: Included Studies Table

Reference	Study Design§	Strain/stock	Age/Weight	Sex	Type of Intervention	Intervention (model created or procedure investigated)	Pain Classification Assigned	Intervention effect on MGS score*
Akintola et al., 2017	RCT	C57BL/6	10-12 weeks	M	Animal Model	Chronic constriction injury model for pain	Neuropathic	↑
Bu et al., 2015	RCT	BALB/c	6-8 weeks	F	Animal Model	Chronic pelvic pain	Visceral	↑
Burgos-Vega et al., 2019	RCT	ICR	6-8 weeks	M,F	Animal Model	Migraine	Neuropathic	↑
Chartier et al., 2020	RCT	C57BL/ 6JArc	8 wks	F	Animal Model	Colitis-associated colorectal cancer	Visceral	Nil
Cho et al., 2019	RCT	CD-1, C57BL/6N	7-9 weeks	M,F	Animal Model	Craniotomy with different analgesics	Neuropathic	↑ (reduced by analgesics)
de Almeida et al., 2019	Pre-test, post-test	BALB/c	20-30g	F	Animal Model	Cancer-induced nociception	Mixed	↑
de Almeida et al., 2020	RCT	BALB/c		F	Animal Model	Cancer-induced nociception	Mixed	↑
Duffy et al., 2016	RCT	C57BL/6J	10-12 weeks	F	Animal Model	Experimental Autoimmune Encephalomyelitis (EAE)	Neuropathic	↑
Dwivedi et al., 2016	RCT	Transgenic (BL/6 background) PCSK9 KO mice and PCSK9 overexpression	10-12 wks	M,F	Animal Model	Caecal Ligation and Puncture model of sepsis	Visceral	↑
Faller et al., 2015	RCT	C57BL/6J, transgenic overexpressing creatine transporter in the heart (BL/6 background)	12-16 wks	F	Animal Model	Myocardial infarction created through thoracotomy	Visceral	↑ (reduced by analgesics)
Gallo et al., 2020	RCT (factorial)	Crl:CD1(ICR)	8-9 wks	M	Husbandry/Procedural	Carotid artery catheterisation	Acute	↑

Guo et al., 2019	RCT	C57BL/6	9 wks	M,F	Animal Model	Orofacial pain	Acute	
Hassan et al., 2017	RCT	C57BL/6N, PYY knockout	10 wks	M	Animal Model	Colonic nociception	Visceral	↑
Hassler et al., 2019	RCT	ICR, C57BL/6J, and PAR2 (BL/6 background)	20 - 30 g	M	Animal Model	Migraine	Neuropathic	↑
Herrera et al., 2018	RCT	CD-1	18-20g	nr	Animal Model	Bothops Asper venom	Visceral	↑
Hohlbaum et al., 2017	RCT	C57BL/6JRj	11-13 wks	M, F	Husbandry/Procedural	Isoflurane anaesthesia	None/momentary	↑ (female mice only)
Hohlbaum et al., 2018	RCT	C57BL/6JRj	11-13 wks	M, F	Husbandry/Procedural	Ketamine/xylazine anaesthesia	None/momentary	↑
Hohlbaum et al., 2020 [⁋]	Quasi-experimental	C57BL/6JRj	11-13 wks	M, F	Biological	Ketamine /xylazine anaesthesia	N/A	Study investigated inter-observer reliability in scoring
Hsi et al., 2020	RCT	C57BL/6N	7-9 wks	F	Animal Model	Animals with hypoglycaemia following Roux-en-Y Gastric Bypass surgery	None/momentary (outcome of interest is the hypoglycaemia)	Nil[‡]
Jirkof et al., 2020	RCT	C57BL/6J		F	Husbandry/Procedural	Tramadol treatment effect on MGS between laboratories	None/momentary	Nil
Jurik et al., 2014	RCT	TRPV1 knock-out (BL/6 background))	8-16 wks	M	Animal Model	Abdominal Constriction Test and Acute pancreatitis as models of pain. Effects of knockout versus wildtype genotype	Visceral	No effect of genotype on MGS scores
Kim et al., 2015	RCT	ICR		M,F	Animal Model	Hyperalgesic priming via IL-6 and Carrageenan injection	Acute	↑
Langford et al., 2010	RCT	CD-1 (ICR:Crl)	6-18 weeks	M,F	Animal Model	14 models of pain	Acute	↑
Leach et al., 2012	RCT	CD-1		M	Husbandry/Procedural	Vasectomy surgery	Acute	↑ (reduced by analgesics)
Mai et al., 2018	RCT	C57BL/6J	8-12 wks	M	Animal Model	Caecal Ligation and Puncture model of sepsis	Visceral	↑
Matsumiya et al., 2012	RCT	CD1 (ICR:Crl)	6-8 wks	M,F	Husbandry/Procedural	Ventral ovariectomy and response to analgesics	Acute	↑(reduced by analgesics)
Meyer et al., 2020	RCT	C57BL/6J	10-12 wks	M	Husbandry/Procedural	Common recovery blood sampling routes (facial vein,	None/momentary	↑- anaesthetic, facial vein bleeding, or

						retrobulbar, tail vein with anaesthetic and handling control)		retrobulbar compared to handling
Miller et al., 2015	Pre-test, post-test	CBA, DBA/2		M	Husbandry/Procedural	Isoflurane anaesthesia and buprenorphine analgesic	None/momentary	Nil (↑ by isoflurane in DBA/2 strain)
Miller and Leach, 2015a	RCT	C57BL/6, C3H/He, CD-1 BALB/c	8 wks	M,F	Biological	Impact of sex, strain, time of day or habituation	None/momentary	Nil-order ↑- males compared to females (but strain dependant and not always consistent) Strain effects present Time of day effects with sex and strain differences
Miller and Leach, 2015b	RCT	C57BL/6	8 wks	M	Husbandry/Procedural	Ear notching and analgesic effects	None/momentary	Nil
Miller et al., 2016	Pre-test, post-test	CBA	25.6-28.7g	M	Husbandry/Procedural	Vasectomy surgery	Acute	↑
Miller and Leach, 2016	RCT	CBA, DBA/2		M	Husbandry/Procedural	Handling method: tail versus tube	None/momentary	Nil
Mitchell et al., 2020	RCT	ArcCrl:CD	12 wks	F	Animal Model	TNBS-induced Crohn's-like colitis	Visceral	↑
Mittal et al., 2016	RCT	Transgenic HbSS-BERK (with relevant controls)		M,F	Animal Model	Sickle cell disease and effects of cold	Acute	↑- in females, cold also had impact
Rea et al., 2018	RCT	C57BL/6J, CD1	10-14 wks	M,F	Animal Model	Pain as result of migraine	Neuropathic	↑
Rosen et al., 2017	RCT	CD-1 (Crl:ICR) , Nude (Crl:CD1-Foxn1 nu), C57BL/6J, C57BL/6-Rag1 tm1Mom, mutant mice lacking expression of the	7- 12 wks	M,F	Animal Model	Pregnancy analgesia after inflammatory insult induced by administration of Complete Freund's Adjuvant (CFA)	Visceral (pregnancy state).	↑-in late-pregnant mice compared to nulliparous females

		Oprd1 (-opioid receptor) gene						
Rossi et al., 2020	RCT	Mixed CD1 and C57BL6/J background	17-21 wks	M,F	Animal Model	Tooth pulp injury	Acute	↑
Roughan et al., 2016	RCT	BALB/c	25-30 g	M	Husbandry/Procedural	Handling method: tail versus cupping at time of surgery	None/momentary	Nil (although surgery itself increased MGS)
Roughan and Sevenoaks, 2019	RCT	BALB/cAnNCrl	10-13 wks	M,F	Husbandry/Procedural	Ear tattooing and tagging, with tail handling method or tunnel	None/momentary	↑ tail versus tunnel ↑ males versus females ↑ ear tagging versus tattoo
Sorge et al., 2014	RCT	CD-1 (ICR:Crl), C57BL/6J	6-12 wks	M,F	Husbandry/Procedural	Effect of gender/gender-specific and other animal pheromones on response to nociceptive assays	None/momentary (intervention of interest is the pheromones)	↓ with male observer or male's T-shirt compared to no observer
Serizawa et al., 2019	RCT	C57BL/6J	7 wks	F	Animal Model	Experimental Autoimmune Encephalomyelitis (EAE)	Neuropathic	↑
Tillu et al., 2015	RCT	ICR, C57Bl/6	20-25g	M	Animal Model	Hyperalgesic priming	Acute	↑
Tuttle et al., 2018	RCT	CD-1 (ICR:Crl)	6-12 wks	M,F	Husbandry/Procedural	Ventral ovariectomy and response to analgesics, xymogen assay (validation of automated scoring)	Acute	↑ (reduced by analgesic)
Wang et al., 2017	RCT	C57BL/6, TRPV1 KO	8-12 wks	M	Animal Model	Masseter inflammation	Acute	↑
Wang et al., 2018	RCT	C57BL/6, TRPV1 KO (C57BL/6 background), TRPA1 KO (mixed B6; 129 background)	8-12 wks	M	Animal Model	Masseter inflammation	Acute	↑
Wang et al., 2019	RCT	C57BL/6	12 wks	M	Animal Model	Orthodontic tooth movement	Acute	↑

Wu et al., 2016	RCT	C57BL/6	8-19 wks	M	Animal Model	Spinal cord injury	Neuropathic	↑
Zhu et al., 2017	RCT	Balb/c	25-30g	M	Animal Model	Orthodontic tooth movement	Acute	↑

§ The terminology Randomised Control Trial has been used to indicate use of a comparator with a parallel arrangement of study groups, however randomisation was not necessarily performed at all or to a high standard in all studies. This was not specifically investigated as part of this review.

*General consistent direction of effect

ǂ no sham control so effect of model unknown

⁋ Study used photos obtained from Hohlbaum 2017 study. In reporting we have not considered this to represent an additional animal study for presentation of animal-focussed data