Université de Montréal

Advancing pain research and animal welfare: focusing on the Rat Grimace Scale and reporting standards

par Vivian Leung

Département de sciences cliniques
Faculté de médecine vétérinaire

Thèse présentée à la Faculté de médecine vétérinaire
en vue de l'obtention du grade de *Philosophiae Doctor* (Ph. D.)
en sciences vétérinaires

Octobre, 2018

Université de Montréal

Faculté des études supérieures et posdoctorales

Cette thèse intitulée:

Advancing pain research and animal welfare: focusing on the Rat Grimace Scale and reporting standards

Présentéé par

Vivian Leung

A été évaluée par un jury composé des personnes suivantes :

Dre. Elizabeth O'Toole, présidente-rapporteuse

Dr. Daniel Pang, directeur de recherche

Dr. Éric Troncy, membre du jury

Dre. Gilly Griffin, examinatrice externe

Résumé

Les comportements non-stimulés pour évaluer la douleur chez les animaux suscitent un intérêt grandissant. Un exemple est la « Rat Grimace Scale » (RGS), une échelle de douleur basée sur 4 unités d'action d'expressions faciales: resserrement orbital, aplatissement nez / joue, changements d'oreilles et vibrisses. Le potentiel de cette échelle et ses limites demeurent à déterminer.

La RGS standard est laborieuse à compléter (enregistrement de vidéos et extraction manuelle des images). Par conséquent, cette thèse a évalué si l'application en temps réel était possible. En comparant les résultats obtenus en temps réel à la méthode standard, il a été constaté que les résultats étaient similaires. Ainsi, la fiabilité de la RGS en temps réel élargit grandement son applicabilité en tant qu'outil clinique et de bien-être.

Toutefois, l'applicabilité de la RGS dans la douleur viscérale aiguë et chronique demeurent inexplorée. Par conséquent, cette thèse a évalué si la RGS pouvait évaluer la douleur à partir d'un modèle de colite aiguë et chronique de « dextran sulfate sodium » (DSS). Deux autres outils comportementaux (enfouissage et « Composite Behaviour Score » [CBS]) ont également été évalués. Ils ont été comparés au « Disease Activity Index » (DAI), un outil commun d'évaluation de la sévérité de la maladie. La RGS et l'enfouissage ont augmenté et diminué respectivement lorsque le DAI a augmenté. De futures études sont nécessaires pour valider le CBS. Cette étude démontre que le RGS peut évaluer la douleur viscérale et potentiellement plusieurs types de douleur.

La nécessité d'une formation avant la notation RGS a également été investiguée en évaluant la fiabilité de l'évaluateur après avoir reçu une formation ou aucune formation. Il a été constaté que la formation était bénéfique pour améliorer la fiabilité en plus de réduire la variabilité alors que la notation de plusieurs images seulement ne l'était pas. En outre, les évaluateurs obtenaient des résultats fiables après une période d'inactivité. Cette étude démontre donc le besoin de former les nouveaux évaluateurs.

Par ailleurs, cette thèse a également étudié si la publication des directives ARRIVE « Animal Research: Reporting of *In Vivo* Experiments » avait entraîné une amélioration des normes de déclaration. Cette étude a montré que les normes de déclaration ne s'étaient pas améliorées de manière significative, mais aussi que les articles publiés dans des revues qui soutiennent les directives ARRIVE n'ont pas de meilleurs standards. Par conséquent, cette étude souligne la nécessité d'imposer les directives ARRIVE pour assurer une amélioration significative.

Dans l'ensemble, cette thèse a démontré l'utilité de la RGS en temps réel comme outil d'évaluation clinique de la douleur viscérale chronique ainsi qu'en recherche. Elle souligne également le besoin de former les évaluateurs avant la notation RGS. Enfin, il a été démontré que les normes de déclaration restent faibles et que les directives ARRIVE doivent être imposées. Il est à espérer que ces études encourageront la progression de la recherche sur la douleur par l'amélioration des standards de déclaration ainsi que par l'utilisation de la RGS et autres comportements spontanés pour évaluer la douleur.

Mots-clés : échelles de grimace, expression faciale, douleur, comportements animaux, développement d'outils, état affectif, rats, modèles animaux, directives ARRIVE, normes de déclaration

Abstract

There is growing interest in the use of non-evoked spontaneous behaviours to assess pain in animals. A tool that measures such behaviours is the Rat Grimace Scale (RGS), a validated facial expression pain scale consisting of four "action units": orbital tightening, nose/cheek flattening, ear changes and whisker changes.

The strengths and limitations of the RGS are not fully explored. One limitation of the RGS (using the standard scoring method) is its time- and labour-intensive nature (video recording and manual image extraction are required). A primary goal of my research was to evaluate the feasibility of real-time RGS scoring. To accomplish this, the standard and real-time assessment methods were compared. It was found that both scoring methods were comparable and demonstrated the utility of real-time RGS scoring. This provides evidence that the RGS may be utilised not only as a research tool, but a useful clinical and welfare tool as well.

A further goal was to explore the use of the RGS in a visceral and chronic pain model. The RGS was hence tested in a dextran sulfate sodium (DSS) colitis model. Two other behavioural tools (burrowing and the composite behaviour score [CBS]) were also evaluated. These behavioural tools were compared to the Disease Activity Index (DAI), a common tool assessing disease severity in colitis models. The RGS and DAI scores increased and decreased concurrently. This study demonstrates that the RGS can be applied to assess chronic visceral pain and may be used to assess the mechanisms of different pain types.

The need for training prior to RGS scoring was explored by assessing the rater reliability after receiving training or no training (scoring of multiple images only). This study demonstrates that training is beneficial; training improved scoring reliability and reduced variability. This was not observed in raters who received no training. Additionally, trained raters could still score reliably four years later. Therefore, this study demonstrates the need for new raters to be trained in RGS use to improve reliability.

Lastly, this thesis explores whether the publication of the ARRIVE (Animal Research: Reporting of *In Vivo* Experiments) guidelines improved the reporting standards of animal studies. This study found that reporting standards had not improved meaningfully, and the standard of reporting was no better in papers published in journals that support the ARRIVE guidelines. Therefore, this highlights the need for the enforcement or refinement of ARRIVE guidelines to ensure meaningful improvement of reporting standards.

Overall, this thesis demonstrates the utility of the RGS as a practical pain assessment tool, with real-time application and the ability to assess chronic visceral pain. It highlights the need for raters to be trained prior to RGS scoring. Lastly, it demonstrates that the implementation of reporting standards in line with the ARRIVE guidelines are low, and enforcement may be required to ensure widespread application. It is the hope that these studies will encourage the use of the RGS and other non-evoked spontaneous behavioural pain assessment tools and will improve reporting standards in medical literature that advance pain research.

Keywords : Grimace Scale, Facial Expression, Pain, Animal Behaviour, Tool Development, Affective State, Rat, Animal Models, ARRIVE Guidelines, Reporting Standard

Table of Contents

List of tables

List of figures

List of acronyms

°C	Celsius
ADL	Active Daily Living
*a*MGS	Automatic Mouse Grimace Scale
AMPA	α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid
ARRIVE	Animal Research: Reporting of *In Vivo* Experiments
ASIC-1	Acid Sending Ionic Channel-1
ATP	Adenosine Triphosphate
AU	Action Unit
AWR	Abdominal Withdrawal Reflex
B2	Bradykinin-2
BL	Baseline
CaMkII	Calmodulin-depending Protein Kinase II
CBS	Composite Behaviour Scale
CFA	Complete Freund's Adjuvant
CGRP	Calcitonin gene-related peptide
CI	Confidence Interval
CONSORT	Consolidated Standards of Reporting Trials
COX-2	Cyclooxygenase-2
CPP	Conditioned Place Preference
CRD	Colorectal Distension
DAI	Disease Activity Index
DSS	Dextran Sulfate Sodium
e.g.	*exempli gratia*
EMLA	Eutectic Mixture of Local Anesthetics
ERK	Extracellular Signa-regulated Kinase
et al.	*et alia*
etc.	e*t cetera*
FACS	Facial Action Coding System
fMRI	Functional Magnetic Resonance Imaging

GABA	Gamma-Animo Butyric Acid
HARRP	Harmonised Animal Research Reporting Principles
HIV	Human Immunodeficiency Virus
i.e.	*id est*
IBD	Inflammatory Bowel Disease
ICC	Intra-Class Correlation
IFN	Interferon
IL	Interleukin
IMG	Image
kDa	kilodalton
kg	kilogram
kHz	kilohertz
mg	milligram
MGluR	Metabotropic Glutamate Receptor
MGS	Mouse Grimace Scale
MIA	Monoiodoacetate
MPO	Myeloperoxidase
NGF	Nerve Growth Factor
NK	Neurokinin
NMDA	N-methyl-D-aspartate
nonSUPP	Non-Supporting
NSAID	Nonsteroidal anti-inflammatory drugs
O+V	Observer + video
p.o.	*per os*
PET	Positron Emission Tomography
PKA	Protein Kinase A
PKB	Protein Kinase B
PREPARE	Planning Research and Experimental Procedures on Animal Recommendation for Excellence
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QST	Quantitative Sensory Testing

RCT	Randomised Controlled Trials
RFF	Rodent Face Finder
RGS	Rat Grimace Scale
RT	Real-Time
s	Seconds
s.c.	Subcutaneous
SD	Standard Deviation
SEM	Standard Error of the Mean
STROBE	Strengthening the Reporting of Observational studies in Epidemiology
SUPP	Supporting
TINT	Time to Integrate Nest Test
TNBS	2,4,6-trinitrobenzenesulfonic acid
TrkA	Tropomyosin A
TrkB	Tropomyosin B
TRPA1	Transient Receptor Potential Ankyrin 1
V1	Video only
V2	Video only
VMR	Visceromotor Response
VR	Vanilloid Receptor
α	alpha
µg	microgram

Dedication

This is for Daddy, Mummy and Russell.

Thank you for always supporting me in whatever I do.

Acknowledgements

First and foremost, I am tremendously thankful to Dr. Daniel Pang for this *PhD* opportunity. Thank you for your continued support, confidence and patience throughout this journey. Thank you for being an amazing teacher and mentor, always willing to impart advice for my projects, career and life in general.

Secondly, I am grateful to my parents for their amazing love and support throughout my life. I would not be where I am if you did not encourage me to pursue my passions and interests.

Thank you to all the people who have been on my supervisory committee, comité conseil and my examination committee for their invaluable feedback and encouragement throughout my *PhD* and with the writing of this thesis: Dr Douglas Morck, Dr. Elizabeth O'Toole, Dr. Éric Troncy, Dr. Gregory Muench, Dr. Gilly Griffin, Dr. Lee Niel and Dr. Milagros Freire Gonzalez. I would like to especially thank Dr. Lee Niel for introducing and encouraging me into the world of research during my undergraduate years.

I am very grateful to my fellow Pang Lab members: Cassandra Klune, Chelsea Schuster, Colin Laferrière, Dr. Amy Larkin, Dr. Frédérik Rousseau-Blass, Dr. Geneviève Fortin-Simard, Dr. Hayley Robbins, Dr. Julie Reimer, Dr. Katrina Frost, Dr. Maxime Rufiange. Emily Zhang and Maaria Shah. Thank you for always helping me with my projects, looking out for my rats and being fantastic company. Without all of you my *PhD* experience would not have been as enjoyable. I would like to especially thank Frédérik and Maxime for being abundantly patient with my poor French and always willing to lend a hand with translation.

Thank you to Auntie Audrey and Auntie Clara for supporting me all the way from Singapore by reading some parts of my thesis and instructing me on how to write better.

Thank you also to the 90 rats that were part of my actual or preliminary studies.

And last but most of all, I want to thank my Heavenly Father for guiding me through this path. It has been an enjoyable and trying experience. I know not what lies ahead, but I will do my best to continue to trust in Him.

1. Introduction

Pain is a sensation that everyone experiences at some point in their lives. In humans, the gold standard of pain assessment is by direct verbal report. We can communicate with one another and the medical staff when we are in pain and provide feedback regarding treatment efficacy. This gold standard in pain assessment reporting is not possible in animals as they cannot communicate with us directly. Therefore, pain assessment in animals has largely relied on inferences from their behaviours. Traditionally, pain assessment in preclinical animal research has largely relied on nociceptive tests which assess an animal's reflexive response to an external stimulus. These nociceptive tests are favoured for their practicality: they can be replicated easily and reliably (Mogil and Crager, 2004). However, when the novel analgesics that had been reported as efficacious in preclinical animal trials failed during human clinical trials, it was proposed that nociceptive tests were not the appropriate pain measurement tools they were thought to be (Mogil and Crager, 2004; Rice *et al.,* 2008 and Mogil *et al*., 2010). This is because nociceptive tests only assess the sensory component of pain and are unable to evaluate the affective component of pain, the primary concern reported by human patients (Backonja and Stacey, 2004; Mogil and Crager, 2004; Rice *et al.,* 2008 and Mogil *et al*., 2010). This means that there is a mismatch of pain type assessed during animal and human trials (*i.e.* animal trials assess evoked-reflexive responses to nociceptive tests while human trials assess ongoing pain via verbal report). Therefore, it has been proposed that non-evoked spontaneous behaviours from animals should be used to assess the ongoing pain experienced (affective pain) by the animal (Mogil and Crager, 2004). Other issues with preclinical research have also been identified, one of which is the deficiencies in reporting standards in published papers (Kilkenny *et al*., 2009). The consequences of poorly reported published papers are the obstruction of study replication and validation of findings and will adversely skew systematic reviews and meta-analysis (Kilkenny *et al*., 2009; MacCallum 2010; du Sert, 2011 and Freedman *et al*., 2015).

At the beginning of this PhD, the Rat Grimace Scale, had just been developed by Sotocinal *et al*. (2011). This pain assessment tool is a promising method which utilises spontaneous changes in facial expression to assess ongoing pain experienced by rats. The studies described in this thesis explore the strengths and limitations of the RGS. Specifically, the studies explore:

1) if real-time application of the RGS is feasible, if so this would drastically reduce the time and labour required to obtain pain scores and also expand the usefulness of RGS from a research tool to a clinical tool; 2) if the RGS can be utilised to assess more pain types than originally thought (*i.e.* acute and chronic visceral pain); and 3) if training in RGS use prior to RGS scoring is beneficial by improving reliability.

A study was published by Kilkenny *et al*. (2009) which highlighted the deficiencies of reporting standards in published animal studies. This subsequently prompted the publication of the ARRIVE (Animal Research: Reporting of *In Vivo* Experiments) guidelines a year later by Kilkenny *et al*. (2010) which contained 20 key items that all papers should include for a paper to be well reported and accurate. This thesis will also assess: 1) if reporting standards have improved since the publication of the ARRIVE guidelines and 2) if journals that support the ARRIVE guidelines publish papers with higher reporting standards than journals that do not support the ARRIVE guidelines.

1.1. What is pain?

Pain has been defined by the International Association for the Study of Pain as ‘an unpleasant sensory and emotional experience associated with actual or potential tissue damage or described in terms of such damage’. Pain is an adaptive and protective response to tissue damage and maintains body integrity by preventing further contact with a noxious stimulus (nociceptive pain) that increases tissue damage (inflammatory pain; Woolf, 2010). An injured area is more sensitive to external stimuli and ongoing pain is present, making the injured individual aware of the damage and encouraging protection of the injured area to allow healing. This adaptive pain response dissipates when the injury is healed, and the protective function is no longer needed. However, pain can also be maladaptive (pathological pain) where the sensation of pain outlasts the injury due to structural damages or changes to the nervous system.

1.1.1. The basic mechanism of pain

After an injury, various inflammatory agents are released which activate and sensitise nociceptors (peripheral sensitisation; Coutaux *et al.*, 2005). There are also alterations to neuron properties and nociceptive pathways within the central nervous system that enhance the nociceptive response (central sensitisation; Latremoliere and Woolf, 2009).

Peripheral sensitisation is mediated by nociceptors: Aδ and C fibers. Aδ fibers are lightly myelinated, small in diameter and conduct action potentials slowly (4 – 30 m/s). C fibers are unmyelinated, smaller in diameter and conduct action potentials even more slowly (0.3 – 1.5 m/s). Both fiber types have a high activation threshold and respond to both thermal and mechanical stimuli. C fibers can also be polymodal, allowing them to respond to all stimuli types (thermal, mechanical and chemical). After an injury, damaged cells, platelets, mast cells, macrophages and nerves release neuropeptides and pro-inflammatory cytokines at the injury site and within the central nervous system. This results in a cascade of events that lead to primary hyperalgesia. Damaged cells release H^+ and adenosine triphosphate (ATP) which interact with ASIC-1 (acid-sensing ionic channel), VR-1 (vanilloid receptor) and ATP receptors. These interactions cause cation channels to open and depolarise the nociceptor. Inflammatory agents (*i.e.* bradykinin, prostaglandins, leukotrienes, proinflammatory agents and nerve growth factor

(NGF)) sensitise other receptors, resulting in primary hyperalgesia. During platelet aggregation and mastocyte degranulation, 5-HT and histamine are released and as their concentration increases, pain develops. At the nociceptor, the inflammatory substances bind to their respective receptors and induce the phosphorylation of protein kinases A and C (PKA and PKC). This enhances the efficiency of tetrodotoxin-resistant sodium channels and lowers the thresholds of other receptors (*e.g*. VR-1). NGF forms a complex with TrkA (tropomyosin receptor kinase A) and, within the nociceptor, induces protein synthesis of more tetrodotoxin-resistant sodium channels. The nociceptor itself also releases substance P and calcitonin gene-related peptide (CGRP) which further activates nociceptors. Additionally, substance P induces mastocyte granules to release histamine. The histamine release causes vasodilation, resulting in further mastocyte degranulation and the release of more histamine. The increased release of histamine increases vasodilation and sensitises the nociceptors near the damaged tissues as well as in the surrounding healthy tissues (secondary or spreading hyperalgesia). Overall, there is an amplified response to a stimulus from lower thresholds resulting in the opening of voltage dependent sodium channels open and depolarising of the nociceptors.

The action potential generated spreads through the cell bodies of the nociceptors to the dorsal root ganglion (Mello and Dickenson, 2008). Most Aδ and C fibers terminate in laminae I-II and V and transmit impulses to nociceptive specific cells (in laminae I-II) and to wide- dynamic range neurones (laminae V; Mello and Dickenson, 2008). These impulses are transmitted by releasing glutamate which diffuses across the synapse to activate the AMPA (α-animo-3-hydroxy 5-methyl-4-isoxazeloproprionic acid) and NMDA (N-methyl-D-aspartate) receptors of nociceptive specific cells and wide-dynamic range neurones (Mello and Dickenson, 2008). The acute stimulation of fibers results in the activation of AMPA receptors which set the initial response to noxious and tactile stimuli or the first pain response. The influx of Ca^{2+} into dorsal horn neurons activates PKC and CaMKII (calmodulin-depending protein kinase II), both of which are major effectors of central sensitisation. Repetitive and high-frequency stimulation of C fibers will cause nociceptors to release more glutamate, substance P and CGRP, causing a slow depolarisation of the neurones and removal of the Mg^{2+} NMDA block (long term potentiation). Activation of the NMDA receptors with other receptors (*i.e*. G-coupled MGluR (metabotropic glutamate receptor), NK1 (neurokinin-1), B2 (bradykinin-2) and CGRP1 receptors) result in an increased influx of Ca^{2+} that amplifies and prolongs the spinal dorsal horn neurons to subsequent input – the wind-up phenomenon. The wind-up phenomenon continues for as long as the wide dynamic range neurones receive input, resulting in elevated activation and responsiveness of the dorsal horn neurones. This is the second pain response to the ongoing unpleasant pain sensation.

The spinal neurones in laminae I-II then convey the signal to the parabrachial area, the amygdala and hypothalamus, the periaqueductal grey and the rostral ventromedial medulla, in that order. The affective component of pain (unpleasant ongoing emotional experience) is processed in the amygdala and hypothalamus. Neurones from laminae V in the spinal cord dorsal horn transmit the signal to the thalamus, then to the somatosensory cortex where the sensory component of pain is processed (location and duration of the injury). These effects can be either facilitated or inhibited. In the facilitatory pathway, there is a release of 5HT-3 from the rostral ventromedial medulla which activates 5HT-3 receptors, exerting a pronociceptive effect at the spinal cord level by allowing an influx of Na^{+} (Mello and Dickenson, 2008). The inhibitory pathway begins from the release of norepinephrine into the spinal cord from the brain stem nuclei, resulting in the inhibition of transmitter releases from primary afferent terminals and the suppressed firing of projection neurones in the dorsal horn (Mello and Dickenson, 2008). Inhibition also occurs at the level of the spinal cord when inhibitory neurones in the spinal cord dorsal horn release GABA (gamma-aminobutyric acid) and glycine. These bind to their respective receptors to re-polarise the nociceptors via opening ion channels that allow negatively-charged Cl^{-} and bicarbonate ions to go through the plasma membrane.

Pain may also be modulated via central sensitisation. This is the abnormal enhancement of neurone properties and nociceptive pathways within the central nervous system that results in increased membrane excitability, synaptic efficacy or reduced inhibition (Latremoliere and Woolf, 2009). This may only be triggered if the stimulus is intense, of a long duration and repeated. During this time, phosphorylation by protein kinases (*i.e.* PKA, PKC, CaMKII and ERK [extracellular signal-regulated kinases]) increases synaptic efficacy and membrane excitability by reducing the depolarization threshold and activation of NMDA and AMPA receptors. These protein kinases also recruit and insert more AMPA receptors at the membrane and reduce the outflow of K^+. Lastly, these kinases mediate transcription factors drive gene expression (*e.g.* c-fos, NK1, TrkB [tropomyosin receptor kinase B], Cox2 [cyclooxygenase-2]) which causes the strengthening of the synapse to last longer. The resultant effect is an increased pain response due to increases in spontaneous activity in action potential firing, a reduction in threshold to external stimuli and the enlargement of receptive fields.

1.1.2. Behavioural pain assessment methods in rodents

Laboratory rats are frequently used in pain research as models for human pain. However, it is difficult to assess pain in animals because, unlike humans, they cannot provide verbal feedback to indicate if they experience pain or if a novel analgesic treatment is working. Therefore, pain evaluation in animals is based on inferences from their behaviours. Pain assessments in animals range from the use of nociceptive tests, reflexive responses to an external stimulus (*e.g.* von Frey testing), to specific and non-specific pain behaviours, behaviours that increase or decrease in response to pain or analgesic administration (*e.g.* grimacing). A few of these assessment methods with an emphasis on rodents are described below.

1.1.2.1. Nociceptive tests

1.1.2.1.1. von Frey test

The von Frey test is considered the gold standard method for assessing someone's mechanical threshold (Deuis *et al.*, 2017). It was originally designed to assess the itching sensation in humans and has now been standardised to assess mechanical sensitivity in both humans and animals (von Frey, 1922 as cited by Bove, 2006). This method involves manually applying a nylon monofilament at a right angle to the area of interest, such as the hind paw of a rat, until the filament bends (Barrot, 2012). The monofilaments are of varying diameter; a thicker filament is stiffer and applies a greater force than a thinner filament. During the test, rats are placed on a platform with a mesh bottom that allows them to move about freely, the von Frey filament is applied when the rats are stationary. Filaments of increasing force are applied until the animal withdraws its hind paw or licks or shakes its hind paw (Deuis *et al.*, 2017). The force that evoked the response is noted as the mechanical threshold of the rat. This is performed multiple times to ensure reliability and consistency. This usually results in 5-10 applications of von Frey filaments (Deuis *et al.*, 2017).

The advantages of the von Frey test is that it is simple to use and inexpensive (Bove, 2006). However, the force generated by the application of a von Frey filament can be affected by differences in protocol (speed of application, degree of bending, number of applications) and biases (experimenter and environmental; Bove, 2006). The speed of application can affect the force applied. A quicker application applies the maximum force sooner or may even overshoot the intended force, thus prompting a greater response and underestimates the actual mechanical threshold. The applied mechanical threshold may also be reduced if the filament is flexed too far with over application of strength (filament tip applied at an angle) as the filament tip will apply less force. Repeated applications of the von Frey filaments may also sensitise the area and result in a reduced threshold with more applications. Experimenter bias may also be an issue when experimenters need to make a judgement about whether an observed response is a true positive because they usually know the paw that is affected and will expect a positive response (Wallas *et al.*, 2003 and Bove, 2006). Tested animals may also be affected by environmental factors (*e.g.* stress) resulting in an increase of the threshold which reduces the response. Another factor that may affect the force applied is tissue compliance at the site of application because it affects the way the filament bends and the reaction may be affected by the rat's shift in weight from its inflamed paw to the healthy one, resulting in a systematic error of requiring a greater force to evoke a response on the healthy paw (Bove, 2006).

The von Frey test may also be assessed with an electronic von Frey system. An electronic von Frey system consists of a single filament that applies an increasing force until a paw withdrawal is observed (Deuis *et al.*, 2017). The machine records the force applied automatically and sets it as the paw withdrawal threshold. This requires fewer applications (3-4 applications) compared to the manual von Frey method. The electronic von Frey systems can also analyse the rate at which the force was applied to ensure consistency between applications (Deuis *et al.*, 2017). Overall, the electronic von Frey test method produces fewer data variabilities and may be able to counteract many of the flaws of the manual von Frey method. However, the electronic systems produce different values from the manual von Frey test and comparisons between the two methods is difficult.

1.1.2.1.2. Tail flick test

The tail flick test was first proposed to assess thermal nociception by D'amour and Smith (1941) by applying a radiant heat source on the tip of the rat's tail. The heat source was turned on and the latency for the rat to twitch or move its tail (called a tail flick) was noted. An increased latency of a tail flick was described as analgesia. The intensity of the radiant heat source was set up so that the tail flick was observed after 5s at baseline. The authors performed the test 10,000 times on hundreds of rats and found individual variability to be low. The authors used this test to assess the analgesic properties of various drugs (hydromorphone, heroin, morphine, codeine and pantopon) and observed that increasing doses resulted in an increased latency to tail flick. The tail flick behaviour did not occur at higher doses (4 mg/kg hydomorphone; 4 mg/kg heroin; 12 mg/kg morphine; 30 mg/kg codeine and 24 mg/kg pantopon) even when the tails became burnt. This was described as a loss of reaction to pain by the authors. The latency to perform a tail flick is also affected by heat intensity and the area being stimulated (*e.g.* sensitivity increases when the most distal part of the rat's tail is stimulated and when the heating rate is high; Le Bars *et al.*, 2001). Interestingly, when the rate of heating was slow, the tail flick response did not occur even when the tail was burnt. Therefore, when conducting a tail flick test, a cut-off time of 10-20s must be set to avoid skin burns.

An advantage of this test is the effectiveness of assessing the activity of opioid analgesics with increasing sensitivity when the heat is applied more distally (Le Bars *et al.*, 2001). A disadvantage of the test is that the tail flick response is considered a spinal reflex and may be affected by motor processing changes (Deuis *et al.*, 2017). Additionally, the latency to perform the tail flick behaviour is impacted by stress and requires proper habituation and acclimatisation to ensure reliable and repeatable testing (Le Bars *et al.*, 2001). This test is also affected by the ambient temperature during testing as the rat's tail is important for thermoregulation and responsible for dissipating up to 20% of the rat's body heat (Berge *et al.*, 1988; Le Bars *et al.*, 2001). Consequently, it has been found that a higher ambient temperature increases the temperature of the tail and reduces the tail flick latency.

1.1.2.1.3. Paw withdrawal test/ Hargreaves test

Hargreaves *et al*. first proposed this test in 1988. A rat was placed in a plastic box with a glass floor and a radiant heat source was aimed at the rat's hind paws from below. The following four measurements were assessed: 1) paw withdrawal latency, 2) whether the withdrawal reflex was completed within a second; 3) licking behaviour and 4) duration of hind paw withdrawal. The authors observed that there was a decreased latency after the rat was treated with an intra-plantar carrageenan injection (decreasing from around 10s to 4s) at 1, 2.5 and 4 hours later. It was also observed that a higher carrageenan dose of 2 mg/kg had a shorter latency compared to 0.5 or 1 mg/kg of carrageenan. Withdrawal latencies returned to baseline levels when the rats were administered 3 mg/kg of morphine. It was also observed that the carrageenan-injected rats had a slower withdrawal movement, were more likely to lick their hind paw and withdraw their hind paws for a longer period of time in comparison to the saline-injected rats. The test was unaffected by repeated testing as the latency of paw withdrawal on the contralateral paw and in saline injected animals remained stable. When the heat was applied quickly (6.5 °C/s), the paw withdrawal reaction time was short and the skin surface temperature reached a higher level (Le Bars *et al.*, 2001). This suggests that the Aδ fibers are activated when heat was applied quickly. However, when the heating process was slow, the reaction time was longer and the skin temperature increased less, thus activating only the C fibers. The effects of morphine were more evident during the second phase compared to the first phase (Le Bars *et al.*, 2001). When the temperature increased slowly (1 °C/min) the paw and plate temperatures were close and paw withdrawal was observed at around 39-40 ˚C, which corresponds to the temperature at which thermo-nociceptors are activated (Yeomans and Proudfit, 1996).

This test is useful for assessing unilateral models of pain. The inflamed and contralateral hind paws can be compared allowing each animal to act as its own control, thus reducing variability (Barrot, 2012 and Deuis *et al.,* 2017). Furthermore, this test allows animals to be free ranging (within the testing apparatus) and therefore reduces the possibility of stress-induced analgesia (Barrot, 2012 and Deuis *et al.*, 2017). The disadvantage of this test is the need for the animal to acclimatise to the testing apparatus to minimise exploratory behaviours. Also, behavioural responses can differ between species and strains (Barrot, 2012 and Deuis *et al.*, 2017). An alternative to this test has been proposed where the radiant applied heat is increased by 2.5 °C increments every 10s until a paw withdrawal is observed (Banik and Kabadi, 2013). However, this method takes a longer time and is not available commercially (Deuis *et al.*, 2017).

1.1.2.1.4. Hot plate test

This test was first proposed by Woolfe and MacDonald (1944). Mice were placed on a hot metal plate by trapping them within an overturned glass beaker. The mice were observed for a series of behaviours that were performed chronologically: the mice sat on their hind paws, licked their hind paws, kicked up their hind paws, and then attempted to escape. The first behaviour of paw licking was observed 30s after placement on the hot plate. When increasing doses of analgesics were administered, fewer and fewer animals reacted after 30s on the hot plate: the authors described this as analgesia. At temperatures below 50 °C, there was a large variation - some mice presented signs of discomfort while others appeared comfortable and did not attempt to escape. It was at the higher temperatures (55 °C) that all the mice reacted consistently within 30s and latency of reactions was shorter. Unlike mice, rats did not display a predictable chain of behaviour. Therefore, the sensitivity to assess the analgesic efficacy can be improved by assessing the rats' latency to perform any behaviour (sniffing, grooming, stamping, freezing, licking and jumping; Plone *et al.*, 1996). Further improvements in sensitivity are achieved if lower temperatures (50 °C vs 55 °C) are used (Plone *et al.*, 1996). An alternative procedure is the dynamic hot plate test where a rat is placed on a hot plate at a comfortable temperature (< 42 °C) and the temperature is increased consistently until the licking behaviour is observed (Ogren and Berge, 1984). The temperature at which the response is observed is designated as the response temperature. This is affected by the temperature at the beginning of the study, the room temperature and the heating rate (Tjolsen *et al.*, 1991). Overall, a variety of behaviours has been observed during the conduct of this test (sniffing, grooming, stamping, freezing, licking, leaning and jumping). However, the data was less variable if assessment simply consisted of assessing the latency to perform any of the behaviours mentioned above, and if lower temperatures were utilised (Plone *et al.*, 1996). The licking and jumping behaviours were considered to be supraspinally mediated responses because the rats no longer performed these behaviours after a spinal transection (Le Bars *et al.*, 2001 and Giglio *et al.*, 2016). It was also observed that different types of analgesics affected different behaviours (*i.e*. the increased latency to licking was observed with opioid administration, and the latency to jump was observed with less potent analgesics like acetylsalicylic acid or paracetamol; Le Bars *et al.*, 2001). While these behaviours were fairly stereotypical in mice, the behaviours of rats were more irregular and did not seem to follow a specific pattern (Le Bars *et al.*, 2001). The disadvantages of this test are the variability observed in the data generated even within a single laboratory, and that the animals learn over multiple testing sessions that the performance of certain behaviours result in their removal from the apparatus (Plone *et al.*, 1996; Le Bars *et al.*, 2001 and Barrot, 2012).

1.1.2.2. Non-evoked spontaneous behaviours

1.1.2.2.1. Conditioned place preference

Conditioned place preference (CPP) has traditionally been used to assess the reinforcing and associated rewarding effects of drugs (Sufka, 1994). It was proposed as an assessment for the rewarding effects of analgesics during pain by Sufka (1994). This test utilises the idea that pain is an unpleasant sensation and that alleviation of that pain with analgesia is rewarding. In this test, rats are placed in an apparatus with three compartments: two stimuli-distinct compartments that usually differ by colour (black and white) and could also differ by floor type and bedding (Sufka, 1994). The third compartment is built in between two distinct stimuli-compartments. This compartment tends to be grey in colour with doors leading to the stimuli-distinct compartments. During this test, the pain model is induced in the animals and they go through three conditioning and testing trials: 1) pre-conditioning – the rats are allowed to explore all three compartments freely for 15 minutes; 2) drug conditioning trials – the animals are administered the drug immediately before being confined to the animal's non-preferred compartment (usually the white compartment) for 60 minutes and this is repeated with the vehicle control, but with the black compartment. These drug conditioning trials are repeated four times for each drug and vehicle control and a trial is only performed once per day. Lastly, 3) testing trials – the animals are confined in the neutral compartment and allowed to explore all the compartments. Two assessments are made: 1) time assessment – measuring the duration the animals spend in the drug or vehicle associated compartment and 2) choice assessment –

assessing the compartment that is entered first by the rat when it is allowed to leave the middle neutral compartment (Sufka, 1994). In a subsequent study, it was found that a single drug conditioning trial was sufficient to assess the efficacy of various analgesics (King *et al*., 2009). Animals could associate the analgesic or vehicle controls with their respective stimuli distinct compartments after one drug conditioning trial of 15 minutes for each analgesic and vehicle control. These drug conditioning trials can also be conducted on the same day and this greatly reduced the time required to perform this test.

During the initial study by Sufka (1994), rats were administered an intra-plantar CFA (complete Freund's adjuvant) injection on day 1. They then went through the drug conditioning trials on days 2-9 and were assessed on days 10-13. It was expected that CFA-treated animals would prefer the compartments associated with the drugs (morphine, MK-801 (an NMDA receptor antagonist) and indomethacin (a non-steroidal anti-inflammatory drug)) and non-CFA treated animals would show no preference. However, it was found that CFA-treated animals were more likely to enter and preferred to spend more time in the drug associated compartment with a low dose of MK-801 (0.03 mg/kg) while non-CFA-treated animals showed no preference at this dose. Interestingly, when a high dose of MK-801 (0.3 mg/kg) was administered, both CFA- and non-CFA treated animals entered the vehicle associated compartment and also spent more time in there. This suggests that MK-801 is aversive at this dose and that any analgesic property is insufficient to offset this aversiveness. When rats were administered a high dose of morphine (10 mg/kg) both CFA- and non-CFA-treated animals entered the drug-associated compartment first and spent more time there. While CFA-treated animals were more likely to enter the morphine-associated compartment at a lower dose (3 mg/kg), they did not spend significantly more time in this compartment, and non-CFA-treated animals did not prefer this compartment. This suggests that at a high dose, morphine has rewarding effects that are not associated with its analgesic properties as non-CFA-treated animals also found it rewarding. At a lower dose, CFA-treated animals may find it rewarding enough to enter that compartment, but it was not rewarding enough for them to remain in there. When animals were administered indomethacin, a preference for the drug-paired compartment was not observed in either CFA or non-CFA-treated animals. From these results, the author concluded that assessing the animal's choice of which compartment to enter first was more sensitive than assessing how long the animals spent in each compartment as the animal may be influenced by other factors when choosing whether to remain in the drug-associated compartment (Sufka, 1994). However, the observed negative result with morphine and indomethacin is probably better explained by the chosen assessment times. Ongoing pain was observed to peak at 6 hours after an intra-plantar CFA injection; it decreased to below an intervention threshold 48 hours later; and it returned to baseline levels 7 days later when assessed with the Rat Grimace Scale (Sotocinal *et al*., 2011 and De Rantere *et al*., 2016). Furthermore, it was demonstrated that morphine alleviated ongoing pain at 2 and 5 mg/kg doses (Sotocinal *et al.*, 2011). Therefore, when the CPP assessments were performed by Sufka (1994), ongoing pain was likely no longer present and therefore, the analgesic properties of drugs were not rewarding enough to keep the animals in the drug-associated compartment. This is further supported by later CPP studies that also utilised the intra-plantar CFA pain model where the animals were tested on the same day the CFA was administered. These animals consistently spent more time in the drug (lidocaine and clonidine) paired compartment (Okun *et al*., 2011 and He *et al*., 2012).

A variety of chronic pain models has since been assessed with this test (spared nerve ligation, spinal lesion, osteoarthritis via intra-articular injections of MIA (monoiodoacetate), sciatic nerve axotomy; King *et al.*, 2009; Davoody *et al.*, 2011; Liu *et al.*, 2011; Qu *et al.*, 2011; He *et al.*, 2012 and Okun *et al.*, 2012). The animals consistently spent more time in the drug paired compartment at doses of analgesic drugs that alleviate mechanical hyperalgesia (Clonidine, conotoxin, lidocaine; King *et al.*, 2009). Interestingly, rats did not prefer the compartment associated with adenosine, a drug which was demonstrated to reduce hyperalgesia but not for ongoing pain in humans (King *et al.*, 2009). Additionally, CPP no longer occurred when neurons from the rostral anterior cingulate cortex (the part of the brain that processes ongoing pain) were severed (Johansen *et al.*, 2001). These results demonstrate that CPP is motivated by ongoing pain.

The use of CPP for pain assessment is advantageous because it is relatively reliable, can be performed readily, is quite easy to interpret, and may also assess movement-evoked pain since voluntary limb movement is required to travel within the apparatus (Li, 2013). This test is not only able to assess if a potential new drug is able to alleviate pain, it is also able to assess the possible aversive, rewarding, or potentially abusive effects (Sufka, 1994 and Li, 2013). However, this test is quite time consuming as rats need to be trained to associate a compartment with a drug or its vehicle. Furthermore, this test requires the presence of ongoing aversive pain to assess the rewarding effects of an analgesic (Li, 2013). Lastly, this test may be useful for pain assessment in nerve injury models, but not for pain models that are short and paroxysmal (Li, 2013).

1.1.2.2.2. Composite behaviour score

The composite behaviour score (CBS) is a rat ethogram composed of multiple behaviours that are present after a laparotomy and which subside with analgesics. This pain assessment score was developed in 2000 by Roughan and Flecknell by observing the frequency and duration of 150 behaviours in rats which were moved to the surgery room and placed in an induction box with oxygen only or with oxygen and isoflurane. Rats were also observed after they were administered subcutaneous injections of saline, ketoprofen (5 mg/kg) or buprenorphine (0.05 mg/kg). Observations were over 24-hour intervals. The 150 behaviours were then reduced to 40 by identifying behaviours that were different after the above-mentioned procedures. These 40 behaviours were then categorised as: active, inactive, attentive, grooming and sleeping behaviours. At baseline, the rats displayed similar patterns of behaviours and spent a similar proportion of time performing the different categories of behaviours throughout the day. Each procedure affected how the rats spent the next 24 hours. After being moved to the surgery room with or without exposure to anesthesia, the rats displayed less active, inactive and grooming behaviours and slept more compared to their baseline. These activities were further affected by analgesics: the animals which were administered ketoprofen performed less attentive and grooming behaviours compared to the baseline; and the animals which were administered buprenorphine displayed an increase in inactive, active and attentive behaviour and a decrease in sleep behaviour compared to the other groups and to their own baseline. Animals that underwent a laparotomy surgery and administered saline performed more inactive behaviours compared to animals that received ketoprofen prior to the laparotomy procedure. When animals received buprenorphine prior to surgery, they displayed more attentive behaviours, decreased sleep and grooming behaviours compared to their baseline behaviours and to rats which received saline or ketoprofen prior to surgery. Overall, the authors concluded that the move to the surgery room and the preparation for surgery were stressful and resulted in an alteration of behavioural patterns. They also concluded that buprenorphine caused behavioural changes in rats and that it was likely to affect any behaviour related to pain and, therefore, NSAIDs (non-steroidal anti-inflammatory drugs) like ketoprofen were probably better as they did not affect the rats' behaviour.

In subsequent studies, many behaviours were again observed and assessed if they increased after a laparotomy surgery and decreased with analgesic administration (Roughan and Flecknell, 2001). Five key behaviours were identified: twitch, stagger, back arch, writhing and belly pressing (Roughan and Flecknell, 2001; 2003; 2006 and Thomas *et al.*, 2016). However, behaviours were excluded if they were found to be absent (*e.g.* back arch) or were too variable (*e.g.* twitch) within a particular study (Roughan and Flecknell, 2003 and Thomas *et al.*, 2016). In general, these behaviours were observed to increase after a laparotomy and decreased or occurred at a lower frequency when an analgesic was administered (ketoprofen (5, 10, 15 mg/kg), carprofen (5, 10, 15 mg/kg) or meloxicam (1 or 2 mg/kg)). The alteration in behaviours lasted 4-5 hours after the laparotomy surgery (Roughan and Flecknell, 2001; 2003). The frequency of these behaviours also decreased with a dose of analgesia, thus displaying a dose dependent change which could be identified with a 5- or 10-minute observation time (Roughan and Flecknell, 2003). The observers were also able to accurately distinguish if the animals received an analgesic (buprenorphine 0.05 mg/kg or carprofen 5 mg/kg) 90% of the time (Roughan and Flecknell, 2004). This is a practical approach to the assessment of pain as these behaviours are distinct and a 5-minute observation period is sufficient to differentiate between groups (Roughan and Flecknell, 2003). It is also user-friendly as inexperienced observers were able to recognise these behaviours after a short training session (Roughan and Flecknell, 2006). Lastly, there is no need to train the rats as the behaviours occur naturally. The limitations of this ethogram are that there is no single behaviour that can predict analgesic dosage or pain severity (Roughan and Flecknell, 2003); and it is has been criticized as complicated, time-consuming and impractical (Waite *et al.*, 2015). Additionally, while some of the behaviours identified were also observed in a ureteral calculosi and an intestinal mucositis model (Giamberardino *et al.*, 1995 and Whittaker *et al.*, 2014), they were not consistently observed in a bladder cancer model

(Roughan *et al*., 2004). The authors did observe twitching and back arching behaviours in the bladder cancer model, but these behaviours did not decrease following analgesic administration (carprofen (5 mg/kg) or meloxicam (2 mg/kg)). Instead, behaviours that could differentiate between the groups were: abdominal licking, circulatory ambulation, digging, coprophagy and shaking. However, the frequency of these behaviours was very variable. The authors attributed the ineffectiveness of their scale to the severe pain that prevented the animals from performing the behaviours because the animals were inactive during observation (Roughan *et al*., 2004).

1.1.2.2.3. Grimace scales

Grimace scales utilise the facial expressions of animals to assess the severity of pain experienced. This is not a novel idea as facial expressions have been well defined in humans with "action units" (AUs; Cohn *et al*., 2007) and have been used to assess pain in patients who cannot verbally communicate (*i.e.* infants and dementia patients; Williams, 2002, Kunz *et al.*, 2007 and Kohut *et al*., 2012). This is possible as facial expressions are innate and usually occur spontaneously. Therefore, observation of someone's facial expressions allow others to perceive his emotions or pain severity (Williams, 2002). The use of facial expressions to assess pain in animals was first proposed by Langford *et al*. (2010) with the creation of the Mouse Grimace Scale (MGS).

The MGS was first assessed by first video recording the mice for 30 minutes (baseline/ no 'pain' videos) followed by an acetic acid abdominal constriction test (*i.e.* acetic acid injected intraperitoneally) and a video recording again ('pain videos'). Ten images of the faces of the mice were extracted from the videos and sent to human facial pain expression experts. The experts then identified five AUs that were most likely to assess pain: orbital tightening – tightly closed eye squeeze, nose bulge – rounded appearance of the nose pad, cheek bulge – rounded appearance of the cheeks, ear position – ears pulled away from the front of the face, and whisker changes – whiskers clumped together and pulled towards the cheeks. Each of these AUs were then assigned a score of either 0, 1 or 2 to indicate the degree of its presence. The authors noted three similar AUs were present in mice and in humans (*i.e.* orbital tightening and nose and cheek bulge; Prkachin, 1992 and Langford *et al.*, 2010). This provided evidence that facial expressions of pain are evolutionarily conserved.

The MGS was applied to a variety of pain models varying by intensity, duration and pain type. Photos of mice were compared before and after each pain model. The following observations were made: 1) the MGS scores were more likely to increase when noxious stimuli was of moderate duration (10 min to 4 hours) but did not increase in pain lasting more than a day or in neuropathic pain models; 2) the MGS scores were higher when noxious stimuli was applied more deeply compared to superficial stimulus (joint and viscera compared to subcutaneous); 3) the MGS scores increased in a dose dependent manner to the inflammatory stimulus administered; and 4) the MGS scores decreased in a dose dependent manner with morphine administration. The MGS was considered ineffective for the assessment of chronic pain because the MGS scores did not increase in the neuropathic pain model assessed (spared nerve injury and chronic constriction model) up to 14 days after surgery. However, the authors commented that the MGS scores might be confounded by the paroxysmal nature of pain and by stress induced analgesia. Later studies reported increases in the MGS scores after other types of neuropathic pain models (Wu *et al*., 2016 and Akintola *et al*., 2017). In one such study, MGS scores increase 21 to 24 days after the surgery (chronic constriction of the infraorbital nerve; Akintola *et al*., 2017). The authors suggested that the differences in the results to the original MGS paper were probably due to the differences in pain intensity and the involvement of different mechanisms and brain structure.

Overall, the MGS demonstrates good face, content and construct validity as well as good inter-rater reliability (intra-class correlations (ICC; average) = 0.90; Langford *et al*., 2010). The AUs display good internal consistency (Cronbach's α = 0.89) and high accuracy to discriminate between 'pain' and 'no pain' animals (accuracy = 72%; Langford *et al*., 2010).

After the development of the MGS, the Rat Grimace Scale (RGS) was developed by Sotocinal *et al*. (2011). The RGS was developed with acute inflammatory pain models (intra-plantar CFA, intra-plantar carrageenan/kaolin and laparotomy model). Like the MGS, the RGS consisted of similar AUs: orbital tightening, ear and whisker changes. However, in the RGS, the nose and cheeks were observed to flatten simultaneously when rats were in pain. These two action units were merged to form a single AU: nose/cheek flattening. In the tested models, the RGS scores increased over time before decreasing, displaying changes in pain intensity over time. The RGS scores also decreased in a dose dependent manner to morphine administration. Like the MGS, the RGS displayed construct validity, good reliability between raters (ICC = 0.90) and a high accuracy to discriminate 'pain' and 'no pain' animals (accuracy = 82%; Sotocinal *et al*., 2011).

Since then, facial grimace scales in animals have been a promising method to assess pain and many complete or partial grimace scales have been developed for different species of animals (*i.e.* rabbits, horses, lamb, sheep, piglets, cats, ferrets and seals; Keating *et al*., 2012; Dalla Costa *et al*., 2014; Holden *et al*., 2014; Di Giminiani *et al*., 2016; Guesgen *et al*., 2016; McLennan *et al*., 2016; Hager *et al*., 2017; Mullard *et al*., 2017; Reijgwart *et al*., 2017; Viscardi *et al*., 2017 and MacRae *et al*., 2018). The AUs for each grimace scale and the pain model used to develop the scales as well as the validation methods have been summarised in Appendix A.

Construct validity was demonstrated for the majority of the grimace scales with scores increasing after a painful procedure or during a painful disease process (Appendix A). However, some scales have not been successfully completed or validated, such as the piglet and ferret grimace scales, with only an observed increase in a single AU (orbital tightening; Di Giminiani *et al*., 2016 and Reijgwart *et al*., 2017). Therefore, more work is evidently required for some grimace scales. In general, scoring is found to be difficult when performed in low lighting, with low quality photos and with dark-coated animals. Furthermore, confounding factors that may influence changes in facial expressions must be taken into consideration. This has been observed in some non-painful situations (*e.g.* fear, aggression and stress; Defensor *et al*., 2012; Boissy *et al*., 2014; Sorge *et al*., 2014; Dalla Costa *et al*., 2017 and Senko *et al*., 2017).

1.1.2.3. Non-specific behaviours/welfare measures

1.1.2.3.1. Burrowing

Burrowing is an evolutionarily conserved behaviour observed in many laboratory rodent species (Deacon 2006; 2009). Burrowing functions as protection from predators, the weather and for food storage in wild rodents (Deacon, 2006). Burrowing is no longer a functional behaviour in laboratory rodents because they have a steady supply of food and are not exposed to predators. However, they will still burrow even when they are provided with shelters or preexisting burrows but will not utilise the burrows (*i.e.* for sleeping or food stashing; Sherwin *et al*., 2004; Stryjek *et al*., 2012; Makowska and Weary, 2016 and Gould *et al*., 2016). Therefore, burrowing seems to be a self-motivating and self-rewarding behaviour in laboratory rodents. Burrowing behaviour has been suggested as the rodent's equivalent to the human activities of daily living (ADL; *i.e.* working, performing chores; Deacon, 2009). Burrowing may, therefore, be described as a measure of well-being in rodents (Deacon, 2006). While burrowing is not input specific to pain (*i.e.* pain is not the only factor that affects burrowing), it has a pain specific component (Bryden *et al*., 2015). Burrowing behaviour is quantified by measuring the weight of the substrate remaining in the burrowing tube (*e.g.* 2.5 kg of gravel) or by measuring the latency to initiate burrowing (Deacon, 2006; Jirkof *et al*., 2010 and Andrews *et al*., 2012).

So far, alterations in burrowing behaviour have primarily been used to assess models of inflammatory and neuropathic pain that occur in a rodent's limb (Andrews *et al*., 2012; Lau *et al*., 2013; Rutten *et al*., 2014ab; Bryden *et al*., 2015; Gould *et al*., 2016; Muralidharan *et al*., 2016 and Wodarski *et al*., 2016) and one model of generalised neuropathic pain from HIV (Human Immunodeficiency Virus) drug treatment (Huang *et al*., 2013). Other pain models not localised to the limbs have observed limited success in quantifying pain with burrowing behaviour (*i.e.* chemotherapy induced mucositis and migraine pain model; Whittaker *et al*., 2015 and Christensen *et al*., 2016). The use of burrowing as an indicator of pain cannot be discounted from these pain models as it was possibly confounded by the study design. In the migraine model, the authors commented that the model may not have been severe enough because they administered a lower dose (Christensen *et al*., 2016). In the mucositis model, large variations within the data during baseline assessments may have masked any differences from baseline (Whittaker *et al*., 2015). Burrowing behaviour at baseline was observed to increase with exposure to the burrowing tube, and therefore, selection of the number of baseline days is vital (Deacon, 2006 and Whittaker *et al*., 2015). This can be potentially corrected by averaging the amounts burrowed over the baseline days (Andrews *et al*., 2012). Additionally, it has been reported that while the variability within individual rats is low, the variability between rats is high, therefore, rats should be used as their own controls (Andrews *et al*., 2012 and Bryden *et al*., 2015). Although the motivation to burrow is not motivated by anxiety/stress or shelter seeking (Gould *et al*., 2016) it may still be affected by stress (Whittaker *et al*., 2015).

Additionally, burrowing behaviour in female rats is dependent on their estrous cycle where rats in proestrus burrow significantly less than female rats in their estrous or diestrus phase (Christensen *et al*., 2016).

Burrowing was demonstrated to have good face validity as a measure of pain as it is a spontaneous and self-rewarding behaviour (Deacon, 2006) and decreases following an injury, similar to a patient's avoidance of ADL when in pain (Andrews *et al.*, 2012). It also demonstrates good construct validity as burrowing decreases with pain model severity, and recovers following analgesic administration (Andrews *et al*., 2012; Lau *et al*., 2013; Rutten *et al.*, 2014ab and Gould *et al*., 2016). Interestingly, because indication of analgesic efficacy is dependent on recovering burrowing behaviour and not on depressing responses (*e.g.* reflex response with nociceptive threshold testing), burrowing behaviour is not susceptible to false positives from drugs that cause motor impairment or sedative effects (Andrews *et al*., 2012; Bryden *et al*., 2015; Rutten *et al*., 2014ab and Gould *et al*., 2016). It may also be unique as a pain assessment tool because it may encompass both affective (pain on motivation) and sensory (pain from burrowing) components of pain (Bryden *et al*., 2015). It has demonstrated good repeatability in different laboratories and is a robust assessment of pain (Andrews *et al*., 2012 and Wodarski *et al*., 2016). While no strain differences have been observed, it has only been assessed in two strains of rats to date (Sprague Dawley and Wistar rats; Wodarski *et al*., 2016). Despite this, burrowing proficiency varied between laboratory mice strains and wild rats burrowed more than Wistar rats (Deacon, 2009 and Stryjek *et al*., 2012).

A disadvantage of burrowing as an assessment of pain is that it does not correlate with nociception threshold testing (Muralidharan *et al.*, 2016 and Lau *et al*., 2013). Therefore, it requires comparison to other behavioural assessments methods to ascertain criterion validity (Wodarski *et al*, 2016).

1.1.2.3.2. Nest building

Much like burrowing, nest building is an innate and a highly motivated behaviour in rodents (*i.e*. mice will work to obtain nesting materials) and is considered an ADL event (Roper, 1973). This behaviour is well characterised in laboratory mice who will build complex nests from any materials available (Van der Weerd *et al*., 1998). Mice that have undergone an invasive or painful procedure will be unable to build nests or will trample and destroy nests (Deacon *et al*., 2002, Arras *et al*., 2007, Jirkof *et al*., 2013, Hager *et al*., 2015 and Negus *et al.,* 2015). Injured mice also took a longer time to initiate nest building behaviour (Jirkof *et al.,* 2013 and Rock *et al*., 2014). Nest building behaviour recovered or was maintained if analgesics were administered (Arras *et al*., 2007, Negus *et al*., 2015).

Mice are highly motivated to engage in nest building behaviour and will consistently perform this behaviour. Nest building behaviour is more variable in laboratory rats. Laboratory rats do not usually build nests unless pregnant. Denenberg et al. (1968) found that, if pregnant, the females were highly motivated to build a nest and this behaviour increased when they were close to parturition. This behaviour by pregnant females is likely to reduce environmental fluctuations. Nest building behaviour has also been reported in 10-week old male rats of various strains who will utilise nesting materials and incorporate it into a nesting box (Jegstrup *et al*., 2005). In another study, it was reported that rats of both sexes only displayed nest building behaviour if exposed to nesting materials at 4 weeks of age (Van Loo and Baumans, 2004). If the exposure was delayed until 8 weeks of age, this behaviour was not observed (Van Loo and Baumans, 2004). Shredding and utilisation of nesting material was observed in castrated males and ovariectomized females, especially when the ambient temperature was low (13 °C) but not when it was high (22 °C; Deneberg *et al*., 1968). This suggests that rats are not highly motivated to engage in nest building behaviour but under certain conditions or stimuli, rats will make use of nesting materials.

Nest building behaviour in mice has been assessed three ways: 1) naturalistic nest scoring method, 2) time to integrate nest test (TINT) and 3) nest consolidation test (Hess *et al*., 2008; Rock *et al*., 2014; Negus *et al*., 2015). With the naturalistic nest scoring method, the quality of the nest is scored from 0 to 5, from untouched nesting materials to a fully built nest with an enclosed dome (Hess *et al*., 2008). The TINT method assesses the latency for mice to initiate the nest building behaviour and mice will usually initiate nest building behaviour minutes after being presented with nesting materials (Jirkof *et al*., 2013 and Rock *et al*., 2014). Lastly, nest building behaviour can also be assessed by dividing the cage floor equally by six, separating nesting materials into each area and assessing the number of pieces of nesting materials incorporated into a nest (Negus *et al*., 2015).

This behaviour decreases with pain, increases with analgesic administration and has been demonstrated as a more sensitive measure of pain compared to weight loss and the von Frey test (Oliver *et al*., 2018). This behaviour can be used as an initial assessment of pain with the TINT method because it can be performed quickly. It can also be used as a follow up pain assessment method as nest quality can be assessed a few hours later. It was demonstrated that if nesting materials were given in the morning, assessments of the nest can be performed at the end of the day and the condition of the mice assessed retrospectively (Jirkof *et al*., 2013). The requirement to wait hours for mice to construct their nest before assessments is disadvantageous as the mice might experience pain during those hours between assessments. Another disadvantage is that nest building behaviour is affected by a variety of factors such as strain, type of nesting materials given and exposure to anesthesia (Hess *et al*., 2008; Jirkof *et al.,* 2013 and Rock *et al*., 2014).

1.1.2.3.3. Grooming

Grooming behaviour is another innate behaviour frequently performed by rodents, and rats spend around 40% of their waking moments grooming (Bolles, 1960). Grooming consists of fur licking, scratching with hind legs and face washing (Bolles, 1960). Grooming is performed to clean the fur, for thermoregulation, and as a stress displacement activity (Jolles *et al.*, 1979, Cohn and Price, 1979 and Denmark *et al.*, 2010). Grooming behaviour can be modulated by experimental manipulation, dopaminergic drugs, genetic mutations and psychological stress (Kalueff *et al.*, 2016). Grooming behaviour is a highly stereotyped behaviour which follows a fixed action pattern which will be completed without sensory feedback once initiated (Kalueff *et al*., 2016).

For pain assessment, grooming behaviour has been assessed two ways: 1) frequency or duration spent grooming after a procedure and 2) the pattern of grooming performed. In general, rats groomed more after a painful procedure (*i.e.* laparotomy surgery, spinal cord injury and facial pain via injection of formalin; Clavelou *et al*., 1989; Roughan and Flecknell, 2000; Gorman *et al*., 2001) and grooming decreases in a dose dependent manner with analgesic administration (*i.e.* buprenorphine, IL-10, memantine (an NMDA-antagonist), morphine, acetysalic acid and paracetamol; Eisenberg *et al*., 1993; Roughan and Flecknell, 2000; Clavelou *et al* 1989). Grooming duration or intensity may be assessed by direct observation or assessment of hair loss and subcutaneous damage (Roughan and Flecknell, 2000 and Plunkett *et al.,* 2001). Interestingly, in a model of facial pain by injection of formalin into the muzzle, a biphasic facial grooming response (phasic and tonic phase) was observed (Eisenberg *et al.,* 1993). During the tonic phase, rats groomed during the first six minutes followed by a period of no grooming before grooming more intensely and for a longer period for 12-42 minutes after formalin injection.

Pain can also be assessed by observing the pattern or type of grooming performed. For example, directed asymmetrical grooming was observed after formalin injection in one muzzle pad or chronic constriction to the infraorbital nerve (*i.e.* grooming focused on injured side; Eisenberg *et al.,* 1993; Vos *et al*., 1994; 1998). Under non-painful conditions, rats groom primarily with small strokes (*i.e.* grooming focused only on muzzle area) but with pain (*e.g*. after formalin injection or chronic constriction of the infraorbital nerve) grooming is directed over a larger area (from below the eye area to muzzle; Vos *et al*., 1994; 1998). Conversely, when rats were exposed to non-painful stimuli (*i.e.* mineral oil dripped on whiskers, whisker clipping or injection of bupivacaine to muzzle) grooming with small strokes on both sides of the face occurred (Vos *et al*., 1998). More recently, grooming behaviour after a painful laparotomy procedure was assessed by the grooming transfer method in mice (Oliver *et al*., 2018). This method took advantage of the predictable pattern of grooming by observing the transfer of a fluorescent liquid (first applied to the head, between the ears) during grooming. The transfer of the fluorescent substance was assessed in the dark with an ultraviolet light. With this method, mice were found to have lower grooming transfer scores at 8 and 12 hours after the laparotomy procedure compared to their baseline scores. This study also reported strain (CD-1 and C57BL/6 mice) differences but this was negligible within group and between sex. These scores differed temporally from nesting scores: pain depressed nesting scores for up to 24 hours after the laparotomy procedure but for up to 48 hours in the case of grooming transfer scores. The differences between the two methods were also evident with analgesic treatment: buprenorphine (0.1 mg/kg s.c.) alone or in combination with carprofen (30 mg/kg *p.o.*). This analgesic protocol improved grooming transfer scores but did not affect nesting scores. The authors attributed this difference to the two assessment methods potentially assessing different aspects of pain or motivations to perform these behaviours.

The advantage of this method is its innate nature and the frequency it is performed by rodents. Using the grooming transfer method, it is easily quantifiable and identifiable. However, to reduce variability, baseline scores must be obtained. Additionally, because the frequency at which rodents groom varies over the course of a day (Bolles, 1960 and Oliver *et al.*, 2018), baseline scores must be obtained over a similar time frame as the experimental days. This reduces the practicality of the method. Furthermore, grooming behaviour is affected by external factors (*i.e.* stress and drugs), which may confound the scores. Lastly, while reliability to asses pain with the grooming transfer method was reportedly excellent, inter-rater reliability scores were poor (Oliver *et al.*, 2018). Therefore, the rater performing the assessment should be consistent throughout the study or additional training may be required to improve inter-rater reliability.

1.1.2.3.4. Ultrasonic vocalisation

Ultrasonic vocalisations are emitted by rats during play, mating, agonistic behaviour and to warn conspecifics of danger (Thomas and Barfield, 1985; Blanchard *et al.*, 1991; Haney and Miczek, 1993; Vivian and Miczek, 1993; Jourdan *et al.*, 1995 and Knutson *et al.*, 1998). In general, ultrasonic vocalizations in the range between 22 – 28 kHz are associated with a negative affective state (*i.e.* during stressful or painful situations; Calvino *et al.,* 1996) while vocalisations in the range of 50-55 kHz are associated with a positive affective state (*i.e.* during play; Knutson *et al.*, 1998 and Cloutier *et al.*, 2012)).

Ultrasonic vocalisations are recorded with a microphone and bat detector and processed with a computer program before being measured manually by an experimenter. Rats with adjuvant-induced arthritis vocalised ultrasonically when approached by a healthy and heavier conspecific but were silent when alone (Calvino *et al.*, 1996). The duration of ultrasonic vocalisations decreased when analgesics (aspirin and morphine) were administered to the painful rats. However, the occurrence of these ultrasonic vocalisations were observed in some studies (Han *et al.*, 2005; Naito *et al.*, 2006 and Kurejova *et al.*, 2010) but not in others (Jourdan *et al.*, 2002; Wallace *et al.*, 2005; Williams *et al.*, 2008). This disparity of results was observed in both rats and mice studies and with different types of pain models: acute or chronic inflammatory and neuropathic pain (*i.e.* ultrasonic vocalisations were observed in acute arthritis via kaolin/carrageenan injection, adjuvant-induced arthritis and spared nerve injury models but not after ear notching and tail snipping, or in CFA induced arthritis, partial sciatic nerve ligation and neuropathic diabetic models). Some concluded that ultrasonic vocalisation cannot reliably assess pain in rodents and the differences observed could be related to protocol differences (Kurejova *et al.*, 2010). Some modifications were suggested, such as the acclimatisation to the recording chamber to reduce stress induced analgesia, exclusion of recordings if rodents were moving, the assessment of rodents individually and the careful adjustments of bat detectors while recording (*i.e.* minimise the number of bat detectors and reduce ultrasonic noises; Kurejova *et al.*, 2010). The recording at higher frequencies (at 50 and 37 kHz) has also been suggested, however, this has only been reported by one study so far (Blanchard *et al.*, 1991). The possibility of strain differences also needs to be accounted for because certain strains may be more vocal (Schwarting, 2018). Overall, the usage of ultrasonic vocalisations to assess pain in rodents requires further validation and standardisation of protocols before it can be used as a pain assessment tool in rodents.

1.1.2.4. What should be used to assess pain?

Traditionally, the most commonly accepted and favoured measurement of pain in animals has been nociceptive threshold or evoked testing because it is practical, highly reliable, repeatable and has also been reported in human patients (Mogil and Crager, 2004). It was reported that 90% of pain studies utilised evoked tests of either mechanical or thermal stimuli, and only 10% utilised behavioural measurements (Mogil and Crager, 2004). The use of Quantitative Sensory Testing (QST) has also been suggested. This consists of utilising multiple nociceptive tests in a standardised manner to assess the sensory profile of patients (*i.e.* loss or gain of function in response to different types of stimuli; Backonja *et al.*, 2013). Due to the standardised nature, this method is useful to quantify and observe changes in the patient's sensory profile over time, diagnose diseases and identify patients who are at risk of further sensory loss (Backonja *et al.*, 2013). The use of QST to identify abnormal mechanisms behind neuropathic diseases may improve predictions of therapeutic efficacy (Attal *et al.*, 2011). In a retrospective analysis of 902 QST profiles of chronic pain patients, it was found that patients could generally be grouped into one of three clusters: 1) sensory loss with loss of small and larger fiber function; 2) thermal hyperalgesia but with preserved small and large fiber sensory function; and 3) mechanical hyperalgesia with loss of thermal sensitive small fiber function (Baron *et al.,* 2017). It was also noted that certain analgesics were associated with higher efficacy in certain patient clusters (*e.g*. patients within the first cluster had higher efficacy with opioids but not oxcarbazepine). Additionally, all three types of sensory profiles were present in patients with the same type of neuropathic disease (Baron *et al*., 2017 and Vollert *et al*., 2017). These studies demonstrate that similar mechanisms may be at play across different neuropathic diseases, therefore, developing novel analgesics by understanding the mechanisms behind abnormal sensory profiles may yield more efficacious therapies. As a result, it was recently proposed that back translation of identified sensory profiles into animal models should be performed (Rice *et al*., 2018). This involves redeploying traditional nociceptive assessments into a sensory profiling protocol that reflects one used for humans, and the development of animal pain models that accurately reflect the identified sensory profiles in neuropathic patients.

While nociceptive assessment methods can be highly standardized, allow direct comparisons between animal and human studies and are useful for sensory profiling, they are limited to assessing nociception and not the assessment of ongoing "spontaneous" pain (Backonja *et al*., 2013). It has been reported that the greatest and most prominent complaint of pain in patients was "spontaneous pain" (the ongoing unpleasant sensation that is not evoked by an external stimulus; reported by 96% of patients) while increased mechanical or thermal hypersensitivity was reported significantly less frequently by patients, 64% and 38% respectively (Backonja and Stacey, 2004). Therefore, the overreliance on nociceptive evoked tests has been attributed as one of the major reasons for the lack of successful novel analgesic developments in translational pain research (Mogil and Crager *et al*., 2004; Rice *et al*., 2008 and Mogil *et al*., 2010).

To recap, pain is defined as 'an unpleasant sensory and emotional experience associated with actual or potential tissue damage or described in terms of such damage' (International Association for the Study of Pain, 1979). This definition breaks down the pain experience into two parts: the sensory and the emotional components. Nociceptive evoked tests only assess the sensory component of pain by assessing a response to an external stimulus. These tests are unable to assess the ongoing emotional experience of pain and are therefore, unable to fully capture the whole pain experience. Furthermore, it seems that the more relevant component of pain, particularly in humans and most likely in animals as well, would be the affective component rather than the sensory component assessed by nociceptive evoked testing. This results in a disproportionate number of pain studies that focused on sensory pain, which is of less clinical relevance. It has been suggested that nociceptive evoked testing may be utilised as a substitute measurement of ongoing pain experienced by animals. However, differences between ongoing pain (as assessed with the RGS) and mechanical allodynia (as assessed with von Frey filaments) has been demonstrated (De Rantere *et al*., 2016). Both measurements increase and peak similarly, but the duration of mechanical allodynia outlasted the ongoing pain. This phenomenon of mechanical allodynia outlasting the ongoing pain has also been corroborated by a human account (a researcher documented his experience after an accidental self-injection of CFA; Gould, 2000), thus suggesting that the two pain experiences are separate in humans as well. Therefore, the emotional component and the sensory component of pain are distinct mechanistically. Also, the emotional and sensory components of pain are processed in different parts of the central nervous system: ongoing pain is processed in the higher centers of the brain and the reflex withdrawal assessed in most nociceptive evoked testing is only a spinal response. Therefore, measures of nociceptive evoked testing cannot be used as a substitute for the assessment of ongoing pain and should not be used as the sole assessment of pain. The overreliance on nociceptive evoked testing is likely the reason some drugs (*e.g.* neurokinin-1 antagonists) which showed promise in animal trials failed in human clinical trials (Goldstein *et al*., 2001 and Pitcher and Henry, 2004). These analgesics were most likely effective in the modulation of the sensory aspect of pain (hypersensitivity) but not in ongoing pain (Mogil and Crager, 2004).

Consequently, critical evaluation of these pain assessment methods has led to a search for outcome measures that account for “spontaneous” pain. To address these issues, it was proposed that spontaneous and non-evoked behaviours should be assessed (*i.e.* behaviours that animals voluntarily display without an external stimulus; Mogil *et al*., 2010). The spontaneous behaviours utilised in pain assessments should have the following characteristics: be frequently observed, specific to pain, not related to comorbidities, consistent across a variety of pain models, repeatable across rodent strains and be sufficiently sensitive to detect changes in types and doses of analgesic medications. Finally, the results should correlate well with the results from nociceptive evoked tests (Mogil and Crager, 2004).

The use of spontaneous behaviours to assess pain may be difficult in rodents as they are prey species, and as such, will instinctively camouflage signs of distress or pain. Some behaviours have been proposed to directly assess pain in rodents, three such behavioural tools have been described above: the CPP, the CBS and the grimace scales. Each of these behavioural tools can be useful in pain assessments, however, each has their own shortcomings. The CPP test is useful for pain assessment as the unpleasant sensation of pain is naturally aversive and analgesics are unsurprisingly positively reinforcing. However, the CPP test assesses the animal's motivation to remain in one of two compartments. Pain or alleviation from pain can be a powerful motivator, this motivation can also be affected by confounding factors (*e.g.* positive reinforcement from certain drugs and activity levels). Additionally, animals are only able to pick between two options and demonstration of a dose dependent analgesic effect requires repeated trials with multiple animals. The CBS is also able to assess pain as the behaviours in the ethogram were observed to increase after a laparotomy surgery and decreased with analgesia. However, so far, the behaviours displayed are most likely specific to visceral pain as all behaviours are centered around the animal's abdomen. Therefore, this ethogram is of limited use for other types of pain. Lastly, the grimace scales have demonstrated utility across multiple types of pain models at least in the application of the MGS and RGS and may be useful for assessing ongoing pain. However, these scales are influenced by stress induced analgesia and changes in facial expressions may also occur in non-painful situations.

General measures of well-being may also be utilised to assess pain in animals as a decrease of well-being is generally observed in patients experiencing pain. Measures of burrowing, nest-building and grooming behaviours are ADL activities as under non-painful conditions, rodents are highly motivated to perform these behaviours. These behaviours are generally observed to decrease when animals are believed to be in pain and improve when analgesics are administered. However, unless latency to initiate burrowing or nest building activity is used as the assessment, assessment of animals can only be performed hours later and during that time animals may be in pain. Furthermore, latency to perform these behaviours are probably affected by exposure to anesthesia and varies between animals. These behaviours are also affected by the types of burrowing substrate and nesting materials and may vary between strains and species. Ultrasonic vocalisations may also be useful for pain assessments, however, ultrasonic vocalisations due to pain has not been a consistent finding in published studies. Therefore, unless ultrasonic vocalisation can be consistently observed from rodents while in pain, this method will not be useful for pain assessment.

Overall, there are a variety of pain assessment methods that are available for use as translational pain and animal welfare assessments. No single assessment method may be classified as the gold-standard. Instead, the most appropriate pain assessment method depends on the specific type of pain that one is trying to assess. For example, does a researcher want to assess the sensory or affective component of pain? If it is the sensory component of pain, does the researcher want to assess mechanical, thermal or cold allodynia? These may be assessed with the von Frey, hot plate and acetone tests respectively. If the researcher wishes to assess the affective component of pain, does the researcher wish to assess ongoing pain (pain from a constant stimulation by inflammatory stimuli), inflammatory pain (ongoing pain in addition to allodynia and hyperalgesia), spontaneous pain (internal pain that is self-contained and caused by changes to neurons) or summated pain (pain caused by the accumulating stimuli from activities of daily life; Bennett, 2012)? Ongoing, inflammatory and spontaneous pain may be assessed with the CPP, CBS, RGS and grooming behaviour while summated pain may be assessed with the CPP, burrowing, nest building and grooming behaviours. In addition to considering the appropriate pain assessment method for the desired pain type, one must also consider what is practical. Nociceptive evoked tests, the RGS and CBS may be assessed within a short duration of time (around 10 minutes) but requires personnel to be present during the entire assessment period. The time required by the personnel to assess each animal can quickly add up as each animal needs to be tested individually. On the other hand, burrowing, nest building, and grooming behaviour do not require personnel to be present during the entire duration, personnel can simply begin the process (*i.e.* provide burrowing tube, nesting materials or apply the fluorescent substance) and return at the appropriate times. Lastly, the CPP is the most time and labour intensive as personnel are required to train the animals and are required to be present during the entire assessment period.

To further illustrate this, consider if a researcher wants to assess the analgesic efficacy of a novel analgesic in a rat model of arthritis. They would be interested in the analgesic efficacy, duration of analgesic action, if mobility improves with analgesic administration and if there are any negative or abusive potentials of the analgesics. Analgesic efficacy and duration of analgesic action may be assessed with the RGS and von Frey test to assess ongoing pain and mechanical hypersensitivity respectively. Both assessments may be assessed rapidly and allow for repeated assessments of pain at set intervals after analgesic administration. These two measurements allow for a chronological assessment of analgesic efficacy and the maintenance of efficacy to reduce ongoing and hypersensitivity pain. These assessments may also be utilised to assess the dose required to achieve analgesia. Assessment for improvements in summated pain or mobility may be assessed with the burrowing test. Rats administered the novel analgesic would be expected to burrow more than rats that did not because rats that received the analgesic should retain their ability and motivation to burrow. Lastly, any negative or potentially abusive properties of the novel analgesic can be assessed with the CPP test. Animals with arthritis should prefer the compartment paired with the analgesic and non-arthritic animals should have no preference. However, if they avoid the drug-paired compartment, the researcher will know that the novel analgesic causes an aversive reaction.

Another example would be a veterinary technician or researcher who wants to ensure the welfare of mice after a laparotomy surgery. They would likely be interested in ensuring that the mice received proper analgesia and were able to perform normal activities after the surgery. However, they would have little time to monitor the animal throughout the day and therefore, would want to utilise assessments that can be performed quickly and allow for reassessment before the end of the day. Therefore, they could utilise the MGS, nest building and grooming behaviour. After surgery, they can quickly assess the animals with the MGS and assess the latency for the mice to initiate nest building activities. Analgesic interventions may be decided at this time or animals suspected as experiencing pain reassessed a few hours later. At the end of the day, they may also assess the willingness of the mice to perform nest building or grooming behaviour by examining the quality of the nests and spread of the fluorescent substance on the mouse. They can then decide if additional analgesia should be administered.

In conclusion, multiple assessment methods should be utilised to assess the different modalities of pain, be it hypersensitivity, ongoing pain, the unpleasant aversive component or the disincentive to perform normal activities. The types of assessment tools used should correspond with what is most relevant for the type of pain being assessed, for what reason it is being assessed and how practical it is to perform the assessment type.

1.1.2.5. What about biomarkers of pain?

It has also been proposed that cellular or genetic biological markers of pain will be useful to objectively assess pain and used as the target for novel analgesics (Mao, 2009). While the idea of biological markers as an objective measure is appealing, to date no such biomarker has been found that is sensitive and specific (Mogil, 2009; 2010). For example, four spinal neuropeptides (substance P [a tachykinin neuropeptide that acts as a neurotransmitter and neuromodulator], CGRP [a peptide vasodilator involved in nociceptive transmission], somatostatin [an inhibitory neuromodulator] and bradykinin [an inflammatory mediator]) were assessed in a commonly used MIA induced model of osteoarthritis (Otis *et al*., 2017). These neuropeptides are evidently involved in central sensitisation as they increased after induction of the pain model and the increase coincides with mechanical hypersensitivity. However, if these neuropeptides are directly related to pain, they should also display a dose dependent increase after the MIA injection and should decrease when analgesic is administered. In this study, only one neuropeptide, somatostatin, displayed a dose dependent increase, however, it did not decrease significantly when lidocaine, an analgesic, was administered. Alternatively, substance P and bradykinin decreased with lidocaine administration but did not increase in a dose dependent manner to the injected MIA dose. Interestingly, CGRP levels and mechanical hypersensitivity were higher in sham rats (rats administered saline injection instead of MIA) compared to naïve rats. It would have been interesting if the spinal neuropeptides assessed in this study had also been compared to non-evoked spontaneous behaviours that assess ongoing pain. For example, RGS scores and CGRP levels were observed to increase simultaneously after an experimental tooth movement model and both remained low when a CGRP antagonist was administered

(Long *et al*., 2015). Therefore, the currently identified biomarkers are evidently associated with pain and central sensitisation, however, these biomarkers on an individual basis do not increase or decrease in a dose dependent manner to stimulus intensity and analgesia respectively. As such, it cannot be stated that these biomarkers are sensitive to pain, nor may they be used to capture the entire pain experience.

Furthermore, the achievement of such a feat is complicated by the fact that multiple cellular and genetic mechanisms are likely involved in a single type of pain and patients tend to suffer from multiple types of pain at any given time (Hill, 2000; Mao, 2009 and Mogil, 2010). Comorbidities of pain and changes in neural plasticity during chronic pain conditions will likely affect the processing and the modulation of pain (Hill, 2000; Mao, 2009; Mogil, 2010). Neuroimaging via fMRI (function Magnetic Resonance Imaging) and PET (Positron Emission Tomography) have also been proposed as objective measures of pain as these methods have provided significant information on the processing of pain by the brain (Mao, 2009). However, these methods are unable to provide real-time information on pain and are difficult and costly to perform (Chizh *et al*., 2008 and Mao, 2009). Nonetheless, even if possible, biomarkers of pain still need to be validated and confirmed with spontaneous measurements of pain (Mogil, 2010)

1.1.3. The ethics of pain assessments and creation of pain scales

Before any research study can begin, researchers are ethically obligated to ensure that pain or distress is minimized or averted (Griffin *et al.*, 2014). However, when pain is the subject matter studied, this is difficult to achieve. Researchers need to induce measurable pain in order to study its effects and be able to quantify it. Therefore, pain researchers are ethically required to minimise the pain experienced by the animals in terms of intensity and duration. This includes the use of appropriate endpoints assessed with appropriate pain assessment methods and the use of appropriate analgesics which do not interfere with the study's objectives. Researchers should always apply the 3Rs when planning studies: 1) Replacement of the animal model whenever possible, 2) Reduction of the number of animals used and 3) Refinement of the study design to minimise pain and suffering.

The replacement of animal models for the study of pain has been proposed with the use of human neuroimaging with fMRIs and the study of human genes related to pain (Mogil *et al*., 2010). However, it must first be recognized that pain is a complex phenomenon that only occurs in a living conscious creature. Therefore, whole animal models are required to capture and assess the pain experience. Additionally, the use of animal models allows for the standardization of the subject (*i.e.* genetic and environmental factors). This reduces data variability as the human population is more genetically diverse with different lifestyles and potential comorbidities that could interfere with understanding complex pain mechanisms. As such, animal models are also more practical, economical and ethical (than experimenting on human subjects, Mogil *et al.*, 2010). Instead, of opting for the complete replacement of animal models, data from animal models should be used to complement human studies and vice versa.

The reduction of animal use can be accomplished by re-using animals that were previously used in another study if the previous study performed does not interfere with the objectives of the new study (*e.g.* use of animals scheduled for euthanasia in a study to assess pain during euthanasia) or the use of animals already undergoing a painful procedure (*e.g.* the assessment of the CBS in a bladder cancer model; Roughan *et al*., 2004). Furthermore, use of surplus animals that would otherwise be euthanized without being enrolled in a study is also a possibility. The performance of power calculations to estimate necessary sample size before a study ensures the optimal number of animals are included in the study to appropriately meet the study objectives. Lastly, when possible, the study design chosen can be carefully selected to minimise animal numbers (*e.g.* use of a factorial design with animals acting as their own controls).

Refinement of a pain study can be accomplished in many ways. Firstly, researchers can utilise routine husbandry procedures that are known to be painful as their pain model (*e.g*. the Horse Grimace Scale was developed with horses undergoing castration; Dalla Costa *et al.,* 2014) or utilise animals that are experiencing naturally occurring painful diseases (*e.g.* the sheep pain facial expression scale was developed with naturally occurring mastitis and foot rot, McLennan *et al*., 2016). Refinement of the study design is also possible with careful consideration of the type of pain model utilised in order to minimise the intensity and duration of the pain stimulus. For example, the RGS was created with acute pain models that lasted for less than a day (*i.e.* intra-plantar carrageenan/kaolin and CFA injection and the laparotomy model; Sotocinal *et al*., 2011). Another example was the choice to utilise a colitis model instead of a cancer model to assess the utility of the RGS in this thesis. The colitis model was favoured as it was not terminal, and the severity of the model could be controlled with the concentration of the stimulant. This is discussed further in chapter 1.3. Lastly, pain studies may be refined by the inclusion of more sensitive endpoints as well as the administration of analgesics during non-testing periods or immediately after testing.

To conclude, in order to minimise, prevent or understand pain that animals experience during research, pain in animals needs to be recognised and quantified. Without a validated pain assessment method, researchers are reliant upon non-specific, potentially insensitive, physiologic responses to pain (*e.g.* weight loss and food and water consumption). Therefore, research to create and validate pain assessment scales that are sensitive and specific to pain is necessary. These pain assessment methods may also aid in identifying efficacious analgesics, their optimal doses and duration of action for different species and strains of research animals. Additionally, use of validated pain scales can inform researchers of the duration of pain expected with painful procedures (*e.g.* pain from a laparotomy surgery in rats lasts 4-6 hours as assessed with the RGS and CBS; Roughan and Flecknell, 2004 and Sotocinal *et al*., 2004) thus allowing researchers to make informed choices on the requirement for analgesic intervention. Finally, validated pain assessment methods will more accurately quantify the experience of pain in animal models of human pain, increasing the success of developing novel analgesics and understanding of pain mechanisms.

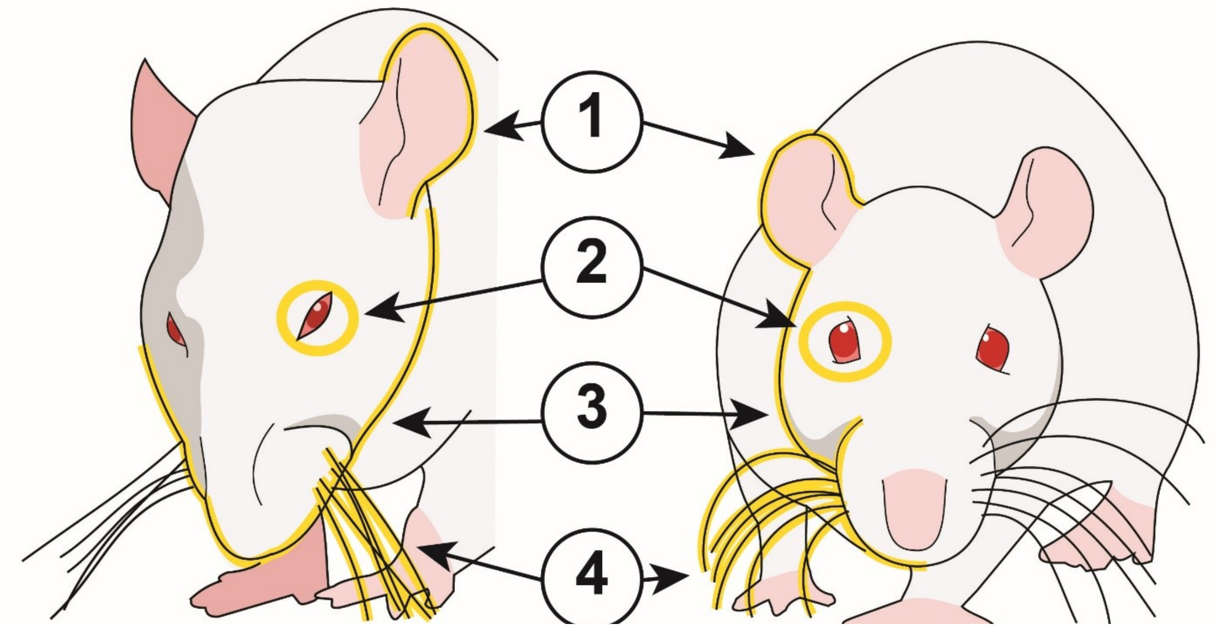
1.1.4. Other factors to consider during pain research

There are several other factors that should be considered as well, these include: ensuring the experimenters are properly trained to use the assessment tool. The practicality of an assessment tool will decide if it will be utilised by many researchers or if it is too cumbersome to be utilised. Proper training of experimenters is important to ensure accurate and reliable interpretation and collection of data. Lastly, accurate reporting of all parts of a study is important to ensure the legitimacy of a study's results by allowing study replication and critical evaluation of methods. These topics will be discussed further in the following chapters.

1.2. The Rat Grimace Scale

After the development of the RGS by Sotocinal *et al*. (2011; Fig. 1.2), the RGS has been utilised in many studies which can be broadly categorised into three types: 1) to refine laboratory rat welfare; 2) to assess different pain types and processes and 3) to improve the practicality and utility of the RGS. These studies have been summarised in Table 1.2.

Fig. 1.2: Cartoon of the Rat Grimace Scale.



Legend: The 'pain' rat (left) has 1) ears folded and angled outwards; 2) partial eye closure; 3) nose flattens and elongates and 4) whiskers bunched and directed away from the face. The 'no pain' rat (right) has 1) ear round and facing forward; 2) no eye squeeze; 3) nose rounded and slightly puffed out and 4) whiskers relaxed and droopy.

1.2.1. Applications of the RGS

1.2.1.1. RGS as a welfare refinement tool

Use of the RGS as a welfare refinement tool has manifested two ways, by: 1) reassessing efficacy of different analgesic protocols and 2) assessment and characterising of pain associated with different pain models. Analgesic efficacy assessments were undertaken by utilising the RGS with other pain assessment methods (Waite *et al*., 2015; Thomas *et al*., 2016; Philips *et al*., 2017 and Nunamaker *et al*., 2018) or by including the RGS in addition to physiological assessment methods (Korat and Kupupara *et al*., 2017; Gao *et al*., 2017; Jeger *et al*., 2017 and Chaves *et al*., 2018). These studies have also assessed the efficacy of singular or multi-modal analgesic protocols. For example, one study assessed the efficacy of various analgesics (buprenorphine, ketoprofen, acetaminophen, ibuprofen and carprofen) to treat pain after a laparotomy procedure (Waite *et al.*, 2015). These different types of analgesics were administered at various doses and at different times points (15 minutes before surgery or during surgery). It was found that certain analgesics were effective at lower doses if administered 15 minutes prior to a laparotomy surgery (0.01 mg/kg buprenorphine and 25 mg/kg ketoprofen) and higher doses were required (0.025 mg/kg buprenorphine and ketoprofen were ineffective at 5-25 mg/kg) if administered during the surgical procedure. Interestingly, the opposite was observed for acetaminophen whereby RGS scores only decreased significantly if a dose of 100 mg/kg was administered before surgery and 50 mg/kg was administered after surgery. It was also found that ibuprofen administered at 15 mg/kg was similarly effective if administered before or after surgery while carprofen at 5-25 mg/kg was ineffective no matter what time it was administered at. This study also concurrently assessed analgesic efficacy with a hot plate test. With the hot plate test, analgesic efficacy was demonstrated at similar doses (0.025 mg/kg buprenorphine and 50 mg/kg acetaminophen) while others required an increased dose (30 mg/kg ibuprofen). Therefore, this study demonstrated that previously derived analgesic efficacy doses that were assessed with nociceptive evoked testing methods may require re-assessment. A limitation of this study was that it only assessed analgesic efficacy at a single time point, it would have been interesting if the authors had also assessed the RGS 2.5-4.5h after laparotomy surgery as it has been demonstrated that RGS scores and the frequency of pain behaviours peak during this time (Roughan and Flecknell, 2004 and Sotocinal *et al*., 2011). Assessment at later time points would also have allowed for the assessment of the duration of action for each analgesic which can inform researchers if re-administration is required. Later studies have also assessed if buprenorphine, meloxicam, nalbuphine or morphine could effectively reduce RGS scores (Sotocinal *et al*., 2011; Jeger *et al*., 2017; Philips *et al*., 2017; Thomas *et al*., 2016; Nunamaker *et al*., 2018). These studies demonstrated that the use of a single analgesic can effectively reduce RGS scores, however, the RGS scores were usually elevated above baseline levels. This indicates that these analgesics are unable to completely abolish the pain experience and a multi-modal analgesic protocol may be required. A few studies have assessed the efficacy of multiple analgesics protocols with the RGS (Gao *et al*., 2017; Korat and Kupupara, 2017 and Chaves *et al*., 2018). However, none of these studies assessed if the use of the analgesics separately was also effective. This prevents readers from assessing if the administration of multiple analgesics was truly more effective than the administration of only one.

The RGS has also been used to assess pain in different procedures, such as the comparison of pain that may result after a laparoscopy procedure and laparotomy (Prefontaine et al., 2015). This study demonstrated that the less invasive laparoscopy procedure resulted in a lower RGS score when compared to a laparotomy procedure. This suggests that a laparoscopy procedure is less painful and should be considered over a laparotomy procedure whenever possible. The RGS was also used to characterise pain from a procedure to inject viral vectors into rats (Long *et al*., 2017). The RGS scores increased after the procedure which suggests that analgesic administration should accompany this procedure. Another endeavour to employ the RGS to refine laboratory rat welfare was the assessment for pain during euthanasia with intraperitoneal pentobarbital (Khoo *et al*., 2018). In this study, the RGS scores did not increase after pentobarbital injection but writhing behaviour did increase and was reduced if lidocaine or bupivacaine was also administered. It is unknown if the lack of an increase in RGS scores was truly due to a lack of pain or the sedative effects of the pentobarbital injection. The effectiveness of two methods of morphine administration (intrathecal and subcutaneous) were compared and pain was assessed with the application for the RGS after a laparotomy surgery (Thomas *et al*., 2016). However, the RGS data in this study was too variable to draw any conclusions and the authors themselves admitted that the sedative side effects of morphine may have been the cause.

Additionally, their RGS scores may have been confounded by the presence of an observer in the room who was taking pictures. The presence of an observer has been observed to affect the display of pain behaviours in mice and guinea pigs (Sorge *et al*., 2014 and Oliver *et al*., 2017).

In general, the RGS has been demonstrated to be a useful tool for the re-assessment of analgesic protocols routinely used to reduce the experience of pain in laboratory rats. It has also been used to characterise and compare the different pain levels that may be caused by different types of husbandry and experimental procedures. The information gained from these studies may inform researchers if certain analgesic protocols are effective and if the level of pain experienced by certain procedures require analgesic treatment. These improved analgesic protocols and ability to assess pain after various procedures will improve the ability of researchers to treat pain effectively in laboratory rats.

1.2.1.2. RGS to assess different pain types

The RGS was originally developed in acute inflammatory pain models (*i.e.* intra-plantar CFA and kaolin/carrageenan and laparotomy model; Sotocinal *et al*., 2011) and has since been used to characterise acute and chronic neuropathic pain (Akintola *et al*., 2017; Philips *et al*., 2017; Schneider *et al*., 2017), orofacial pain (experimental tooth movement, Liao *et al*., 2014; temporomandibular joint pain, Sperry et al., 2018) and several other painful models (*i.e*. muscle pain, Asgar *et al*., 2015; chronic migraine, Harris *et al*., 2017; intracerebral hemorrhage, Saine *et al*., 2016; sepsis, Jeger *et al*., 2017; traumatic brain injury, Studlack *et al*., 2018; Table 1.2). All these studies observed an increase in RGS scores after the painful stimulus was initiated, demonstrating that the RGS could be used to discriminate between painful and nonpainful animals in these models.

However, while the RGS seems to be an effective pain assessment method for many different pain types, there are conflicting reports on the effectiveness of grimace scales for neuropathic pain assessment. In the original MGS study, it was observed that MGS scores did not increase 14 days after chronic neuropathic pain models (*i.e.* spared nerve injury or chronic constriction injury; Langford *et al*., 2010). It was assumed that the grimace scales may be unusable for pain lasting for more than a day or for neuropathic pain, limiting their usefulness.

However, subsequent studies have demonstrated that the RGS and MGS scores increase after different acute and chronic neuropathic pain models (*i.e.* chronic constriction injury, cervical spinal cord injury or nerve root compression; Akintola *et al*., 2017; Philips *et al*., 2017; Schneider *et al*., 2017). In one such study, it was observed that RGS and MGS scores were elevated 10 to 27 days or 21 to 24 days respectively after a chronic constriction of the infraorbital nerve (Akintola *et al*., 2017). The authors attributed the different results from Langford *et al*. (2010) to the differences in the pain intensities induced, differing pain pathophysiology and alterations in brain pathways induced by the differing neuropathic pain models (Akintola *et al.,* 2017). These studies demonstrate that the RGS may not be useful in certain pain models and the continued evaluation of the applicability of the RGS is needed.

The RGS has also been assessed alongside many other pain assessment methods: evoked nociceptive tests (De Rantere *et al*., 2016; Waite *et al*., 2015; Kawano *et al*., 2016; Akintola *et al*., 2017; Philips *et al*., 2017; Schneider *et al*., 2017 and Fujita *et al.,* 2018; Khoo *et al*., 2018; Sperry *et al.*, 2018; Studlack *et al*., 2018), behavioural measurements (Prefontaine *et al*., 2015; Whittaker *et al*., 2016; Thomas *et al*., 2016 and Jeger *et al*., 2017; Nunamaker *et al*., 2018) and expressed levels of various pain biomarkers (Asgar *et al*., 2015; Long *et al*., 2015; Gao *et al*., 2016 and Long *et al*., 2017). When compared to nociceptive tests, it was found that the RGS scores tended to increase and peak at the same time (De Rantere *et al*., 2016; Kawano *et al*., 2016; Akintola *et al*., 2017 and Philips *et al*., 2017). However, mechanical hypersensitivity increased before RGS scores increased and remained elevated when RGS scores returned to baseline levels (De Rantere *et al*., 2016). Furthermore, results from nociceptive evoked tests and RGS scoring may conflict when the efficacy of novel analgesics are tested (*i.e*. allogregnanolone, a neurosteroid, reduces von Frey test scores but not RGS; Fujita *et al*., 2018). These studies demonstrate that RGS scores tend to increase with hypersensitivity, however, the presence of hypersensitivity does not necessitate an increase in RGS scores and therefore, the presence of ongoing pain. These studies also highlight that the different components of pain, nociception and ongoing pain (as assessed with the RGS), are fundamentally different and need to be assessed with specific pain assessment methods. When compared to other behavioural assessments, it has been demonstrated that the RGS may be confounded by factors unrelated to pain such as the sedative effects of morphine (Thomas *et al*., 2016 and Khoo *et al*., 2018). These studies have also demonstrated that the RGS may not be as suitable as other behavioural methods for pain assessment in certain pain models, such as an intestinal mucositis model (Whittaker *et al*., 2016 and Khoo *et al*., 2018). Therefore, more studies are needed to assess the possible confounding factors that will affect RGS scores and models that the RGS may not be applicable for.

The RGS was also compared to various molecular pain biomarkers where an observed increase in expressed levels of various biomarkers also resulted in an increase in RGS scores (Asgar *et al*., 2015; Long *et al*., 2015; Gao *et al*., 2016 and Long *et al*., 2017). For example, the direct relationship of CGRP to pain was observed when increased CGRP levels were accompanied by increased RGS scores (Long *et al*., 2015). In these rats, it was observed that tooth movement increased from day 1 to day 14, however, CGRP levels and RGS scores peaked on day 3 then decreased to baseline levels on day 14. The direct injection of CGRP into the periodontal tissues of rats already experiencing pain from the tooth movement model resulted in exacerbated RGS scores (Long *et al*., 2015). Furthermore, RGS scores decreased when rats were administered olcegepant, a CGRP antagonist (Long *et al.,* 2015). Taken together, it seems that CGRP levels and not tooth movement was associated with pain from a tooth movement model. The RGS has also demonstrated a relationship between ongoing pain with TRPA1 (Transient Receptor Potential Ankyrin 1) and acid sensing ion channel expression levels (Asgar *et al*., 2015 and Gao *et al*., 2016).

The RGS has also been utilised as a translational research tool to assess if certain analgesics may improve cognitive function in elderly patients (*i.e.* elderly rats were used as models of elderly patients; Chi *et al*., 2013 and Kawano *et al*., 2014; Saine *et al*., 2016 and Guo and Hu, 2017), to identify if novel analgesics are efficacious for post-operation pain (Fujita *et al*., 2018) and to compare two different medical techniques (*i.e*. nerve reconstruction via nerve lengthening or nerve autografting, Yousef *et al*., 2015).

It seems the RGS is a robust pain assessment method for the assessment of different types of pain. It has been compared with other pain assessment methods such as nociceptive threshold testing, other behavioural ethograms and molecular biomarkers. The observed differences of pain duration and the effectiveness of certain analgesics when assessed with both the RGS and a nociceptive evoked test demonstrate that these assessment methods measure different components of pain. Therefore, discernment and consideration of different pain assessment methods are required before the beginning of a new study. The use of the RGS in the study of pain mechanisms and molecular biomarkers is also highlighted with the concurrent increase of pain biomarkers and the RGS observed in some studies. Overall, the RGS seems to be a useful tool for pain assessment of different pain types and the study of pain mechanisms.

1.2.1.3. Refinement of the RGS as a research and welfare refinement tool

The RGS has proven to be a useful and a viable pain assessment method, has important implications in translational research and is an important welfare refinement tool in rodents. However, the standard method of collecting images for the RGS and other grimace scales is very labour intensive. A researcher must first video tape the animal for 10 minutes and then watch the videos to manually extract and crop images before any scoring can be accomplished (Sotocinal *et al.,* 2011). This causes a time delay of hours to days before a pain score may be generated. Therefore, the standard method of RGS scoring is only useful as a research tool and not appropriate for clinical application. Two refinements to the RGS and other grimace scales have been developed and proposed to reduce the labour intensiveness of these scales: The Rodent Face Finder (RFF) and the *a*MGS (automatic Mouse Grimace Scale, Sotocinal *et al*., 2011 and Tuttle *et al*., 2018). The RFF is a program that can automatically collect images of rodent faces by detection of the eyes and ears of rodents (Sotocinal *et al*., 2011). The program is then able to extract and crop images in preparation for scoring by a human rater. This program can also exclude images that may be blurred due to movement. However, if no image was captured for an interval, manual extraction by a researcher is still required. The *a*MGS is a computer program that can classify pictures of mice into ‘pain’ or ‘no pain’ categories (Tuttle *et al*., 2018). The *a*MGS has a high accuracy and reliability to human raters. The high accuracy of 94% is only possible if images of low confidence, usually images of intermediate MGS score, are removed. The authors justified the removal of these images by deeming the images of intermediate scores to be of limited clinical value. The removal of these images reduces the sensitivity of the *a*MGS, as intermediate MGS scores may indicate that the mice are experiencing low levels of pain and may require analgesic intervention in the near future. Given this, the *a*MGS cannot discriminate mice with low or moderate levels of pain or those demonstrating subtle signs of pain. Additionally, the current version of *a*MGS is only able to categorise images binarily as ‘pain’ or no ‘pain’ and as such, it cannot be used to monitor changes in pain intensity over time. Still, the RFF and the *a*MGS demonstrate the capabilities of technology to reduce the labour intensiveness of the grimace scales. Both advancements will allow large data sets of animals to be scored automatically. This reduces the labour intensiveness of the RGS and vastly improves it as a research tool, however, it will still be limited as a clinical tool because these tools cannot identify animals with low pain intensities and animals that require rescue analgesia.

Other developments related to the use of the RGS has been the development of an analgesic intervention threshold which may guide when analgesics should be administered (Oliver *et al*., 2014). There has also been the development of a ‘classifier’ to assess if individual AUs within a grimace scale should be given more weight (Dalla Costa *et al*., 2018). This ‘classifier’ may be useful for pain models or experimental conditions that confound the scores of individual AUs and should be excluded if they are not specific to pain. For example, the ears of horses, sheep and rats have been shown to change when the animals are experiencing negative emotional states like fear and/or aggression (Defensor *et al*., 2012; Boissy *et al*., 2014 and Dalla Costa *et al.,* 2017). Taken together, these developments in the grimace scales will allow researchers and clinicians to better utilise the RGS and other grimace scales to determine the pain intensity an animal is experiencing.

Lastly, studies have been performed to assess if there are confounding factors that may affect RGS scores. One study found that long and repeated exposures to isoflurane will increase RGS scores when rats are assessed 15 minutes after exposure to isoflurane (Miller *et al*., 2016). It has also been observed that restraint stress will inflate RGS scores (Senko *et al.,* 2016).

Overall, the RGS appears to be a robust tool for pain assessment in a research setting. Various developments have also increased the practicality of the RGS by reducing the manual labour required to score animals. However, it is still difficult to utilise the RGS in a clinical setting as animals cannot be assessed immediately. Real-time application of the RGS would allow researchers or animal technicians to observe the animal and decide if a humane endpoint has been reached. Furthermore, even when automated scoring is utilised, an observer will still be required to decide if the animal requires an analgesic intervention. Real-time application of the RGS has previously been proposed (Sotocinal *et al*., 2011 and Waite *et al*., 2015), however, it has not yet been assessed or applied in rats.

1.2.2. Real-time application of the RGS and potential challenges

Real-time application of the MGS was previously attempted in various strains of laboratory mice; however, no pain model was used (Miller and Leach, 2015). The aim of this study was to assess if different strains of mice displayed different grimace scores at baseline. The authors found that MGS scores from real-time scoring were lower than scores from the standard method. This suggests that real-time application of rodent grimace scales may not be possible as mice tend to display lower MGS scores in the presence of an observer (Sorge *et al*., 2014). However, it should also be noted that in this study, the observer only had three separate 5s windows to score the mice (Miller and Leach., 2015). These three 5s windows may have been too short of a duration to competently score the mice. Furthermore, an observer effect may have been present as the animals were only habituated to the observation chamber and not to the observer. Therefore, real-time application of the RGS may require a longer observation time and the animals may need to be habituated to the observer.

Other potential challenges associated with real-time scoring include the fact that scoring is no longer performed on static and tightly cropped images that only reveal the rat's face. With the standard method of scoring images, the observer can muse over each image and consult the RGS manual if required. The standard method also allows for the observer to be blinded to other factors that may influence scoring (*e.g.* body posture and visible tissue damage/trauma). In real-time, the observer will have to score quickly while the animal may be mobile and changing its facial expression. Also, the observer is privy to seeing the whole rat and the presence of other behaviours or signs of injury which may bias scoring (*e.g.* a rat with a heavily inflamed swollen foot may look more painful than a rat that has no foot swelling). Secondly, it is largely preferable to only score frontal images of rats as all AUs are clearly displayed. This is possible for image scoring as only one image is grabbed every three minutes and two cameras are placed on either side of the rat to maximise the chances of obtaining a good quality frontal image. However, during real-time scoring, the observer is only able to see the rat from a single angle and will not be able to obtain a score if the rat has its face turned away. Thirdly, during real-time scoring the observer will not be blinded to the time point. This increases the risk of bias if certain levels of pain are expected. Lastly, with real-time scoring, it is difficult to assess reliability. With the standard method of scoring, the images can be easily re-scored by the same observer or another person to check intra- and inter-rater reliability respectively. In real-time, the observer cannot go back in time to re-score the rat at that exact point in time. It is, however, possible to rescore videos playing at normal speed. Inter-rater reliability may also be potentially tested in the same way or two observers may enter the room simultaneously. However, the rat may react differently with two people in the room and it may be difficult to ensure a similar viewing angle with two observers.

Therefore, if possible, real-time application of the RGS will allow pain assessments to be performed quickly. This will increase the practicality of the RGS as a research tool and allows the RGS to be applied in a clinical setting. However, real-time application of the RGS has many potential challenges that must be considered.

Table 1.2: Summary of all RGS papers published to date

Pain model	Other assessments	Key findings	References
Studies that focused on the refinement of the RGS			
Intraplantar CFA Intraplantar carrageenan/kaolin Laparotomy	-	Aim: Develop the RGS and quantify pain in three pain models and develop the Rodent Face Finder • RGS developed with four action units (orbital tightening, nose cheek flattening, ear and whisker changes) • RGS scores increase after each model is induced and reduces in a dose dependent manner to morphine (1-5 mg/kg) • Development of the RFF which automatically captures images of rats from videos	Sotocinal *et al*., 2011
Implantation of telemetry	-	Aim: assess reliability of RGS and identify an analgesic intervention threshold • Intervention threshold identified at RGS score of 0.67 (sensitivity = 84.6%; specificity = 88.6%) • Very good intra-rater reliability possible after 6 months of disuse (ICC = 0.83)	Oliver *et al*., 2014
Exposure to isoflurane	-	Aim: assess if duration of exposure and repeated exposure to isoflurane influences RGS scores • Repeated exposure of short duration of isoflurane exposure (2 minutes	Miller *et al*., 2016

		duration, induction at 5%, 2.4L/min; maintenance at 2%, 2.4L/min) did not significantly increase RGS scores • Repeated and long exposure to isoflurane (12 minutes) increased RGS scores	
Restraint stress	Defecation during open field test Elevated plus maze (both assesses anxiety)	Aim: assess how prenatal exposure to elevated levels of angiotensin II (Ang II) influences the rats' emotionality • RGS scores were lower in Ang II animals compared to controls when in home cage • Both Ang II and control rats displayed increased RGS scores from restraint stress. The RGS scores during stress was similar between groups • Increase in RGS scores was greater in Ang II compared to control rats (interpreted as greater emotional reactivity) • Ang II rats were more likely to defecate and spend time in the closed arms of the elevated plus maze	Senko *et al.*, 2017
Studies that focused on the refinement of laboratory rat welfare			
Laparotomy	Hot plate assay	Aim: assess effective dose for various drugs administered before or during surgery • Drugs that decreased RGS scores when administered before surgery: buprenorphine (0.01 mg/kg), acetaminophen (100 mg/kg),	Waite *et al.,* 2015

		ibuprofen (15 mg/kg), ketoprofen (25 mg/kg). • Drugs that decreased RGS scores when administered during surgery: buprenorphine (0.025 mg/kg), acetaminophen (50 mg/kg), ibuprofen (15 mg/kg) • Carprofen (5-25 mg/kg) did not decrease RGS scores • Drugs that increased time spent on hot plate: buprenorphine (0.025 mg/kg), acetaminophen (50 mg/kg), ibuprofen (30 mg/kg). Ketoprofen (10-25 mg/kg) and carprofen (5-25 mg/kg) did not increase time spent on hot plate.	
Laparotomy Laparoscopy	Behaviour ethogram	Aim: compare pain from rats after a laparotomy or a laparoscopy procedure • Rats were more likely to have lower RGS' action unit scores and display other pain behaviours after a laparoscopy compared to a laparotomy procedure	Prefontaine *et al.*, 2015
Caudal laparotomy with bladder wall injection	Composite Pain Behavioural Scale Activity	Aim: assess if intrathecal administration of morphine (0.2 mg/kg) was more effective than subcutaneous administration (3 mg/kg) • RGS scores of rats administered saline or morphine were significantly different from baseline 1-8h after administration	Thomas *et al.*, 2016

		• Rat without pain displayed lower frequency of rearing, walking and climbing behaviours and were more inactive after intrathecal morphine administration • Up to 4h after surgery, RGS scores from rats administered intrathecal morphine remained similar to baseline; however, RGS scores of saline controlled rats were similar to baseline up to 8h. • After surgery, rats displayed a lower frequency of pain behaviours up to 8h with intrathecal and subcutaneous morphine administration • Authors concluded that side effects of morphine may have confounded RGS scoring and results in large variability of scores	
Laparotomy	Histopathology Tensile strength of wound	Aim: assess if a combination of intravenously administered analgesics administered during surgery reduced pain and improved healing • Intravenous administration of levobupivacaine (0.25% v/v), dexibuprofen (0.2 mg/mL), norepinephrine (0.1 mg/mL) decreases RGS scores after surgery • Analgesic combination if administered subcutaneously during surgery or a	Gao *et al*., 2017

		10x dose administered intravenously increases RGS scores • Rats that received the analgesic combination had better tensile strength and wound healing	
Faecal peritonitis (sepsis)	Heart rate Clinical scoring (behavioural ethogram)	Aim: assess behavioural and cardiovascular effects of nalbuphine in a sepsis model • RGS scores increase after sepsis model • Treatment with nalbuphine (1 mg/kg/h, intravenous) reduces RGS scores at 24h • The heart rate of septic and control animals was similar • Clinical scores did not decrease after 24h	Jeger *et al* 2017
Laparotomy	Tensile strength of wound Histopathology	Aim: assess if administration of an analgesic combination reduced pain after surgery and if tensile strength improves with healing • The combination administration of levobupivacaine (50 µL 0.3% w/v mg/ml), ibuprofen (2 mg/mL) and epinephrine (8 mg/mL) was effective at keeping RGS scores low • Rats that received the analgesic combination had better tensile strength and wound healing three weeks later	Korat and Kupupara 2017
Trigeminal injections to transduce viral vectors (orofacial pain)	CGRP expression level	Aim: assess the pain from a novel technique of delivering viral vectors to rat trigeminal ganglia	Long *et al.*, 2017

		• RGS scores and CGRP expression levels increase after vector injections • The trigeminal injection technique is painful with or without the viral vectors as evidenced by increased RGS scores • RGS scores decreased to baseline levels 7 days later	
Cervical radiculopathy/cervical nerve root compression	Von Frey	Aim: assess the effects of meloxicam after nerve root compression • RGS scores increase after dorsal root compression • Meloxicam (2 mg/kg) treatment decreases RGS scores • Mechanical hypersensitivity threshold decreases after dorsal root compression but increases if meloxicam is administered	Philips *et al.*, 2017
Acute spinal cord injury	Basso, Beattie, Bresnahan (BBB) scale – visual assessment of hindlimb movement	Aim: Assess the effect of tramadol on recovery • Recovery of hindlimb movement was linear with no differences between rats that received tramadol or saline • The RGS scores of rats that received tramadol remained consistently lower than rats treated with saline	Chaves *et al.*, 2018
Euthanasia via pentobarbital	Abdominal writhing Defecation Ultrasonic vocalisation	Aim: compare the effects of lidocaine and bupivacaine on pain during pentobarbital euthanasia	Khoo *et al.*, 2018

		• RGS scores did not increase after pentobarbital injection • Addition of lidocaine to pentobarbital reduced the duration of writhing and the number of feces produced • Rats administered lidocaine performed fewer writhing behaviours compared to bupivacaine	
Ovariohysterectomy	Weight loss Cage-side behaviours Exploratory activity Number of vertical rises	Aim: Evaluate the efficacy of post-operative treatment with buprenorphine and meloxicam • Weight loss could not differentiate between rats that received analgesics or saline • Cage-side behaviours and real-time RGS scoring differentiated between rats that received analgesic or saline • Observation of exploratory activity and number of vertical rises did not differentiate treatment effects	Nunamaker *et al.*, 2018
Studies that focused on the assessment of different pain processes and pain types			
Tooth movement	-	Aim: assess the pain from experimental tooth movement model • RGS scores increases after experimental tooth movement is induced • RGS scores are higher when a greater force is used • Treatment with morphine decreases RGS scores	Liao *et al.*, 2014

CFA induced masseter inflammation Craniofacial muscle pain	Electronic and manual von Frey RT-PCR	Aim: assess if TRPA1 is involved in increasing mechanical hypersensitivity and pain induced by ATP, NMDA and CFA • RGS scores increase after CFA induced masseter inflammation and reduces at 1h if AP18, a TRPA1 antagonist, is administered • Mechanical threshold decreases in ATP-, NMDA- and CFA-induced pain which increases when rats are administered AP18 • TRPA1 mRNA increases after CFA administration	Asgar *et al.*, 2015
Intraplantar CFA Intraplantar carrageenan Intraplantar incision	von Frey	Aim: assess relationship between RGS and von Frey • Mechanical hypersensitivity increases before RGS increase • RGS scores peaks at the same time as mechanical hypersensitivity increases • Mechanical hypersensitivity remained elevated long after RGS returned to baseline	De Rantere *et al.*, 2016
Tooth movement	CGRP expression levels	Aim: assess if periodontal CGRP contributes to pain during experimental tooth movement • RGS scores and CGRP expression levels increase and peaks on day 3 before decreasing to baseline levels on day 14 • When rats are treated with olcegepant, a CGRP antagonist, RGS	Long *et al.*, 2015

		scores and expression levels of CGRP reduced • RGS scores increases when CGRP was injected directly	
Tooth movement	Acid-sensing ion channel expression levels	Aim: assess the roles of acid-sensing ion channel 3 in an orofacial pain model • The amount of tooth movement increases daily (from day 1 to 14) • RGS scores and expression levels of acid-sensing ion channels increase and peak on day 3 before decreases to baseline levels on days 14 and 7 respectively	Gao *et al*., 2016
Chronic migraine	Light/dark box activity Time spent in light Distance travelled	Aim: assess the effect of repeated migraine episodes on photophobia, motor activity and RGS scores • RGS scores increase after each migraine episode • RGS scores of saline and vehicle controls remain low • Light/dark box activity was not discernable between migraine and vehicle control rats • Time spent in light and reduced activity was only evident after five daily migraine episodes	Harris *et al.,* 2017
Intra-plantar incisional model	Von Frey Single fiber recording	Aim: assess the effects and mechanism of endotoxin on post-operative pain • RGS scores increased after surgery and decreased when treated with	Kawano *et al*., 2016

		ketoprofen (15 mg/kg), morphine (0.5 mg/kg) and ropivacaine (300 µL at 0.2%). • RGS scores were higher in rats administered endotoxin with surgery and higher doses of analgesics were needed to reduce RGS scores (ketoprofen (30 mg/kg), morphine (1.5 mg/kg) • Mechanical hypersensitivity and activity of Aδ- and C-fiber increased similarly and over a similar time frame in rats treated with or without endotoxins	
Intestinal Mucositis via intraperitoneal chemotherapy	Behavioural ethogram assessment Enrichment interaction Social interaction test	Aim: characterise pain from an intestinal mucositis model with various behaviours • RGS scores did not increase at any time points • Frequency of back-arching, twitching, writhing behaviours increased from baseline levels • During spent sleeping was decreased from baseline levels • Treated rats gnawed on their Nyla bone enrichment more than control rats • Rats were more likely to spend time exploring, investigating, following, grooming after the pain model was induced	Whittaker *et al.*, 2016

Chronic constriction injury of infraorbital nerve	von Frey	Aim: assess if the RGS and MGS can assess ongoing chronic neuropathic pain • RGS scores increased post-CCI (10 to 27 days later) and reduced with treatment of 24μg/kg of fentanyl • MGS scores increased post-CCI (21 to 24 days later) • von Frey scores decrease post-CCI	Akintola *et al.,* 2017
Cervical spinal cord injury	Acetone test	Aim: characterise pain from cervical spinal cord injury and relationship between ongoing pain and paw withdrawal from acetone test • RGS increases in animals with spinal cord injury • RGS can be used as an indicator of supraspinal sensation	Schneider *et al*., 2017
Intra-plantar incisional pain model	-	Aim: assess the effects of dexmedetomidine on endotoxin-induced pain in an incisional model • RGS scores increased after surgery • Rats administered endotoxin after surgery had exacerbated RGS scores compared to saline animals • Administration of dexmedetomidine decreased endotoxin exacerbated post-operative pain • Attenuation of RGS scores by dexmedetomidine eliminated with atipamezole	Yamanaka *et al*., 2017
Studies that focused on the effects of analgesics and novel techniques for translational research			

Laparotomy	Radial maze Locomotion	Aim: assess impact of post-operative pain on cognitive function in aged rats • RGS increases with laparotomy but attenuates with administration of ropivacaine (300 µL at 0.2%) and morphine (0.8 mg/kg) • Rats that did not receive analgesia were more likely to make errors in the maze • Rats administered memantine, an NMDA-antagonist, performed better on memory tests even when RGS scores did not decrease • Locomotor activity is unaffected by all treatments	Chi *et al.*, 2013
Laparotomy	Open field test (to assess locomotion) 12-radial arm maze	Aim: assess effects of post-surgery administration of ketoprofen or morphine on cognitive function in aged rats • Ketoprofen (40 mg/kg) and morphine (0.8 mg/kg) were effective at keeping RGS scores low from 2-12h after surgery • Locomotion activity was unchanged in rats where analgesia was withheld or administered • Rats performed fewer errors in the maze when they were provided analgesia	Kawano *et al.*, (2014)
Lengthening of the sciatic nerve with novel device	Electrophysiology	Aim: assess effectiveness of a novel device for reconstructing sciatic nerves compared to	Yousef *et al.*, 2015

	Sciatic nerve index (footprint assessment to assess functional recovery) histology	nerve autografting and to assess the pain involved • RGS does not increase after nerve lengthening and remains low days after • Footprints and histology of rats indicated that rats treated with novel device recovered better than rats treated with the nerve autografting method	
Intracerebral hemorrhage (Stereotaxic surgery)	Rotarod test	Aim: assess pain and motor behaviours with different doses of fentanyl in this model • RGS scores increase on day 1 after intracerebral hemorrhage model is induced and reduces in a time dependent manner to baseline levels day 6 • Treatment with 10 µg/kg fentanyl significantly decreased RGS scores • Rotarod test was not an appropriate test as it was physically too demanding on the rats	Saine *et al*., 2016
Laparotomy	Performance error in radial arm maze locomotion	Aim: assess the effects of thalidomide, an anti-inflammatory analgesic, in aged rats • Administration of thalidomide (20 and 50 mg/kg) attenuated RGS scores and reduce performance errors in a dose dependent manner • Locomotion was unaffected	Guo and Hu, 2017

Intra-plantar incisional pain model	Weight bearing von Frey test	Aim: assess the effects of allopregnanolone, a neurosteroid, in an incisional pain model • Administration of allopregnanolone, a neurosteroid, attenuated mechanical threshold but not RGS and weight bearing scores	Fujita *et al.*, 2018
Temporomandibular joint pain (through daily jaw loading)	von Frey test Mankin scoring (assessment of cartilage degradation)	Aim: assess if the RGS can detect pain in this model • RGS scores increased during loading days but decreased when procedure stopped • Mechanical threshold reduced continuously during loading days and remained reduced days after procedure stopped • Cartilage degradation scores increased even when procedure was stopped	Sperry *et al.*, 2018
Traumatic brain injury	Beam walk Accelerating rotarod Open field Elevated plus maze Light-dark box von Frey test Immunohistochemistry Neuroimaging	Aim: assess if rats with this injury will display pain and anxiety behaviour • Experimental group displayed poorer balance compared to control group but did not display more anxiety behaviour • Experimental group displayed higher RGS scores and lower mechanical threshold than control group • Injury was likely caused by microglia-mediated inflammation	Studlack *et al.*, 2018

Legend: Table summarizing all RGS papers published to date. Listed are the pain models and other assessments performed concurrently with the RGS. Included is also the aim of the study as well as a summary of its main findings.

1.3. Utilising the RGS in an untested pain model

The RGS has been used to assess many different types of pain (refer to previous chapter 1.2), however, there are still many questions that remain unanswered in the application of the RGS. One such question is the applicability of the RGS to detect different types of pain such as acute and chronic visceral pain where the injury and pain is not external and cannot be directly stimulated. Therefore, if the RGS demonstrates its utility in the assessment of pain in such models, it will further establish the RGS as a relevant tool for translational pain research in these disease models and for laboratory rat welfare. Therefore, the criteria for selecting a novel untested pain model were as follows:

1. Clinical relevance: the model should be relevant to the human disease it is supposed to replicate
2. Well-recognised and frequently used: a model frequently used by researchers would be ideal as use of the RGS will be more quickly accepted and used if the model is relevant to the researchers
3. Acute: the model should have a short duration of onset and resolution for practicality and to ensure animals do not suffer for long periods of time
4. Chronicity: many human diseases are chronic; therefore, the model should also be chronic in nature
5. Assessment tools already available: the model should have an already established and validated assessment method that can assess pain or disease severity so that the RGS can be compared to it
6. No procedure-related mortality: the model should ideally not be terminal, nor should it cause too much pain
7. Practical: the model should be easy to perform and, therefore, easily reproducible
8. Visceral: visceral pain is difficult to assess as the pain and damage is internal and cannot be stimulated externally

1.3.1. Cancer type pain

Cancer is a painful disease and the pain increases as the disease becomes terminal resulting in a drastic decrease in the patient's quality of life (Pacharinsak and Beitz, 2008 and Currie *et al*., 2013). Cancer pain has been generalised into three types: 1) spontaneous ongoing pain, 2) evoked or weight bearing pain and 3) breakthrough pain (sudden rapid onset of extreme pain; Pacharinsak and Beitz, 2008 and Currie *et al*., 2013). The pain experienced depends on the cancer type and tumour location (Pacharinsak and Beitz, 2008).

1.3.1.1. Bladder cancer pain

As a disease, bladder cancer is the most common cancer type in the urogenital system and ranks 7th and 17th respectively for males and females worldwide as the most common type of cancer (Zhang *et al*., 2015). The quickest and most common way to induce a bladder cancer model is to implant cancer cells either via injection of cancer cells into the targeted organ through a small incision over the bladder (Ibrahiem *et al*., 1963; Roughan *et al*., 2004) or by introducing cancer cells into the bladder through a catheter after an acid wash (Xiao *et al*., 1999; Roughan *et al*., 2014). The success of the cancer model via injection of cancer cells directly into the bladder can be as high as 100% (Ibrahiem *et al*., 1963; Roughan *et al*., 2004 and Zhang *et al*., 2015) while introduction of cancer cells through the urethra with a catheter has an inoculation success rate of around 75% (Roughan *et al*., 2015). Other methods include exposure to carcinogens or a genetically engineered animal model, however, the success rates in these models are more variable and require more time to develop clinical signs (Zhang *et al*., 2015).

An orthotopic bladder cancer model would potentially be a good follow up to Roughan *et al*. (2004)'s study. The authors attempted to apply their previously validated CBS method (Roughan and Flecknell, 2004) to a bladder cancer model in rats. However, they were unsuccessful as none of the behaviours previously identified (twitching, loss of balance, back arching, writhing) reduced significantly when rats were treated with analgesics (meloxicam 2 mg/kg, s.c. or carprofen 5 mg/kg, s.c.). These analgesics were previously observed to be successful in mitigating pain after a laparotomy surgery model (Roughan and Flecknell, 2003). The authors attributed this failure to the rats' inactivity, which was potentially caused by the severe pain the rats experienced from the model. There have been no other papers describing behavioural or nociceptive tests to assess pain in the bladder cancer model in rats. However, a conditioned place paradigm, behavioural (*i.e.* locomotion, rearing, active behaviours, grooming, and resting) and nociception test (Hargreaves) were applied in a mouse bladder cancer model (Roughan *et al.,* 2014). As the mice approached the day of euthanasia (euthanized when presence of palpable tumour >15% of body weight with haematuria detected), they displayed an increasing preference for the morphine (2 mg/kg, s.c.) paired chamber. This chamber preference was correlated with tumour burden post-mortem as well as decreased locomotion, activity levels, rearing behaviours and hyperalgesia compared to control mice. One other method to assess disease progress was to track tumour development with physical examination (palpation) or transurethral cystoscope, hematuria formation, imaging systems such as ultrasound and magnetic resonance imaging system (MRI; Roughan *et al*., 2004; Satoh *et al*., 2007 and Zhang *et al*., 2015). However, once tumours were detectable, the animals usually had to be euthanized as they were in the terminal stages of the disease (Roughan *et al*., 2014). Overall, this cancer model appears to induce a high intensity of pain in animals.

There are no assessment tools that can track the development of this type of cancer until the animal develops palpable tumours and haematuria (Roughan *et al*., 2004; 2014); however, once these symptoms appear, the authors report that animals rapidly decline past their humane endpoints as observed from clinical and behavioural signs and need to be euthanized (Pacharinsak and Beitz, 2008; Roughan *et al*., 2014). Furthermore, there seems to be a large inter-animal variability in the development of the tumours, for example, in the above-mentioned study (Roughan *et al*., 2004) the time it took for the tumour to be palpable ranged from 13-21 days, haematuria development ranged from 9-21 days and the total time animals were in the study ranged from 29-43 days. Due to the significant amount of variation observed, the authors had difficulty in quantifying and reporting their data. They eventually settled on comparing at two time points: five and fourteen days before euthanasia (Roughan *et al*., 2004; 2014).

1.3.1.2. Bone cancer pain

Another type of cancer pain that might be interesting to assess with the RGS is bone cancer pain. Pain associated with bone cancers is the most prevalent type of cancer pain. It tends to present when there is a secondary cancer site in 70% of patients with terminal breast or prostate cancer (Currie *et al.,* 2013). The first model of bone cancer pain was developed by Medhurst *et al*. (2002)

via intra-tibia injections of $3x10^3$ or $3x10^4$ MRMT-1 rat mammary gland cancer cells. Various other models of bone cancer pain have since been developed with different cancer cells however all have been created with a similar protocol of injecting cancer cells into the tibia or femur (Dore-Savard *et al*., 2010 and Remeniuk *et al*., 2015). This model appears to induce pain in rats as demonstrated in a CPP test when the affected rats preferred the chamber paired with an analgesic (lidocaine or morphine). Bone damage from this model may be assessed with radiographs and MRI (Medhurst *et al*., 2002 and Dore-Savard *et al*., 2010). In general, this type of pain model induces progressive damage to the bone by day 10-15 and bone integrity is compromised at day 20-21. At which time, humane euthanasia of the animal is required (Medhurst *et al*., 2002 and Dore-Savard *et al*., 2010).

The degree of damage to the bone correlates well with measurements of hypersensitivity. In one study, comparisons were performed between the rat's mechanical withdrawal threshold and bone damage as assessed by bone remodeling by MRI (percentage of bone volume/tissue volume; Dore-Savard *et al*., 2010). In this study, lower mechanical threshold (assessed with von Frey on the affected hind paw), weight bearing and surface area of affected hind paw in contact with the ground was significantly different from controls or contralateral hind paws starting on day 14 and the decreased mechanical threshold correlated with the degrees of bone remodeling. The affected hind paw had a progressively lower mechanical threshold as the days progressed from days 14 (36g), 18 (30.9g) and 21 (23g) while control animals demonstrated a steady mechanical threshold (~42g). The animals' weight bearing on the hind paws was also affected as the affected hind paw could only bear 15-20% of their body weight, while the hind paw of control animals could bear 40% of their body weight from days 14-21. The surface area of the affected hind paw in contact with the ground also decreased (~40-50 mm^2) compared to the contralateral hind paw (90-100mm^2). These measurements correlated well with bone degradation, whereby rats with increased bone destruction had a reduced mechanical threshold and decreased weight bearing compared to control animals (Medhurst *et al.*, 2002). Activity (characterised by running wheel revolutions) also decreased compared to control animals (days 1-5 after inoculation: MRMT-1: 15000; control: 17500; days 6-10: MRMT-1: 17500; controls: 30000; days 11-15: MRMT-1: 27500; control: 37500). Results from these studies support the notion that bone cancer models are painful, as indicated by the increased hypersensitivity and decreased activity levels as the disease progressed.

It was concerning that the administration of morphine (at 3 mg/kg), did not increase mechanical threshold (Medhurts *et al*., 2002 and Dore-Savard *et al*., 2010). It was only at a higher dose of 10 mg/kg of morphine did the mechanical threshold increase (Medhurst *et al*., 2002). However, as the authors themselves commented, their data may have been confounded by the sedative effects from a high dose of morphine that caused the rats to be nonresponsive to stimuli. It was concluded that bone cancer models may result in such severe pain that morphine was ineffective. However, it may also be possible that opioids are ineffective for this type of pain and other analgesics may be more efficacious. It may also be possible that the presence of the tumour in the hind limb resulted in lameness but did not induce any pain and therefore, the administration of morphine had no effect. Interestingly, the body weights and body temperatures were unaffected as the disease progressed (Medhurst *et al*., 2002).

1.3.2. Inflammatory bowel disease

Inflammatory bowel diseases (IBD) are diseases characterised by chronic inflammation in the gastrointestinal tract of unknown origin but are presumed to be affected by environmental and genetic factors (Nyuyki and Pittman, 2015; Gasparetto and Guariso, 2013). IBD is estimated to affect 1 out of every 1000 people in Western countries and the number of people affected has been rising for the past 50 years, it will become a prominent global disease in the future (Gasparetto and Guariso, 2013). There are two main types of IBD: 1) Ulcerative Colitis and 2) Crohn's Disease; the two types are differentiated by the location of the inflammation. Inflammation in ulcerative colitis is confined to the rectum and colon while Crohn's disease consists of patchy inflammation along the whole gastrointestinal tract, from mouth to colon. Common symptoms of both types of IBD are abdominal pain, diarrhea and rectal bleeding (Shi *et al*., 2011). IBD is a vastly complicated disease believed to alter the central nervous system permanently and has implications in neurological (*i.e.* seizure disorders, cerebrovascular accidents) and behavioural issues (*i.e.* depression; Nyuyki and Pittman, 2015). For example, a study found that people with IBD had a higher probability of suffering from depression and anxiety (Kurina *et al*., 2001). The relationship between these intestinal inflammatory diseases and other systemic issues is unclear and remains speculative (Zois *et al*., 2010).

Animal models are commonly used to assess the complex inflammatory processes of the gastrointestinal tract during colitis and to assess novel medications to treat the disease. Colitis models are commonly created by two methods 1) an intra-rectal injection of the chemical 2, 4, 6-Trinitrobenzenesulfonic acid (TNBS) and 2) supplying water infused with dextran sulfate sodium (DSS) ad libitum (Okayasu *et al*., 1990 and Randhawa *et al*., 2014). TNBS- and DSS- colitis models are among two of the most popular methods of inducing colitis and are well documented. Mortality will occur unless the dose, and consequently the disease severity, are carefully controlled. Common methods to assess disease severity of these colitis models include the Disease Activity Index (DAI; measurement of: weight loss, stool consistency, occult blood or gross bleeding), histology, macroscopic assessments and myeloperoxidase activity (MPO: an expression of neutrophil activity; Zhou *et al*., 2008). However, other than the DAI scoring method, all other methods require animals to be sacrificed at various time points to assess the severity of the colitis models. Endoscopy or MRI have been suggested as alternate assessment methods as animals do not need to be sacrificed and the animals can be used as their own controls (Pohlman *et al*., 2009 and Brenna *et al*., 2013).

Visceral hypersensitivity is commonly assessed with colorectal distension (CRD) followed by assessment of the visceromotor reflex (VMR) or abdominal withdrawal reflex (AWR). Colorectal distention is performed via insertion of a balloon into the anus of a rat while it is anesthetised (Ness and Gebhart, 1987). VMR assessment is then performed in awake rats and this consists of monitoring abdominal muscle contractions measured via electrodes implanted into the peritoneal cavity (Ness and Gebhart, 1987 and Larauche *et al*., 2012). The AWR is a refinement of the VMR method. It simply assesses when the abdominal muscles of the rat contracts; it foregoes the need for surgical implantation of electrodes which may sensitise the rat to pain and confound the results (Al-Chaer *et al*., 2000 and Yang *et al*., 2006). Another assessment frequently used to assess IBD pain is ‘referred hypersensitivity’. It is believed that pain originating from the viscera is often referred to other parts of the body due to an alteration within the central nervous system where there is an overlap between the visceral and somatic pathways (Farrell *et al*., 2014). The use of von Frey filaments on the hind paws of animals has consistently shown decreased withdrawal thresholds in these colitis models (Millecamps *et al*., 2004; Zhou *et al*., 2008 and Farrell *et al*., 2014). Other indirect forms of measuring pain have also been employed. For example, in one study, rats with colitis displayed decreased attention levels (time exploring an unfamiliar object) compared to normal rats (Millecamps *et al*., 2004). Rats were able to demonstrate similar attention levels to normal non-colitis animals when administered an adequate dose of morphine (1 mg/kg; s.c.). Another study was also able to demonstrate that rats with colitis displayed poorer discriminatory learning which was once again able to recover with the administered morphine (1 mg/kg, s.c.; Messaoudi *et al*., 1999).

1.3.3. Comparing pain models: bladder cancer, bone cancer and colitis

To recap, the criteria for a novel pain model were: 1) clinical relevance, 2) well-recognised and frequently used, 3) acute, 4) chronicity, 5) assessment tools already available, 6) no procedure-related mortality, 7) practical and 8) visceral. Each of these criteria are discussed below.

1 and 2) Clinical relevance and well-recognised and frequently used: All three disease models are relevant and mimic debilitating diseases plaguing human society. These disease models are also well-recognised and frequently studied.

3 and 4) Acute and chronic: Between all the mentioned disease models, bone cancer and colitis models seem to have a well-defined predictable onset and progression of disease (20 days and 7-14 days respectively) while a larger variability seems to be involved in bladder cancer (around one month). The natural cycle of remission and relapse of colitis can be mimicked by giving the colitis inducing agents on and off. Therefore, both the acute and chronic phases of the disease may be induced with the application of this model. However, the onset and progression of the bladder cancer model appears to be more variable with large individual variability.

5) Assessment tools already available: Assessment of pain related to bone cancer and colitis models predominantly rely on hypersensitivity or referred hypersensitivity as well as non-specific pain behaviours (CPP, discriminatory learning and attentional levels) which are less practical and more labour intensive because these assessment methods require multiple days/sessions of training and habituation for the animals involved. Both bone cancer and colitis have detectable symptoms that correlate well to measurements of hypersensitivity (*i.e.* bone degeneration and DAI or macroscopic assessment via endoscopy). The bladder cancer model does not have a reliable method to assess pain or disease severity.

6) No procedure related mortality: Bone and bladder cancer models seem to involve severe pain as evidenced from the morphine's ineffectiveness to raise mechanical threshold in the bone cancer models and the appearance of symptoms indicative of terminal bladder cancer (palpable tumour and haematuria). Furthermore, the damage from bone and bladder cancer were reported to be terminal with no recovery possible and euthanasia is the only available humane endpoint. With the colitis model, if the dose is carefully chosen the damage is transient, the animals can recover, and procedure related mortality is preventable.

7) Practicality: The colitis models appear to be the simplest to reproduce as it only requires the addition of DSS into the rats' drinking water or an intra-rectal injection of TNBS to create a colitis model. For bladder cancer, it is possible to create the disease via an intra-urethra injection of cancer cells, however the successful inoculation rate is only 75% and this can only be performed in female rats. A 100% inoculation rate is possible, but surgery is required to inject cancer cells into the bladder of the rat. A surgery may confound the study as side effects from surgery unrelated to the model may affect the results. In bone cancer, surgery is also required to perform the intra-tibia or –femur injection and as mentioned before, the pain detected may be confounded by the pain from the surgery and not only from the disease model itself.

8) Visceral: Of all the three models, only bone cancer is not visceral in nature however, it can still be considered internal as the damage is inside the bone. Visceral was chosen as one of the criteria as it is usually difficult to detect visceral pain. Colitis is unique as one of its symptoms are referred hypersensitivity whereby hypersensitivity can be measured via mechanical hypersensitivity on the rat's back or feet.

Therefore, all three models are clinically relevant, well recognised, frequently used and all models currently rely mainly on nociceptive evoked testing as a pain assessment method. The bladder cancer model is less reliable with its larger variability in onset and it is difficult to detect the presence of the disease until it has reached its terminal stages. This means that all animals must be euthanized soon after detection of disease as there is the potential for severe pain. Bone cancer has a more predictable onset and progression, but its pain severity, the seeming inability to counter the pain with high doses of morphine and the irreversible nature makes it undesirable. Therefore, a colitis model would be the ideal model to use as it can be both acute and chronic and it is recoverable with time and reversible with analgesic.

1.3.4. Comparing colitis models: TNBS- and DSS-colitis

The two favoured methods to induce colitis in rats are via intra-rectal injection of TNBS or providing DSS in water *ad libitum*. The TNBS and DSS models are morphologically similar to Crohn's disease (inflammation with infiltration at the submucosal layer) and ulcerative colitis (inflammation with infiltration focused at the mucosal layer), respectively (Jurjus *et al.*, 2004 and Shi *et al.*, 2011). The TNBS model of colitis is initiated by first fasting the rats overnight or up to 24 hours before intra-rectal injection of TNBS under general anesthesia with a gavage needle or a catheter (Messaoudi *et al.*, 1999; Rawandhawa *et al.*, 2014). The DSS model is induced by providing DSS (2-5%) in the rat's drinking water *ad libitum* for 4-9 days (Randhawa *et al.*, 2014).

Colitis in humans is a chronic disease and usually consists of periods of remission (no active ongoing inflammation) and relapse (inflammation returns; Rawandhawa *et al.*, 2014). Therefore, it is appropriate that the equivalent animal models can mimic the remission and relapse phases. Both TNBS and DSS models of colitis are usually utilised as an acute model, where animals are exposed once to the inflammatory agent to induce a transient inflammation which fades a few days later. However, both model types may create a chronic colitis model by administering TNBS or DSS solutions via their appropriate routes repeatedly.

1.3.4.1. Chronic TNBS-colitis

A TNBS chronic colitis model can be created by giving rats TNBS intra-rectally or via a catheter placed intra-rectally followed by another one a few weeks later (Palmen *et al.*, 1995 and Gambero *et al.*, 2007). A chronic TNBS-colitis model is characterised by inflammation with high MPO activity and ulceration in two or more sites seven days after initial TNBS injection, this is followed by a slow healing period and a slight inflammation before the second intra-rectal injection. After the second intra-rectal injection of TNBS, the inflammation and ulceration reappeared. These chronic TNBS-colitis studies did not assess any pain or hypersensitivity assessments (Palmen *et al.*, 1995 and Gambero *et al.*, 2007).

1.3.4.1.1. Acute TNBS-colitis

Acute TNBS-colitis is created by a single intra-rectal exposure to TNBS. Studies have reported that rats display increased visceral hypersensitivity and somatic ‘referred’ hypersensitivity. Visceral hypersensitivity has been assessed in male Sprague Dawley rats by CRD and observing if the rat’s testicles, tail or abdominal muscles contract when 15-20 mmHg of pressure was applied intra-rectally (Zhou *et al*., 2008). Visceral hypersensitivity was observed 2, 7, 14, 21 and 28 days after exposure to TNBS. In comparison, control rats only respond when 50 mmHg of pressure was applied. Visceral hypersensitivity can also be assessed by assessing VMR during CRD. In studies with male Lewis and female Wistar rats, rats with TNBS-colitis had a higher VMR response compared to control rats when a 60-mmHg pressure was applied (Adam *et al*., 2006 and Deiteren *et al*., 2014). The elevated visceral hypersensitivity resolved on days 10 and 28 in male Lewis and female Wistar rats respectively. Interestingly, visceral hypersensitivity reoccurred in the male Lewis rats on day 28 and 31 with re-elevated VMR responses to CRD (Adam *et al*., 2006). This suggests a strain and sex effect on the inflammatory response in the TNBS-colitis model.

Referred hypersensitivity to the hind paws has also been assessed with von Frey and Hargrave’s method (Zhou *et al*., 2008). A lower mechanical threshold and thermal sensitivity was observed 14 days after TNBS exposure. TNBS-colitis rats had a reduced mechanical threshold (paw withdrawal at 10g) and a reduced latency to withdraw their hind paw from a hot plate (latency of 8s). Comparatively, control rats only displayed a withdrawal response to 50g and their average latency to react while on a hot plate was 15s. A reduced threshold to thermal stimulus was also apparent with the tail flick test (Zhou *et al*., 2008). TNBS-colitis rats had a shorter latency to withdraw their tail from a heat source compared to controls (TNBS-colitis: 2s; controls: 6s). Non-specific pain behaviours (repeated licking of abdomen, testicles and hind paws and hunched posture) occurred more frequently in colitis rats compared to control rats after TNBS injection. These behaviours subsided 5-7 days after TNBS injection. This seems to agree with De Rantere *et al*. (2016)’s study where hypersensitivity lasted well after spontaneous behaviour of pain (as assessed with the RGS) subsided.

1.3.4.1.2. TNBS-colitis and other non-specific indicators of pain

Human chronic pain patients are observed to have reduced attention levels which recover when they are administered an effective dose of morphine, suggesting that a reduced attention level may be pain related (Lorenz *et al*., 1997). Attention levels as a non-specific indicator of pain can also be assessed in rats by assessing if a rat notices a novel object or by assessing their discriminatory learning ability (Messoudi *et al.,* 1999 and Millecamps *et al*., 2003). To assess a rat's attention level to a novel object, rats are first allowed to explore an enclosure with four distinct objects for three days once daily (Millecamps *et al*., 2003). On the testing day, one of the objects was replaced with a novel unfamiliar object and control rats usually spent most of their time exploring the novel object. When rats with TNBS-colitis were put through this test, the rats did not spend more time exploring the novel object. This suggests that TNBS-colitis induced pain caused attentional or discriminatory deficits. Interestingly, TNBS-colitis rats that received an adequate dose of analgesic (1 mg/kg s.c. morphine; 10µg/rat intrathecal morphine; 200 and 400 mg/rat *p.o.* acetaminophen), spent more time exploring the novel object compared to their conspecifics that did not. This suggests that relief from TNBS-colitis induced pain improved the rats' attentional or discriminatory deficits. Discriminatory learning can also be used as an indicator of pain where rats must discriminate between two levers (active and inactive) that granted them 30s of darkness (Messoudi *et al.,* 1999). It was found that male Sprague Dawley rats with TNBS-colitis were less able to discriminate between the two levers compared to control rats or rats treated with morphine (1 mg/kg s.c.). This ability to discriminate was abolished when rats administered morphine were also administered naloxone (0.1 mg/kg), a morphine antagonist. Both groups of TNBS-colitis rats (treated with saline or morphine) had similar microscopic damage from exposure to the TNBS treatment and this suggests that the rats' inability to discriminate between the two levers were due to pain and not the inflammatory damage from the model.

1.3.4.1.3. TNBS-colitis and other measurements

The TNBS-colitis model is also characterised by weight loss, reduced colon length, increased colon weight and increased DAI (Adam *et al*., 2006; Deiteren *et al*., 2011; Brenna *et al*., 2013 and Sun *et al*., 2013). Weight loss was observed three days after TNBS injection (30 mg/ml TNBS in 50% ethanol) in female Sprague Dawley rats (Brenna *et al*., 2013 and Sun *et al*., 2013) which were subsequently recovered some days later. The weight loss seems to correlate with endoscopic evaluations whereby the colon was erythematous and edematous on day 3 and 7 and healing was evident on day 12 with mucosal granulation and ulceration (Brenna *et al*., 2013). Additionally, the DAI scores increased steadily after TNBS exposure while control rats did not display any of the clinical signs (Sun *et al*., 2013).

1.3.4.1.4. Chronic DSS-colitis

A chronic DSS-colitis model can also be created by giving rats an alternating regime of DSS-water and water multiple times (Okayasu *et al*., 1990; Cooper *et al*., 1993 and Gaudio *et al*., 1999). For example, one study exposed rats to DSS for 6 days (acute phase) before switching to 6 days of water and then restarting the DSS treatment again (Gaudio *et al*., 1999). The original study in mice seemed to determine the length of DSS exposure by the appearance of the clinical signs (reduction in stool consistency, weight and presence of blood) and DSS treatment was stopped long enough for the mice to recover completely before DSS treatment was restarted (Okayasu *et al*., 1990). This alternating cycle can be repeated from 2 to 8 times (Gaudio *et al*., 1999 and Kullman *et al*., 2001). During subsequent cycles (chronic phases), the clinical signs of the model were observed to be more severe, had a quicker onset and lasted longer in male Sprague Dawley and Wistar rats (Gaudio *et al*., 1999 and Kullman *et al*., 2001). Macroscopic and microscopic scores of DSS-colitis rats were more severe with shortened colons compared to controls (1 cycle: control: 14 cm; DSS: 12 cm; 2 cycles: control: 16 cm; DSS: 12 cm; 3 cycles: control: 18 cm; DSS: 13 cm) and had more severe and numerous ulcerations and lesions in the distal colon that worsened with the application of more cycles (Gaudio *et al*., 1999; Kullman *et al*., 2001 and Vetuschi *et al.,* 2002). There was also an increase in apoptosis of colonic cells by a factor of 20- or 120-fold after two or three cycles in comparison with one cycle of DSS exposure (Vestuschi *et al*, 2002). Comparatively, apoptosis in control animals was always less than 1% (Vestuschi *et al*., 2002). At higher concentrations of DSS, mortality was more likely to occur if rats were exposed to more cycles (Kullman *et al*., 2001).

1.3.4.1.5. Acute DSS-colitis

Visceral hypersensitivity with CRD has also been measured in DSS animals, however, the results have not been consistently reported in studies on mice. A study on BALB/c and C57BL/6 mice reported that there were no differences in visceral hypersensitivity between DSS-colitis and control mice after an acute exposure to DSS for 5 days and up to 51 days (Larsson *et al*., 2006). This was despite high inflammatory and MPO scores that suggested the DSS-colitis induced was quite severe. However, differences in visceral hypersensitivity were observed in a similar study in male BALB/c mice (Verma-Gandhu *et al.,* 2006) where there was a reduced visceral hypersensitivity in DSS-colitis mice after 5 days of DSS. Interestingly, after a chronic DSS model (repeated alternating regime of DSS-water and water), visceral hypersensitivity was no longer present (*i.e.* the pressure required to elicit a response was similar in DSS-colitis and controls). The differences observed may be due to differences in protocol as the studies utilised different percentages of DSS (4% and 5%), differences in molecular weight of DSS (44 kDA and 40 kDa) and days exposed to DSS (5 and 6 days). The severity of DSS-colitis is known to increase when a higher DSS concentration is used, a larger amount of DSS consumed and the DSS has a larger molecular weight (Kitajima *et al*., 2000; Vowinkel *et al*., 2004; Kullman *et al*., 2011 and Goncalves *et al*., 2013). The authors of the first study attributed their negative finding to their DSS source, potential differences at the level of the spinal cord between species and colon compliance (Larsson *et al*., 2006). The lack of visceral hypersensitivity in the second study after chronic exposure to DSS suggests the inflammatory activity in response to the DSS exposure attenuated visceral hypersensitivity (Verma-Gandhu *et al*., 2006). Nonetheless, this lack of visceral hypersensitivity after DSS exposure is interesting as it seems to mimic the human disease, whereby patients display reduced visceral sensitivity (Chang *et al*., 2000; Sharkey *et al*., 2006). However, ongoing pain is still reported by these patients (Chang *et al*., 2000). This phenomenon has been attributed to an upregulation of sensitivity during acute tissue damage and an activation of counter regulatory mechanisms when the damage is chronic (Chang *et al*., 2000). Alternatively, visceral hypersensitivity can also be assessed by applying von Frey filaments to the rats' abdomens as mice with DSS-colitis were found to withdraw their abdomen when a 0.04 g of force was applied (Jain *et al*., 2015). It has also been observed that male Sprague Dawley rats with DSS-colitis had an increased excitatory neuronal response when they were assessed with the CRD (Qin *et al*., 2008).

Somatic referred hypersensitivity to the hind paw has been observed in animals with DSS-colitis. Male C57BL/6 mice with DSS-colitis withdrew their hind paw from mechanical stimulus or displayed a lower latency to withdraw their hind paws from a thermal stimulus compared to controls (Jain *et al*., 2015).

1.3.4.1.6. DSS-colitis and other non-specific indicators of pain

Mice with DSS-colitis were observed to perform more nocifensive behaviours. These behaviours include locomotion, time spent climbing and rearing in response to a splash test (where mice were sprayed with water and the frequency of rearing was assessed; Lapointe *et al*., 2015 and Jain *et al*., 2015). In C57BL/6 mice with DSS-colitis, mice travelled less and spent less time climbing compared to control animals (DSS-colitis: 20m travelled and 500s spent climbing; controls: 50m travelled and 1750s spent climbing; Lapointe *et al*., 2015). These behaviours continued to be depressed five weeks after DSS treatments were discontinued (30m travelled and 750s spent climbing). In another study, male C57BL/6 mice with DSS-colitis not only had depressed locomotion, they also reared less after the splash test (Jain *et al*., 2015). These data suggest that DSS-colitis animals have a lower motivation to move compared to control animals and this may be because of the pain experienced.

1.3.4.1.7. DSS-colitis and other measurements

The DSS-colitis model has also been characterised by a decrease in food (DSS: 131g; control: 143g) and water consumption (DSS: 157 mL; control: 197 mL), decreased urine output (DSS: 78 mL; control: 125 mL), an increased white blood cell count (DSS: 25000; control: 5000) and an increase in serum inflammatory cytokines (TNF-α, IL-β, IL-6; Togawa *et al*., 2002 and Geier *et al*, 2007)

1.3.4.2. Comparing TNBS- and DSS-colitis

Overall, both types of colitis models are well accepted as models for human colitis. Both colitis models utilise similar assessment methods to assess the severity and progression of the colitis model: visceral and referred somatic hypersensitivity, the DAI, microscopic and macroscopic scores and assessment of inflammatory markers. In both models, procedure related mortality can be avoided provided the doses are carefully chosen. However, both models also have a similar drawback whereby the experimental protocols from published studies differ. For example, in the TNBS-colitis studies the experimental protocol can differ by the different percentages of ethanol, volume and TNBS weight used, animal strains, variation of anal leakage and the duration of fasting (Gambero *et al*., 2007 and Brenna *et al*., 2013). This issue of inconsistency is also evident in the DSS-colitis model as severity of colitis can be impacted by DSS concentration, duration of DSS exposure and molecular weight of the DSS powder (Kitajima *et al*., 2000; Vowinkel *et al*., 2004; Kullman *et al*., 2011 and Goncalves *et al*., 2013). Some studies even fail to report the molecular weight of DSS powder used (Bramhall *et al*., 2015). This is problematic as studies reported using a 5% concentration of DSS with similar duration of exposure to DSS but had different molecular weights reported a larger variation of average DAI scores which ranged from 1 to 3.5 out of 4 (Stucchi *et al*., 2000; Kihara *et al*., 2003; Osman *et al*., 2004 and Dicksved *et al*., 2012). Lastly, in terms of practicality, the DSS-colitis model is preferable to the TNBS-colitis model as it is less intrusive because the creation of a DSS-colitis model does not require anesthesia. Instead, the model is created by simply adding DSS powder into the rats' drinking water. Without the need for anesthesia, the animals are probably less stressed because it has been reported that rats find anesthesia agents to be aversive (Leach *et al*., 2002 and Altholtz *et al*., 2006). This is important as stress has been shown to be involved in modulating visceral pain and hypersensitivity from colitis (Larauche *et al*., 2012). Therefore, the repeated exposure to anesthesia and the repeated administration of the intrarectal injections of TNBS may result in stress of the animal and introduce confounding factors during the pain assessment. Given this, an acute and chronic DSS-colitis model was selected as a visceral pain model to assess the utility of the RGS.

1.3.5. Differences between acute and chronic DSS-colitis

The methodology of acute and chronic DSS-colitis models is very similar. The chronic DSS-colitis model is induced by two or more exposures to DSS, therefore it could be suggested that the subsequent exposures to DSS are a repeat of the acute phase. This begs the question: are the repeated phases of DSS-colitis representative of a chronic process or are they repetitions of the acute phase?

1.3.5.1. What happens during the 'acute' phase of DSS?

During the first phase of DSS (acute phase), animals develop the clinical signs typical of the colitis model (loose stools, rectal bleeding and weight loss) beginning from day 3 after the DSS treatments start (Cooper *et al*., 1993 and Gaudio *et al*., 1999). The colons of these animals also shorten significantly (shrinking ~15% in rats and ~50% in mice) and colon weights increase in comparison with controls (Gaudio *et al*., 1999; Okayasu *et al*., 1990 and Bento *et al*., 2012). When examined microscopically, the colon mucosa displayed crypt loss, thinning epithelium and inflammatory infiltrates at the mucosal and submucosal layers (Cooper *et al*., 1993 and Gaudio *et al*., 1999). The inflammatory cytokine profile was characterised by elevated levels of serum proinflammatory cytokines in comparison to controls (TNF-α, IL-6, IL-12 and IL-17; Alex et al., 2009 and Bento *et al*., 2012). Whereas the pro-inflammatory cytokine, IL-1β, and anti-inflammatory cytokines, IL-4 and IL-10, were similar to control levels (Bento *et al*., 2012).

When animals were allowed to recover with a water only phase (representing the 'remission' phase typical of human ulcerative colitis), these animals usually recovered quickly with clinical signs resolving around 4-10 days after DSS treatment stopped (Gaudio *et al*., 1999 and Bento *et al*., 2012). The colon length of these animals also recovered and were significantly longer than the colons during the acute phase (~30% longer) but remained shorter than controls (~13% shorter; Melgar *et al*., 2005; Hall *et al*., 2011 and Bento *et al*., 2012). During this time, the anti-inflammatory cytokines of IL-10 and IL-4 increased and TNF-α levels remained elevated in comparison to controls (Bento *et al*., 2012). Interestingly, C57BL/6 mice developed a concomitant chronic colitis when allowed to recover with a prolonged water phase after the initial acute phase of DSS (Melgar *et al*., 2005). However, in Swiss Webster and BALB/c mice, a slow regenerative healing (colon not completely regenerated 5 weeks after acute phase ends) and quick regeneration (colon appears normal 3 weeks after acute phase ends) were observed respectively (Dieleman *et al*., 1998 and Melgar *et al*., 2005). During this time, a progressive production of pro-inflammatory cytokines (IL-1β, IL-12, IL-17 and IFN-ϒ) was observed in C57BL/6 mice (Melgar *et al*., 2005). In these mice, loose stools, high inflammatory scores and no crypt healing was observed (Melgar *et al*., 2005). In Swiss Webster mice, low grade dysplasia was evident (Cooper *et al*., 1993) and two weeks after stopping DSS treatment, the colonic mucosa tissues had increased IFN-ϒ, IL-4 and CD4+ T cells in areas of inflammation and regenerating crypt lesions, suggesting that the immune response was triggered and may play a role during the regenerative phase of DSS included colitis (Dieleman *et al*., 1998). Increases in B and T cells were also observed in C57BL/6 mice (Melgar *et al*., 2005).

1.3.5.2. What happens during the subsequent phases of DSS?

During the second phase of DSS administration (chronic phases), bleeding and loose stool consistency appeared earlier (day 2) when the DSS treatment restarted (Okayasu *et al*., 1990; Gaudio *et al*., 1999; Cooper *et al*., 1993). In some studies, after 4-5 cycles, the clinical symptoms no longer resolved during the water phase but continued to worsen (Okayasu *et al*., 1990). Depending on the study, animals began to gain weight during the subsequent phases of DSS but were always about 20% lower than control animals (Gaudio *et al*., 1999) or continued to lose weight (Bento *et al*., 2012). More animals developed dysplasia at the mucosal epithelium when more cycles were induced (Okayasu *et al*., 1990 and Kullman *et al*., 2001). The colons of these animals were shorter in comparison to control animals and similar to those observed in the acute phase (~25% shrinkage in rats and ~50% in mice, Okayasu *et al.,* 1990; Gaudio *et al*., 1999; Bento *et al*., 2012). However, after the second water phase, the colons no longer regenerated (Bento *et al*., 2012). Microscopic changes were similar to the acute phase in terms of crypt loss and inflammation (Cooper *et al*., 1993) and increased focal erosion of the epithelium, greater crypt dilation and goblet cell depletion was evident with more cycles (Gaudio *et al*., 1999). The cytokine profile during the chronic phase (2 and 4 cycles) was characterised by increased serum proinflammatory cytokines (TNF-α, IL-1β, IL-6 and IFN-ϒ) and anti-inflammatory cytokine (IL-10) in comparison to control animals (Bento *et al*., 2012 and Alex *et al*., 2009). However, one study reported an increase in IL-4 in C57BL/6 mice after 4 cycles of DSS (Alex *et al.*, 2009) while the other did not in BALB/c mice after 2 cycles of DSS (Bento *et al*., 2012). This difference may be attributed to the differences in strain and the number of cycles of DSS administered. The increase in anti-inflammatory cytokines (IL-4 and IL-10) during the chronic phases of DSS suggests an anti-inflammatory dominant profile during the chronic phase which is similar to human ulcerative colitis (Alex *et al*., 2009).

During the second water phase, TNF-α, IL-1β remain elevated and CD4+ and CD8+ T cells and TGF-β increased from controls in BALB/c mice (Bento *et al*., 2012). This suggests that the colon is adapting during the chronic phases, with regulatory T-cells now releasing the observed increase in anti-inflammatory cytokines (IL-10, TGF-β and FoxP3; Bento *et al*., 2012).

In conclusion, it seems that changes are initiated after the initial insult of DSS treatment, with increased levels of pro-inflammatory cytokines and immune cells in DSS-treated animals. These pro-inflammatory cytokines were still present when DSS treatment is halted. The cytokine profiles are also different during the acute and chronic phases, with the acute phase characterised by a pro-inflammatory cytokine profile and the chronic phrases characterised by both pro- and anti-inflammatory cytokine profiles. These differences are evident as early as the second phase of DSS. Differences are also evident with clinical signs reappearing earlier and taking longer to resolve and the colon lengths are no longer regenerating after the second phase of DSS treatment. Therefore, these studies suggest that two phases of DSS exposure is sufficient to create a chronic colitis model.

1.4. Pain recognition training

It has been reported that underestimation or failure to recognise pain is the major reason some veterinarians withhold or provide inadequate analgesics to their patients (Hugonnard *et al*., 2004; Raekallio *et al*., 2003; Hewson *et al*., 2006; Dohoo and Dohoo, 1996; Mich *et al*., 2010 and Lim *et al*., 2014). The further training of veterinarians and veterinary students to recognise painful behaviour or the use pain assessment tools has been proposed (Mich *et al*., 2010). Two studies have assessed the effects of training in veterinary students to recognise pain. In both studies, students were given a formal single, brief training assessment (30-40 minutes) of pain in dogs or cats and then asked to assess their level of confidence on the identification of painful behaviours post-training (Mich *et al*., 2010 and Lim *et al*., 2014). After training, students reported improved confidence in assessing pain which was statistically significant. These students also improved and were able to recognise subtle signs of pain that they had missed before. However, despite the formal training sessions, students were unable to score in a similar manner to experienced veterinarians (Mich *et al*., 2010 and Lim *et al*., 2014). This suggests that the training provided was insufficient for students to be as competent as experienced raters and that experience or more training was still required. Another study assessed the effect of experience on scoring ability (Doodnaught *et al*., 2017). This study also found that veterinary students were unable to score similarly to experienced veterinarians (graduate veterinarians and anesthesiologists; Doodnaught *et al*., 2017). In contrast, another study assessed if three behaviours that rats typically displayed after a laparotomy surgery could be recognised by raters with varying levels of experience with rats (*i.e*. administrative staff, animal technicians and researchers; Roughan and Flecknell, 2006). Before a short (10 minute) training session, raters were asked to assess pain with a visual analogue scale from a 5-minute video of a rat. After the training session, most of these raters were able to recognise painful behaviours and demonstrated improvement in differentiating between rats with and without pain. It appears that the correct identification of painful behaviours requires learning and that providing of training tools will not ensure a similar ability in the scoring of pain with raters of different experience levels (Haidet et al., 2009 and Campbell *et al*., 2014).

Interestingly, the underestimation of pain and overestimation of analgesic efficacy by healthcare professionals for human patients has also been observed (Klopfenstein *et al*., 2000).

Like veterinarians, some doctors also report low confidence in their ability to assess pain and to treat pain (Silvoniemi *et al*., 2012). Additional training and use of a systematic assessment tool have also been suggested (Klopfenstein *et al*., 2000 and Silva *et al*., 2013). When nurses were trained and provided with a systemised assessment form (an application form), their patients reported greater pain relief than patients with untrained nurses, thus demonstrating that trained nurses were more effective at pain management (Silva *et al*., 2013). When the trained nurses no longer had access to the systemised assessment form, they were less able to manage pain as effectively. Furthermore, with training and a systemised assessment form, health care professionals reported increased confidence at assessing pain and were more likely to administer analgesics to their patients (Silva *et al*., 2013 and Heinrich *et al*., 2015).

Therefore, training to recognise pain and proper knowledge of pain assessment tools is required for veterinarians and human healthcare professionals to be effective at pain management.

1.4.1. The Rat Grimace Scale and training

As a pain assessment tool, the RGS is simple to use as it consists of only four AUs (Sotocinal *et al*., 2011). These AUs are all located on the rat's face which human observers tend to focus on (Leach *et al*., 2011). Each AU is assigned a score of 0, 1 or 2 depending on the intensity or how obviously present it is. The assessment of AU intensities makes the scoring system subjective and may have an impact on the reliability of the scale (Cohen *et al*., 2007). Therefore, training and assessment of inter-rater reliability should be performed before a rater begins to score. This ensures that raters are proficient and can score reliably with one another. However, when RGS or MGS studies describe the training undertaken by raters, it varies from only reviewing training manuals (Faller *et al*., 2015) to one training session (Langford *et al*., 2010; Sotocinal *et al*., 2011; Oliver *et al*., 2014 and Philips *et al*., 2017) to multiple training sessions (Mittal *et al*., 2016). Additionally, few studies assess and report reliability between raters (Langford *et al*., 2010; Sotocinal *et al*., 2011; Oliver *et al*., 2014 and Mittal *et al*., 2016). Therefore, it is unknown how proficient these raters were during the time of scoring. A simple solution to assess competency would be to train and assess the inter-rater reliability between raters or between a new and experienced rater (Streiner and Norman, 2008). This would ensure that the inter-rater reliability between new and experienced raters was at an acceptable standard and would also help to identify rogue raters who could be excluded or sent for more training (Mittal *et al.*, 2016 and Mullard *et al.*, 2017). Assessment of intra-rater reliability may also be used to ensure raters were scoring consistently with themselves over time (Oliver *et al.*, 2014).

While the usefulness of training has not been assessed adequately with animal grimace scales, it has been assessed in a study on human facial expressions of pain (Solomon *et al.*, 1997). In this study, new raters were given a 30-minute training session to identify four different facial movements: frown, eyes close, nose wrinkle and squint. While these raters were more likely to pick up subtle facial movements after training, raters still tended to underestimate pain. Overall, raters in this study only improved slightly in their abilities to pick up the four facial expressions of pain. Another study also assessed the reliability of two raters experienced with the Facial Action Coding System (FACS) to assess 19 different AUs (Sayette *et al.*, 2001). These experienced raters had excellent reliability at assessing 11 of the AUs when they assessed for the absence and presence of the AUs. However, reliability fell when raters assessed the intensity of each AU. Interestingly, reliability was better when raters assessed each AU on a 3-point scale instead of a 5-point scale. This study suggests that although experienced raters were able to agree if most of the AUs were present or absent, they were less able to agree on the intensity of individual AU.

These data suggest that a single short training session is insufficient for raters to be proficient at scoring behaviours or facial expressions, however, longer and multiple training sessions could be beneficial to learn and recognise multiple AUs. Therefore, proficiency in RGS scoring may require more than a single training session. Ultimately, assessing the proficiency of raters at utilizing the RGS is key to evaluate if training was effective and this can be done by assessing reliability and accuracy of scores. This thesis explores this issue by assessing the reliability and the proficiency of trainee raters in comparison to an experienced rater after multiple training sessions. An additional group of raters which only scored images was also included to assess if the scoring of multiple images without training would results in improved reliability and proficiency in comparison to an experienced rater.

1.5. The importance of good reporting

Published papers are the building blocks of scientific progress, each paper generates new information that future research builds upon. Accurate and transparent reporting within scientific papers is vital for study validation, replication and use in retrospective analysis such as systematic reviews and meta-analyses (MacCallum 2010; du Sert, 2011 and Freedman *et al*., 2015). Study validation requires sufficient information describing the methods and results so that readers may critically evaluate the findings and conclusions made by the authors. The data from a properly reported study may also be used to direct future animal research or human clinical trials (*e.g.* deciding if a novel analgesic has potential for human use). Adequate study reporting is also required for study replication which may be performed to ensure that the reported study is reliable and valid and to determine if the results can be replicated in a similar or different population or environment. This cannot be performed if insufficient information is reported in the original study. For example, when the use of anesthesia and analgesia are unreported, replicate studies cannot reproduce the study in its entirety and may even incorrectly assume anesthesia or analgesia were not used (Carbone and Austin, 2016). Lastly, only well-reported studies may be incorporated into a metanalysis or systematic review (Rice *et al*., 2013). This allows for the generation of a larger data set to test new hypotheses without the need to use more animals. This also allows for the confirmation of findings from smaller studies through increased power. Incorporation into metanalyses or systematic reviews cannot be performed if the original studies were poorly performed or had substandard reporting. Due to these limitations, the impact of poorly reported animal studies affects the number of animals used in research and significantly impacts welfare and ethical considerations.

Furthermore, poorly reported research has important financial implications. It is estimated that 53% of reported preclinical studies report irreproducible results, leading to a loss of $28 billion dollars spent on irreproducible data annually in the United States alone (Freedman *et al*., 2015). This is further compounded by the use of further funds in repeating flawed studies to check, correct and refute findings (Freedman *et al*., 2017).

Importantly, reporting quality has been associated with study design quality (Macleod *et al*., 2008, Sena *et al*., 2010 and Holman *et al*., 2015). For example, studies were more likely to report

NXY-059 (a free radical scavenger with supposed neuroprotective properties) as efficacious when randomisation and blinding were unreported (Macleod *et al.*, 2008). When these two items are unreported in a study, it is suggested that they were either not performed or were poorly performed, thus allowing for the element of bias in the study and likely inflating the changes of obtaining a positive result. Furthermore, when Swiss researchers were asked which measures were effective against different types of biases, many researchers answered incorrectly (Reichlin *et al.*, 2016). Together, these studies demonstrate that readers should not assume a study was performed properly if the expected and necessary steps are unreported. Therefore, poor reporting equates to doubt over the study's validity due to possible issues with experimental bias (Rice *et al*., 2013).

Overall, good reporting is important to maintain scientific progress and has ethical and financial implications. This seems to be in consensus among authors, journal editors and funding agencies (MacCallum 2010; du Sert 2011 and Ma *et al*., 2017).

1.5.1. What is the reporting standard in animal research?

In 2009, a systematic review of 271 *in vivo* papers from 1999-2005 was performed by Kilkenny et al. This study found that studies were generally poorly reported, with fundamental information missing. For example, the sex, age and weight of animals used were reported in 74%, 43% and 46% of the 271 papers respectively. Alarmingly, only 13% reported both the age and weight and 24% did not report either age or weight. Most curiously, items that impacted the study's validity and are important to study design: sample size justification (0% reported), randomisation (12%) and blinding (14%) were also unreported or never reported. Of the studies that reported use of statistical analysis (247/271), adequate reporting of statistical methods and results were only present in 70% of the papers. Overall, most of these papers were missing fundamental items which prevented the readers from judging the validity of the results (*i.e.* description of study design and statistical analysis), reproducing the data or including it in retrospective analysis (*i.e.* basic information such as animal sex, age and weight).

1.5.2. Introducing the ARRIVE (Animal Research: Reporting of In Vivo Experiments) guidelines and its… impact?

In 2010, the ARRIVE guidelines were published to guide and promote improved reporting of animal research papers (Kilkenny *et al*., 2010). The ARRIVE guidelines consist of 20 items with sub-items that all animal research papers should include to be complete and to maximise the data gathered (Kilkenny *et al*., 2010). A recommendation is provided beside each item that suggests how it should be reported. For example, under item number 9 “Housing and husbandry”, authors should provide details of housing, husbandry conditions and welfare assessments that the animals were exposed to during the study. Within these sub-items, authors are recommended to include the cage type, bedding materials, number of cage companions used *etc*. These guidelines were prepared by researchers, statisticians, journal editors and funders of research. The ARRIVE guidelines have since been published by 13 journals, supported by hundreds of journals and by various funding agencies, universities and learned societies organisations (www.nc3rs.org.uk).

Since the publication of the ARRIVE guidelines, there have been various systematic studies on reporting quality to assess the current standards and to evaluate if these reporting standards have improved (Schwarz *et al*., 2012; Baker *et al*., 2014; Delgado-Ruiz *et al*., 2014; Bara and Joffe, 2014; Ting *et al*., 2015; Gulin *et al*., 2015; Lin *et al*., 2016; Avey *et al*., 2016; Nam *et al*., 2018). Overall, these studies reported that reporting standards remain low and the publication of the ARRIVE guidelines does not seem to have made a significant impact. For example, the reporting rates of basic animal information (*i.e.* sex, age and weight) and key items required to assess study validity (*i.e.* sample size justification, randomisation, blinding and statistical analysis) remain poorly reported in many studies (Table 1.5.2.1 and 1.5.2.2).

The reasons for the low reporting standards are probably multifactorial. Firstly, more time may be required for the ARRIVE guidelines to take hold as a relatively short time has passed since they were first published. This may also explain why many authors were not aware of them (Reichlin *et al*., 2016 and Ma *et al.,* 2017). Therefore, more time and perhaps better education and promotion of the ARRIVE guidelines may be required for authors to use the guidelines. The enforcement of core concepts contained in the guidelines (*e.g.* items related to bias and study precision – the use of blinding, randomisation and sample size calculation) by journals and funding agencies may also help. Additionally, it seems that researchers may be reporting only what they think is minimally required for a paper. This can be observed by the fact that items of the ARRIVE guidelines which were frequently well reported were items that constitute the essential 'skeleton' of a paper: the title, abstract, background, objectives and hypothesis, details on experimental animals (methods), experimental outcomes (results), outcomes and estimation, interpretation and implication (discussion; Schwarz *et al*., 2012; Delgado-Ruiz *et al.,* 2014; Bara and Joffe, 2014; Ting *et al*., 2015; Lin *et al*., 2016; Avey *et al*., 2016; Nam *et al*., 2018). In contrast, items that were generally poorly reported are not always deemed necessary or essential: welfare assessments, sample size calculations, baseline data of animals, numbers analysed, adverse events, implications for replacement, refinement or reduction of animal use. Lastly, it is interesting to note that reporting standards within a certain field remain consistent overtime. For example, it was observed that 369 papers on neoplasm in rodents from 2010 to 2012 had very similar reporting rates for all items (Lin *et al*., 2016). Although this was over a short amount of time, this suggests that authors within the same field take cues from one another regarding which items should be reported. This was also observed by Macleod *et al.* (2015) where different fields were more likely to report certain items.

Overall, it seems that the ARRIVE guidelines have not made a substantial impact on the reporting standards of animal research papers. It remains to be seen if this will change over time. Use of the ARRIVE guidelines and the importance of good reporting should continue to be promoted to researchers and enforcement of the entirety or part of the guidelines should be employed.

1.5.3. Items listed in the ARRIVE guidelines related to pain research

Studies have been performed which demonstrate the various factors that can affect outcome measures in pain research. These items highlight the importance of precise descriptions in pain research. These include stress induced analgesia which has been reported to be affected by experimenter (*e.g.* male olfactory cues) and when experiments are conducted in a novel environment (Abbott *et al*., 1986; Chesler *et al*., 2002 and Sorge *et al*., 2014). Other testing conditions such as time of testing (*i.e.* effect of circadian rhythm), testing surfaces and presence of conspecifics have been reported to affect pain assessments (Frederickson *et al*., 1977; Pitcher *et al*., 1999 and Langford *et al.,* 2006). Demographics of experimental animals utilised in the study is also known to affect the results of pain assessment methods such as sex, strain and age (Chesler *et al*., 2002; Mogil *et al*., 2005; Legg *et al*., 2009 and Sorge *et al*., 2011; 2015). The effects of housing have been described to also affect the outcome of pain assessment methods, these include the type of bedding (Moehring *et al*., 2005), cage size, presence of enrichment items (Tall, 2009 and Gabriel *et al*., 2010) and number of companions (isolated housing, overcrowding and aggression; Nishikawa and Tanaka, 1978; Pilcher and Browne, 1982; Puglisi-Allegra and Oliverio, 1983; Gentsch *et al*., 1988; Brown *et al*., 1995; Coudereau *et al*., 1997 and Tuboly *et al*., 2009). Lastly, the effects of dietary differences of fat, sugar content, caloric restriction and iron deficits may also affect pain assessment results (Yehuda *et al*., 1986; Frye *et al*., 1993; and de los Santos-Arteaga *et al*., 2003; Hargraves and Hentall, 2005; Perez *et al*., 2005; Dowling *et al*., 2009; Martin and Avendano, 2009; Ross-Huot *et al*., 2011 and Veigas *et al*., 2011).

1.5.4. Conclusion

In conclusion, good reporting is required to maintain scientific rigor as well as prevention of ethical and financial issues associated with poor reporting. The ARRIVE guidelines were published to give authors clear guidance on what should be reported to maximise the data gathered from each study. However, reporting standards remain poor and the efforts by all stakeholders are required to improve reporting standards. This thesis explores this issue by assessing if papers published in veterinary journals five years after the publication of the ARRIVE guidelines have improved. In addition, a comparison was made between papers that were published in journals that support the ARRIVE guidelines and journals that did not. Both of these comparisons assess if the publication of the ARRIVE guidelines resulted in improved reporting standards of published veterinary papers.

Table 1.5.2.1: Summary of the reporting rates of three basic animal demographic details that should be reported in all papers

	Animal details			
Sex (%)	Age (%)	Weight (%)	N (of papers)	Reference
74	43	46	271	Kilkenny *et al*., 2009
-	30	70	25	Delgado-Ruiz *et al*., 2014
79	44	56	47	Avey *et al*., 2016
60	82	24	50	Nam *et al.,* 2018
77	38	78	77	Bara and Joffe, 2014
90	58	52	83	Gulin *et al.,* 2016

Legend: Table compares the reporting rates of different basic items from various systematic reviews assessing reporting adherence to the ARRIVE guidelines. As there were no differences observed of items reported in papers published before or after the ARRIVE guidelines, the data was merged.

Table 1.5.2.2: Summary of the reporting rates of four key items to the validity of a study design.

Sample size justification (%)	Randomisation (%)	Blinding (%)	Statistics (%)	n	Reference
0	43	46		271	Kilkenny *et al.,* 2009
0	17	29	75	41	Ting *et al*., 2014
88	-	-	72.6	75	Schwarz *et al*., 2012
0	90	0.25	79	396	Lin *et al*., 2016
2	23	30	32	47	Avey *et al*., 2016
8	17	6	42	50	Nam *et al*., 2018
5	61	40	-	77	Bara and Joffe, 2014
0	16	-	61	83	Gulin *et al*., 2016

Legend: Table comparing the reporting rates of 4 key items related to study design from various systematic reviews assessing reporting adherence to the ARRIVE guidelines. As there were no differences observed of items reported in papers published before or after the ARRIVE guidelines, the data was merged

1.6. Research questions, hypotheses and objectives

1.6.1. Can the Rat Grimace Scale be utilised in real-time?

1.6.1.1. Background

The RGS, a facial expression scale, allows for the assessment of the affective component of pain in rats (emotional experience and presence of ongoing pain). This behavioural tool works well as a research tool but is limited as a clinical tool as the standard video-based method is time and labour intensive. The standard method requires the rat to be first video-recorded, then images need to be manually extracted and cropped before scoring can even begin. This results in hours or even days passing before a score can be obtained for a single rat. Overall, it is impractical, and it is difficult to utilise the RGS in a clinical setting where time is limited and decisions to intervene need to be made quickly. Therefore, real-time application of the RGS, where an observer can simply walk in and assess an animal, will drastically reduce the time and labour required as well as lend itself as a clinical tool to assess animals and provide analgesic intervention quickly.

1.6.1.2. Hypothesis and objectives

It was hypothesised that real-time application of the RGS would be as successful as the standard RGS method for the assessment and evaluation of pain in the rat. This was tested through four objectives: 1) assessing if point- and interval-scoring methods would produce scores comparable to the standard method (video-based) method; 2) assessing if real-time scoring was able to identify treatment effects of analgesics over time; 3) assessing the shortest observation time (10, 5 and 2 minute durations) required for real-time scores to be comparable to the RGS scores from the standard method and 4) assessing if there is an effect of an observer being present (do rats display a different facial expressions when an observer is present?).

1.6.2. Can spontaneous behaviours (Rat Grimace Scale, burrowing and Composite Behavioural Score) assess visceral pain in an acute and chronic dextran sulfate sodium colitis model?

1.6.2.1. Background

The RGS was developed with acute inflammatory pain models and its applicability in other pain models has also been demonstrated. However, its applicability in chronic and visceral pain remains unexplored. Visceral pain in animals has been difficult to assess due to the absence of an external injury that can be stimulated. Other spontaneous based behaviours that may allow for the evaluation of visceral pain are: The Composite Behaviour Score (CBS), behavioural ethogram of behaviours (twitch, writhe, back arch and fall/stagger) which rats will display with increased frequency during visceral pain; and burrowing, voluntary spontaneous behaviour that rats are highly motivated to perform which decrease during pain. The visceral pain model selected was the DSS colitis model. This model was chosen because it is a frequently used model, replicable and model induced mortality can be avoided. Up until now, the affective component of pain has not been assessed in this model. Instead, disease progression has been assessed with the DAI (assessment for the presence of bleeding, reduction in stool consistency and weight loss).

1.6.2.2. Hypothesis and objectives

It was hypothesised that the use of spontaneous behaviours would be able to assess pain in an acute and chronic dextran sulfate sodium model. This was tested through two objectives: 1) assessing if the RGS and CBS scores would increase with exposure to dextran sulfate sodium and would follow a similar pattern to the DAI index and 2) assessing if rats would burrow less with exposure to DSS.

1.6.3. What is the effect of training on Rat Grimace Scale scoring?

1.6.3.1. Background

If the RGS is to be used as a tool for the assessment of pain, then its reliability and repeatability between raters is an important consideration. The effects of training and how much training is required to attain proficiency has never been explored and rater training is rarely reported in RGS papers. In general, pain recognition and assessment in animals is difficult and a single training session is usually insufficient for trainee raters to be as proficient as experienced raters. Therefore, not only is undergoing training important, it is also important for some assessment of proficiency. One way to assess proficiency is to assess the inter-rater reliability between trainee and experienced raters. Once proficiency and inter-rater reliability from training is attained, it is important for this proficiency and reliability to be maintained within the individual. Therefore, intra-rater reliability should also be reassessed once a latency period is applied to evaluate an individual's proficiency at retaining the ability to score effectively.

1.6.3.2. Hypothesis and objectives

It was hypothesised that training would improve inter-rater reliability and proficiency would be maintained after a period of disuse. This was tested through three objectives: 1) assessing if trainee raters would have improved inter-rater reliability with one another and with an experienced rater after training sessions, 2) assessing if trainee raters without training would demonstrated improved inter-rater reliability with one another and an experienced rater and 3) assessing if inter-rater reliability with each other and the experienced rater and intra-rater reliability may be maintained after a period of disuse.

1.6.4. What are the reporting standards of papers published five years after the ARRIVE guidelines?

1.6.4.1. Background

In addition to the improvement of pain assessment methods and the training required to use these assessment methods, pain research needs to be well reported. Good and complete reporting in published papers are vital to allow for critical assessment, replication of the study and inclusion in retrospective analyses. However, studies that have assessed reporting standards in published animal studies have indicated that reporting is generally poor in published papers. Poor reporting results in irreproducible animal research that is ethically and financially costly. To combat poor reporting of animal studies, the ARRIVE guidelines, a 20-item checklist, was published in 2010 to inform animal researchers about the information they should include for complete reporting. Since its publication, there has been overwhelming support by hundreds of journals. No studies to date have assessed the adherence of the ARRIVE guidelines in the veterinary literature and none have compared the reporting standards of journals that support or do not support the ARRIVE guidelines.

1.6.4.2. Hypothesis and objectives

It was hypothesised that the publication of the ARRIVE guidelines would result in an increase in the standards of reporting. This was tested through two objectives: 1) assessing if papers published five years after the ARRIVE guidelines were published (2015) would have higher standard of reporting than those papers published prior to the 2009 ARRIVE guidelines and 2) assessing if papers published in journals that support the ARRIVE guidelines have a higher standard of reporting than papers published in journals that do not support the ARRIVE guidelines.

2. Publications

This thesis explored the strengths and limitations of the RGS as a pain assessment tool. Specifically, this thesis explored if practicality of the RGS could be improved with real-time application, if it could be used to assess chronic and visceral pain and the effect of training on the reliability of RGS scoring. This thesis also explored if the publication of the ARRIVE guidelines improved the reporting standards of animal studies.

The first objective was to assess if real-time application of the Rat Grimace Scale could be successfully translated from the standard RGS method. The successful real-time application of the RGS would increase the practicality of the scale in the research setting, establish its usefulness in a clinical setting and to improve the animal welfare of laboratory rodents. This study and results are presented in the first paper titled "*Real-time application of the Rat Grimace Scale as a welfare refinement in laboratory rat*s" (Vivian Leung, Emily Zhang, Daniel SJ Pang), published in Scientific Reports (2016, **6**, 31667; doi: 10.1038/srep31667). Author contributions: Vivian Leung was involved in the data collection of the main study, data interpretation, statistical analysis and manuscript preparation. Emily Zhang was involved in the collection of control animal data, data interpretation and manuscript preparation. Daniel SJ Pang was involved with the study design, data interpretation and manuscript preparation.

The second objective was to assess if the RGS could be used to assess pain in an acute and chronic colitis model. The ability to use the RGS to assess pain in this model opens its utility for visceral pain assessment, a pain type that is difficult to assess in animals because there is no external injury that can be stimulated. This study and results are presented in the second paper titled "*Performance of behavioral assays: The Rat Grimace Scale, burrowing activity and a composite behavior score to identify pain in an acute and chronic colitis model*" (Vivian Leung, Marie-Odile Benoit-Biancamano and Daniel SJ Pang), in press in Pain Reports (2019). Author contributions: Vivian Leung was involved in the study design, data collection, data interpretation, statistical analysis and manuscript preparation. Marie-Odile Benoit-Biancamano was involved in the interpretation of histological data and manuscript preparation. Daniel SJ Pang was involved in the study design, data interpretation and manuscript preparation.

The third objective was to assess the influence of training on the reliability and proficiency of RGS scoring. Training before use of the RGS is rarely described and as such, the influence of training on data reliability and proficiency is unknown. This study and results are presented in the third paper titled "*The influence of rater training on inter- and intra-rater reliability when using the Rat Grimace Scale*", in press in the Journal of the American Association for Laboratory Animal Science (2019). Author contributions: Vivian Leung and Emily Zhang were both involved in the study design, data collection, data interpretation, statistical analysis and manuscript preparation. Daniel SJ Pang was involved in the study design, data interpretation and manuscript preparation.

An additional study was performed to assess if the publication of the ARRIVE guidelines improved reporting standards in animal research papers five years after their publication. It also assessed if journals that support the ARRIVE guidelines were more likely to publish papers with higher reporting standards. This study and results are presented in the fourth paper titled "*ARRIVE has not ARRIVEd: Support for the ARRIVE (Animal Research: Reporting of in vivo Experiments) guidelines does not improve the reporting quality of papers in animal welfare, analgesia or anesthesia*", published in PLOSone (2018, **13**(5): e0197882.https://doi.org/10.1371/journal.pone.0197882). Author contributions: Vivian Leung and Frederik Rousseau-Blass were both invovled in the study design, data assessment, data interpretation and manuscript preparation. Guy Beauchamp was involved with the statistical analysis and manuscript preparation. Daniel SJ Pang was involved in the study design, data interpretation and manuscript preparation.

2.1. Real-time application of the Rat Grimace Scale as a welfare refinement in laboratory rats

Vivian Leung[1], Emily Zhang[1] & Daniel SJ Pang[1, 2]

2.1.1. Abstract

Rodent grimace scales have been recently validated for pain assessment, allowing evaluation of facial expressions associated with pain. The standard scoring method is retrospective, limiting its application beyond pain research. This study aimed to assess if real-time application of the Rat Grimace Scale (RGS) could reliably and accurately assess pain in rats when compared to the standard method. Thirty-two male and female Sprague-Dawley rats were block randomised into three treatment groups: buprenorphine (0.03 mg/kg, subcutaneously), multimodal analgesia (buprenorphine [0.03 mg/kg] and meloxicam [2 mg/kg], subcutaneously), or saline, followed by intra-plantar carrageenan. Real-time observations (interval and point) were compared to the standard RGS method using concurrent video recordings. Real-time interval observations reflected the results from the standard RGS method by successfully discriminating between analgesia and saline treatments. Real-time point observations showed poor discrimination between treatments. Real-time observations showed minimal bias (<0.1) and acceptable limits of agreement. These results indicate that applying the RGS in real-time through an interval scoring method is feasible and effective, allowing refinement of laboratory rat welfare through rapid identification of pain and early intervention.

2.1.2. Introduction

Pain in animals is commonly under-treated. This stems from numerous factors, including the limited availability of validated pain scales (Williams *et al*., 2005; Hewson *et al*., 2006; Hewson *et al*., 2007 and Rialland *et al*., 2012). In laboratory rodents, analgesic administration rates as low as 15% have been reported for invasive procedures (*e.g.* orthopedic surgery, thoracotomy)

[1] Veterinary Clinical and Diagnostic Sciences, Faculty of Veteri ine, University of Calgary, AB, T2N 4Z6 Canada.

[2] Hotchkiss Brain Institute, University of Calgary, AB, T2N 4Z6 Canada.

and data variability related to the presence of pain and sporadic analgesic use is likely to act as a confounding factor during experimental studies (Carbone, 2011 and Stokes *et al*., 2009). Furthermore, some experimental designs allow analgesia to be withheld until established humane endpoints have been reached (Carbone, 2011). These endpoints, such as weight loss, are largely non-specific and little is known about their relationship to pain (Roughan *et al*., 2014). Early recognition of pain coupled with appropriate intervention would address these issues and support refinement of *in vivo* research (Carbone, 2011; Matsumiya *et al*., 2012; Waite *et al*., 2015 and 3rs.ccac.ca). The recent development of rodent grimace scales has expanded our ability to assess pain in rodents (Langford *et al*., 2010 and Sotocinal *et al.,* 2011) and potentially addresses failures in translational pain research resulting from a reliance on evoked-response nociceptive testing (Mogil and Crager, 2004, Rice *et al*., 2008 and De Rantere *et al*., 2016). The Rat Grimace Scale (RGS) consists of four facial "action units" (orbital tightening, nose/cheek appearance, ear and whisker positions) which are scored using still images by an observer (Sotocinal *et al*., 2011). The RGS has been validated, showing content and construct validity and reliability (inter- and intra-observer; Sotocinal *et al*., 2011 and Oliver *et al*., 2014). An analgesia intervention threshold has been derived for the RGS and it has been used to highlight discrepancies between nociception and spontaneous ongoing pain (Oliver *et al*., 2014 and De Rantere *et al*., 2016). The development of both the RGS and Mouse Grimace Scale (MGS) has allowed reappraisals of analgesic efficacy in these species (Matsumiya *et al*., 2012 and Waite *et al*., 2015). In their current form, the RGS and MGS show great potential as research tools in the study of pain. However, the standard method of generating pain scores requires multiple steps: high quality video-recording, automated or manual selection of several images per time point and scoring (Sotocinal *et al*., 2011 and Oliver *et al*., 2014). These steps are time and labour intensive and consequently inhibit wider application of the scales. Performing real-time scoring with the RGS and MGS broadens their applications, facilitating improvements in welfare through rapid, early and accurate identification of pain, thus bridging the gap from research tool to improving rodent care and welfare. Real-time scoring has been attempted in mice (Miller and Leach, 2015) and has been proposed, but remains untested, in rats (Oliver *et al*., 2014). Potential obstacles to real-time scoring are: 1. a change in behaviour in the presence of an observer (observer effect), 2. an inherent bias from the observer being able to observe the whole animal rather than just the head, as performed in the validation studies (observer bias) and 3. limited accuracy of real-time scoring of moving animals without the control offered by video playback. We hypothesised that the standard video-based application of the Rat Grimace Scale could be successfully translated to real-time assessment. This hypothesis was tested through two specific aims: 1) assessing if results from two different real-time scoring methods are comparable to those collected through standard RGS methodology and 2) assessing the shortest observation period possible for real-time scores to remain comparable to standard RGS scores.

2.1.3. Methods and materials

2.1.3.1. Ethical statement

All experiments were approved by the University of Calgary Health Sciences Animal Care Committee and performed in accordance with Canadian Council on Animal Care guidelines.

2.1.3.2. Experimental animals

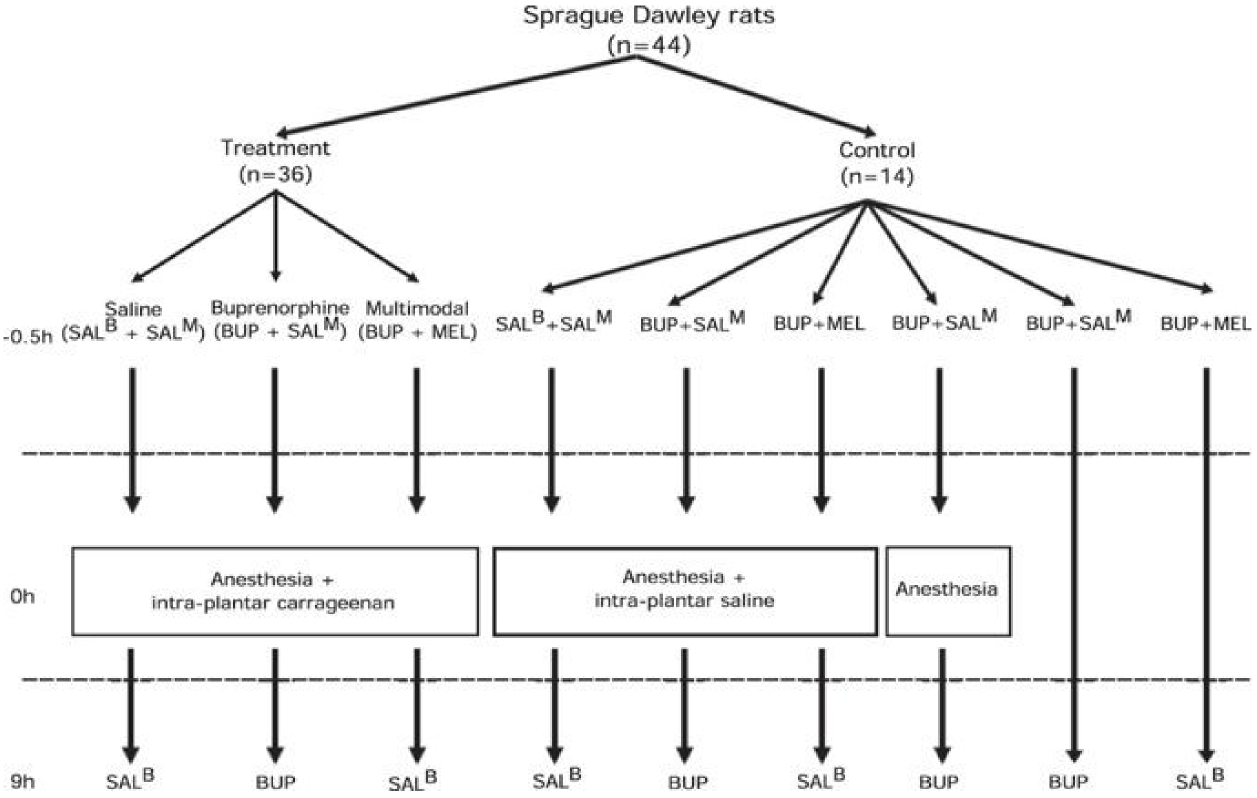
Forty-four male and female Sprague-Dawley rats (224–435 g) were obtained from the University of Calgary Animal Resource Centre surplus stock and Charles River, Canada. Animals were housed in pairs in polycarbonate or polysulfone rat cages (RC88D-UD, Alternate Design Mfg and Supply, Siloam Springs, Arizona, USA) with bedding of wood shavings, shredded paper, sizzle paper and a plastic tube for enrichment. The housing environment was controlled: light cycle of 12 hours on/12 hours off (lights on at 0700) and temperature and humidity settings of 23 °C and 22%, respectively. Laboratory rat pellets (Prolab 2500 Rodent 5P14, LabDiet, PMI Nutrition International, St Louis, MO, USA) and tap water were available ad libitum.

2.1.3.3. Experimental procedures

All animals were habituated to the observer and observation chamber for three days. During these habituation sessions, each animal was placed in the observation chamber for approximately 10 minutes and handled by the observer for at least 20 minutes. Animals were offered a food reward (Honey Nut Cheerios™, General Mills, Inc., Golden Valley, Minnesota, USA) when handled. They were considered habituated when they voluntarily ate the food reward while being held by the observer.

Sample sizes for treatment groups were chosen based on RGS data variability observed in previous publications12,16 with an alpha of 0.05, beta of 0.8 to detect a mean difference of 0.3. Injections were prepared by a third-party not involved in the experiment. All injections were performed between 0700 and 0915 hours and testing completed within the light period. Image scoring and real-time observations were performed by a single observer. Animals were block randomised into one of nine treatment groups (Fig. 2.1.2.3.1). Three treatment groups received intra-plantar carrageenan (100 microlitres of 1% λ -carrageenan dissolved in saline, Sigma-Aldrich, St. Louis, MO, USA) with either buprenorphine (0.03 mg/kg SC, Vetergesic, Champion Alstoe, Whitby, ON, Canada, n = 12), buprenorphine (0.03 mg/kg SC) and meloxicam ("multimodal analgesia group", 2 mg/kg SC, Metacam 0.5% injection, Boehringer Ingelheim, Burlington, ON, Canada, n = 12), or saline (n = 12). A cross-over design was used for the control groups, with each animal receiving three control treatments with a minimum 10-day washout period between treatments (Fig. 2.1.2.3.1).

Figure 2.1.2.3.1: Flow chart depicting experimental pathway for each treatment group.



Legend: SAL^{B}, saline volume equivalent to buprenorphine dose. SAL^{M}, saline volume equivalent to meloxicam dose. BUP, buprenorphine. MEL, meloxicam.

All animals received two sets of injections. The first was given 30 minutes before intra-plantar injection and the second 9 hours after intra-plantar injection (or equivalent time for the control groups). Injections at 9 hours were given after pain assessments were completed.

Intra-plantar injections were performed under brief general anaesthesia. Animals were placed individually in a plexiglass induction chamber and 5% isoflurane carried in oxygen (1 L/min) administered until loss of righting reflex occurred, at which point the animal was transferred to an adjacent counter (anaesthesia maintained by nose cone with 2% isoflurane in 1 L/minute oxygen) and placed in sternal recumbency on a heat pad. The left hind paw was extended caudally, and the plantar surface wiped with 70% ethanol. The assigned treatment (carrageenan or saline) was injected subcutaneously into the plantar surface. Animals were then allowed to recover with 1 L/minute oxygen and returned to their home cages once the righting reflex had returned.

2.1.3.4. Observations

Two video cameras (Panasonic HC-V720P/PC, Panasonic Canada Inc., Mississauga, ON, Canada) were placed at opposite ends of the observation chamber (28 × 15 × 21 cm). During real-time observation the observer was positioned perpendicular to the camera and was free to move around without entering the cameras' field of view. Three observation periods (V1, O+ V, V2) were video-recorded consecutively. V1: video-recording was performed with no observer present. O+ V: real-time observations were performed concurrently with video recording. V2: video-recording was performed with no observer present. Each observation period was 10-minutes long. Observations were performed at baseline (day before procedure) and 3, 6, 9 and 24 h after intra-plantar injections (or equivalent time for control groups).

2.1.3.5. Image RGS scoring

Image scores (IMG) were generated as previously described, by selecting the best image from each consecutive 3-minute period of a 10-minute video (Sotocinal *et al*., 2011). Videos were relabelled by a third party not involved in image grabbing or scoring, blinding the observer to the rat, treatment and time point. The preferred image was a frontal view that clearly showed all action units. A profile view was selected if no frontal image of sufficient quality was available. Images were put into a presentation software (Microsoft PowerPoint, version 15.0, Microsoft Corporation, Redmond, WA, USA) and the slide order randomised before scoring. An average score was calculated from the three images from each video.

2.1.3.6. Real-time RGS scoring

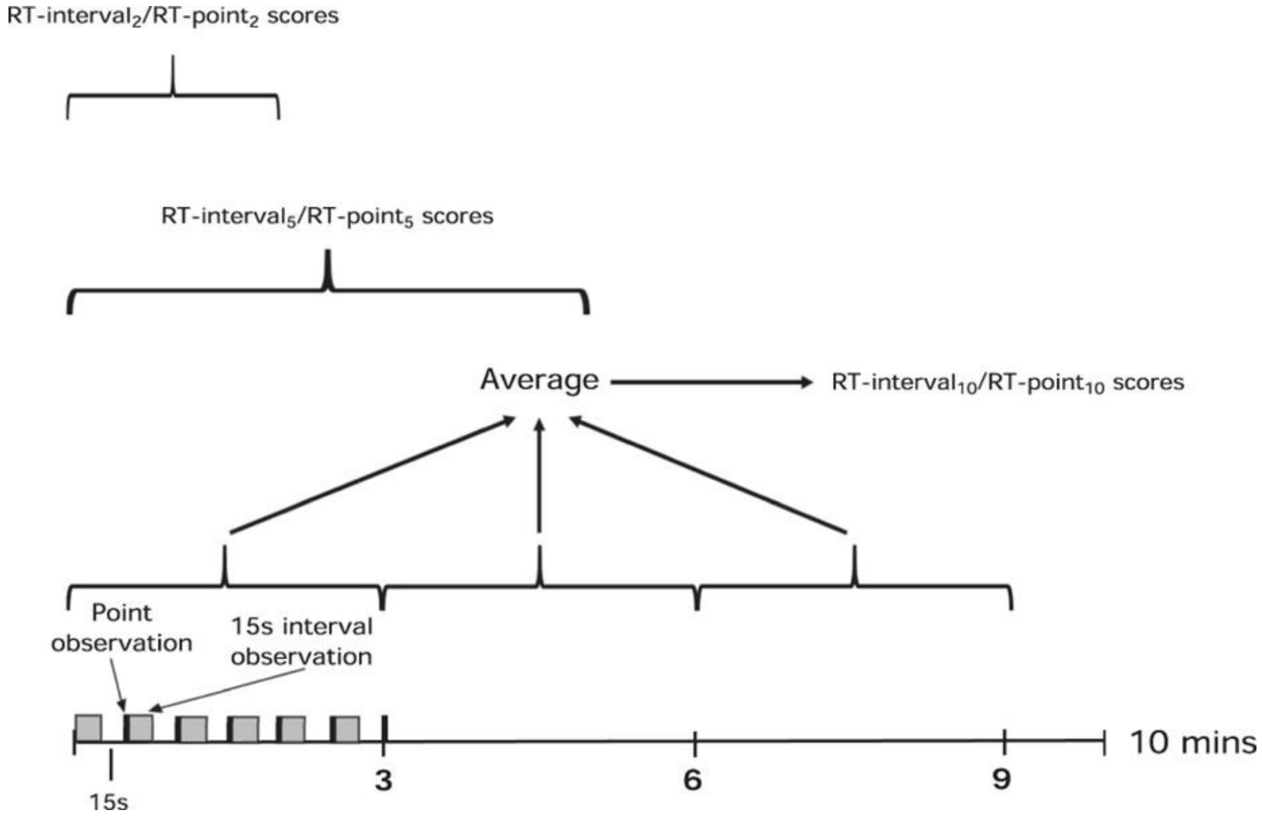
Real-time (RT) scores were obtained using two methods: 1) a point observation alternating with 2) a 15 s interval observation, where the animal was observed for 15 s and assigned a single score for the period. Each method was repeated every 30 s for the 10-minute observation period, generating 18 scores of each type per animal. Similar to the standard method described for RGS scoring (Sotocinal *et al.*, 2011), scores generated from both methods were averaged every three minutes to produce three separate scores and these averaged to yield a single score (RT-interval$_{10}$ or -RT-point$_{10}$). Real-time scores were also averaged from the first five and two minutes of the observation period (RT-interval$_{5}$, RT-point$_{5}$, RT-interval$_{2}$, RT-point$_{2}$) to compare shorter observation periods (Fig. 2.1.2.6.1).

Additionally, five single real-time scores from each 10-minute observation period were randomly selected (single RT-interval and single RT-point) to evaluate variability associated with single observations. Real-time scoring and image grabbing were not performed if a rat was rearing (two paws raised off the chamber floor), sniffing, grooming or sleeping.

2.1.3.7. Pica

A petri dish (given to each cage at the beginning of habituation period) was weighed at baseline and after the experiment as pica is a potential side effect of buprenorphine (Schaap *et al.*, 2012). Pica was confirmed if there was evidence of petri dish fragments at necropsy examination (visual inspection of the stomach contents) or a decrease in the mass of petri dishes (> 0.1 g) was observed.

Fig. 2.1.2.6.1: Cartoon of real-time observation methods


RT-interval$_2$/RT-point$_2$ scores
RT-interval$_5$/RT-point$_5$ scores
Average
RT-interval$_{10}$/RT-point$_{10}$ scores
Point observation
15s interval observation
15s
3
6
9
10 mins

Legend: Observations alternate between point and 15s interval observations. After a 15s pause, the observations are repeated for the 10-minute observation period. Scores from each 3-minute block were averaged and the 3 blocks averaged to give an overall score for the 10-minute period (real-time interval [RT-interval$_{10}$] and real-time point [RT-point$_{10}$]). Raw scores were also averaged over 5 (RT-interval$_5$ and RT-point$_5$) and 2 minutes (RT-interval$_2$ and RT-point^2).

2.1.3.8. Statistical methods

Data analyses were performed using commercial software (Prism 6.07, GraphPad Software, La Jolla, CA, USA). Open source software (R 3.3.0, 'MethComp' package ver. 1.22.2) was used for the Bland and Altman method. Data were assessed for normality with a D'Agostino-Pearson omnibus normality test and parametric tests applied where data approximated a normal distribution. Repeated measures two-way ANOVA was used for between group comparisons with post-hoc tests if a significant main effect was observed: RT-interval and RT-point versus IMG scores (post-hoc Dunnett's test), treatment groups (saline vs buprenorphine vs multimodal; post-hoc Tukey's test), single RT-interval and single RT-point versus IMG scores (post-hoc Dunnett's test), observer effect (RGS scores during observation periods with and without the observer present; post-hoc Tukey's test). When it was not possible to obtain an RGS score for a rat at a given time point, an average of the scores obtained from other rats at the same time point was substituted to allow analysis. The Bland and Altman method for repeated measures was used to assess agreement between IMG scores and RT-interval or RT-point scores (Bland and Altman, 2007). Control data were analysed with Friedman's test with a post-hoc Dunn's test. Differences were considered statistically significant if the computed two-tailed p value was less than 0.05. When available, p values are reported with 95% confidence intervals (95% CI). Data are presented as mean ± SD or median ± interquartile range. Graphs are plotted as mean ± SEM.

2.1.4. Results

2.1.4.1. Multiple interval and point observation scoring methods

Agreement between real-time interval observation scoring methods (RT-interval$_{10}$, RT-interval$_{5}$, RT-interval$_{2}$) were comparable to the standard RGS method (IMG-O+ V, Fig. 2.1.3.1.1). No significant differences were observed between these observation methods at each time point in the saline (F = 1.92, df 3, p = 0.14, Fig. 2.1.3.1.1a) and buprenorphine (F = 1.32, df 3, p = 0.28, Fig. 2.3.1.1.3b) groups. A single difference was observed in the multimodal (buprenorphine and meloxicam) treatment group (F = 13.74, df 3, p < 0.0001) at the 24-hour time point between IMG O+ V and RT-interval$_{10}$ (p = 0.02, 95% CI: 0.02 to 0.35, Fig. 2.1.3.1.1c).

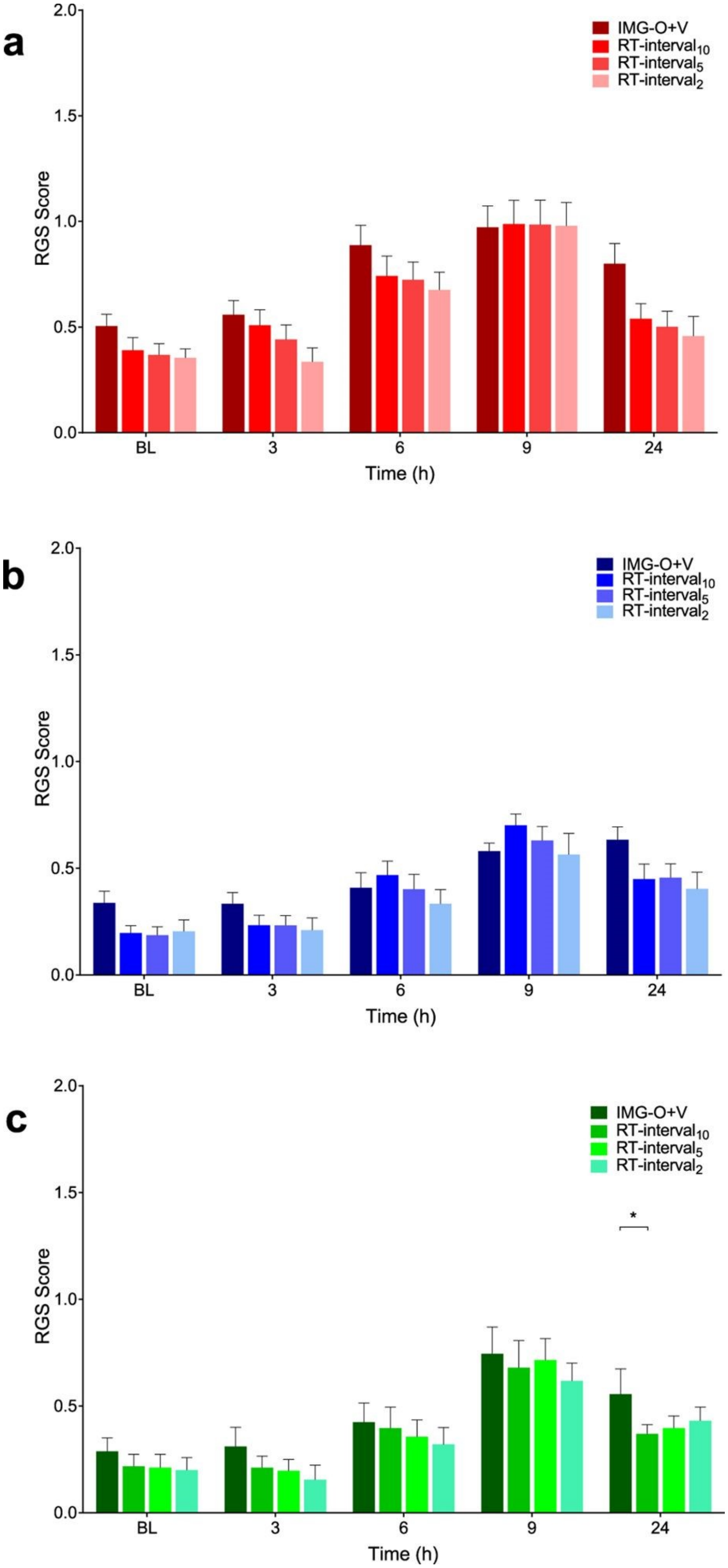


Fig. 2.1.3.1.1: Real-time interval Rat Grimace Scale (RGS) scoring methods were comparable to standard RGS scoring.

Legend: Saline (A) and buprenorphine (B): scoring methods had no significant effect on RGS scores (saline: p = 0.14; buprenorphine: p = 0.28). (C) Scoring method was found to have an effect in the multimodal group. However, the differences was limited to the 24 hour time point (between IMG O+V and RT10, p = 0.02). RT-interval = real-time interval RGS scoring. IMG O+V = standard (video-based) RGS scoring. Data are mean ± SEM.

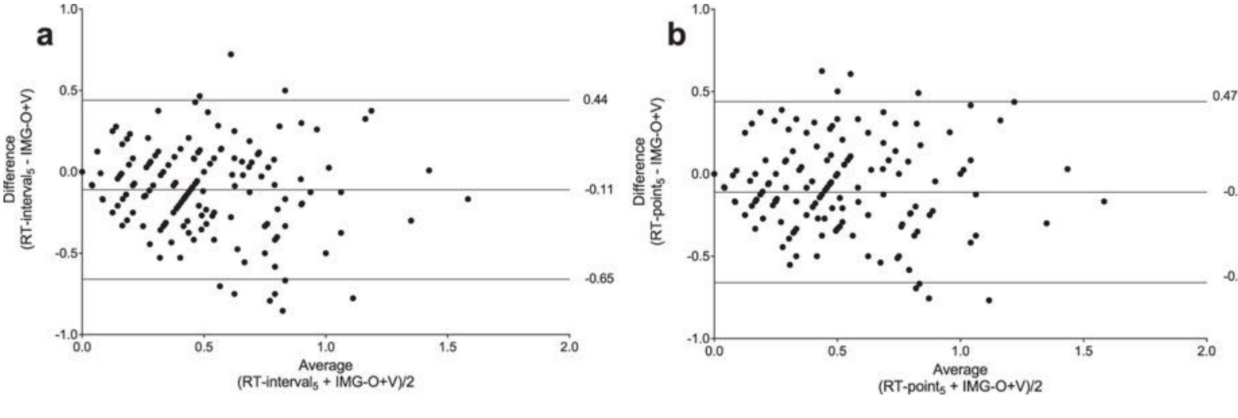
The Bland and Altman analysis revealed that the bias between real-time and standard RGS observation methods was small, regardless of the type or frequency of real-time observations, and represented a systematic underestimation of the standard method by real-time methods of approximately 0.1 (Table 2.1.3.1.1). The limits of agreement (bias ± 2 SD) reflect the distribution of 95% of the measured differences between scoring methods. Observation frequencies of either 5 or 10 minutes showed similar limits of agreement for both interval and point observations (Table 2.1.3.1.1, Fig. 2.1.3.1.2). As observation frequency decreased to 2 minutes, the limits of agreement widened (Table 2.1.3.1.1, Fig. 2.1.3.1.3).

Table 2.1.1.3.1.1: Bland and Altman method comparing each real-time (RT) observation method with image (IMG) scores

Observation type	Bias	Upper limit	Lower limit
RT-interval$_{10}$	-0.09	0.46	-0.63
RT-interval$_5$	-0.11	0.55	-0.65
RT-interval$_2$	-0.14	0.43	-0.71
RT-point$_{10}$	-0.07	0.49	-0.63
RT-point$_5$	-0.08	0.47	-0.63
RT-point$_2$	-0.09	0.50	-0.68

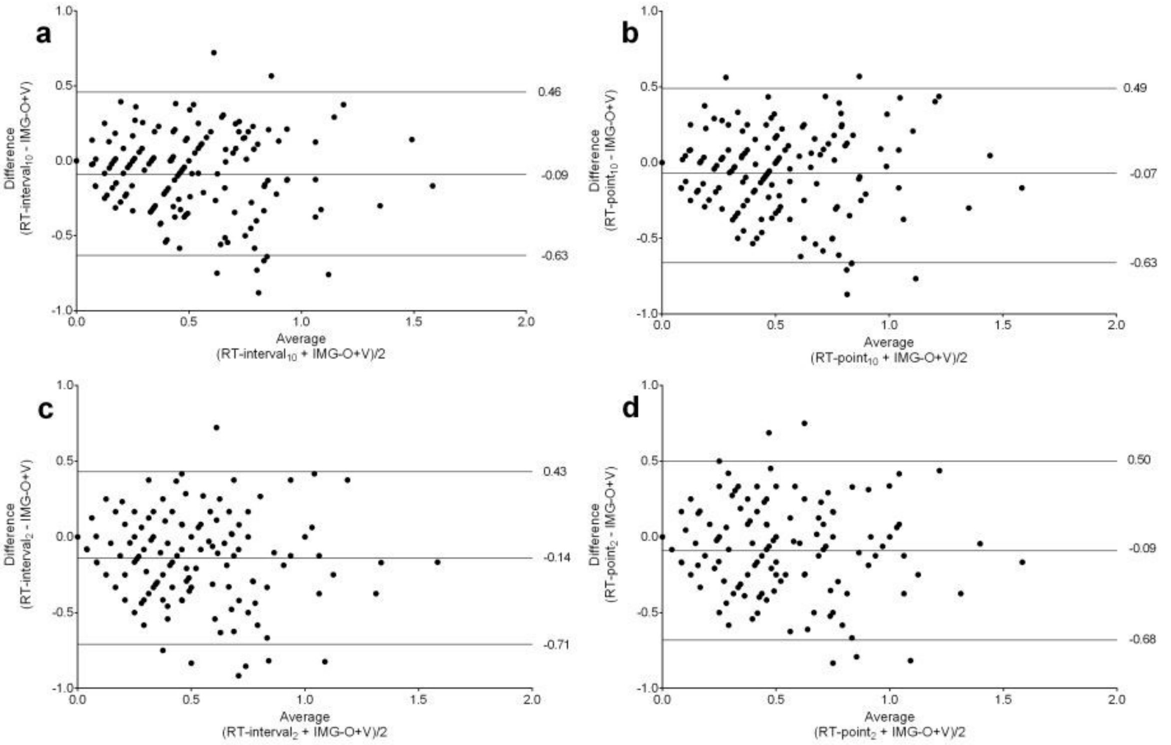
Legend: Bias is the mean difference between RT and IMG Rat Grimace Scale scores. Upper and lower limits of agreement are mean difference ± 2 SD.

Figure 2.1.3.1.2: Bland and Altman plots comparing image and real-time scores.



Legend: The Bland-Altman analysis indicates that the limits of agreement between (A) real-time observation over 5 minutes (RT-interval5) with a bias (underestimation) by real-time scores of -0.11 and limits of agreement ranging from -0.65 to 0.44. (B) Real-time point observation over 5 minutes (RT-point5) with a bias (underestimation) by real-time scores of -0.08 and limits of agreement ranging from -0.63 to 0.50.

Fig. 2.1.3.1.3: Bland and Altman plots comparing real-time (RT) scoring (RT-interval$_{10,2}$ and RT-point$_{10,2}$) to image based (IMG) scores



Legend: (a) RT-interval10 (b) RT-point10, (c) RT-interval2, (d) RT-point2. Data are mean differences (bias, central horizontal line) and limits of agreement (bias ± 2 SD, upper and lower horizontal lines)

Most (4/6) of the real-time observation methods, including all of the interval observation methods, were able to discriminate between saline and analgesic treatments (Fig. 2.1.3.1.4, 2.1.3.1.5). Buprenorphine and the multimodal treatments provided effective analgesia with significant reductions in RGS scores. Coinciding with an expected peak in carrageenan-induced pain at 6 hours (De Rantere *et al.*, 2016), buprenorphine and multimodal analgesia were effective at reducing RGS scores compared with saline in the IMG-O+ V (buprenorphine, $p < 0.0001$, 95% CI: 0.33 to 0.87; multimodal, $p = 0.0003$, 95% CI: 0.19 to 0.74, Fig. 2.1.3.1.4a), RT-interval$_{10}$ (buprenorphine, $p = 0.03$, 95% CI: 0.02 to 0.52; multimodal, $p = 0.004$, 95% CI: 0.09 to 0.60, Fig. 5b), RT-point$_{10}$ (multimodal, $p = 0.02$, 95% CI: 0.05 to 0.59, Fig. 2.1.3.1.4c), RT-interval$_{5}$ (buprenorphine, $p = 0.005$, 95% CI: 0.08 to 0.56; multimodal, $p = 0.001$, 95% CI: 0.13 to 0.61, Fig. 2.1.3.1.4d). The same pattern was observed at 9 hours in the RT-interval$_{10}$ (buprenorphine, $p = 0.02$, 95% CI: 0.03 to 0.54, multimodal, $p = 0.01$, 95% CI: 0.06 to 0.56, Fig. 2.1.3.1.4b), RT-

$point_{10}$ (multimodal, p = 0.007, 95% CI: 0.08 to 0.62, Fig. 2.1.3.1.4c) and RT-$interval_5$ (buprenorphine, p = 0.002, 95% CI: 0.12 to 0.60, multimodal, p = 0.02, 95% CI: 0.03 to 0.51, Fig. 2.1.3.1.4d). At 9 hours the IMG-O+ V method identified a decrease in RGS scores associated with buprenorphine compared with saline (p < 0.0001, 95% CI: 0.23 to 0.78) and multimodal analgesia (p = 0.04, 95% CI: 0.01, 0.54, Fig. 2.1.3.1.4a). Fewer differences were observed at 3 and 24 hours, consistent with the expected time course of carrageenan-induced inflammation. No analgesic effects were identified with RT-$point_5$ (F = 2.73, df 2, p = 0.08, Fig. 2.1.3.1.4e). Ability to discriminate between saline and analgesic treatment groups were identifiable with RT-$interval_2$ but not RT-$point_2$ (Fig. 2.1.3.1.5).

When comparing the RT-point observations with IMG-O+ V, the expected pattern of RGS scores with different treatments is present (Fig. 2.1.3.1.6).

2.1.4.2. Single interval and point observation scoring methods.

The random selection of 5 interval and 5-point observations illustrated that the predicted time course of pain for each treatment group was present but substantial variability was observed between individual scores (Figs 2.1.3.1.7 and 2.1.3.1.8).

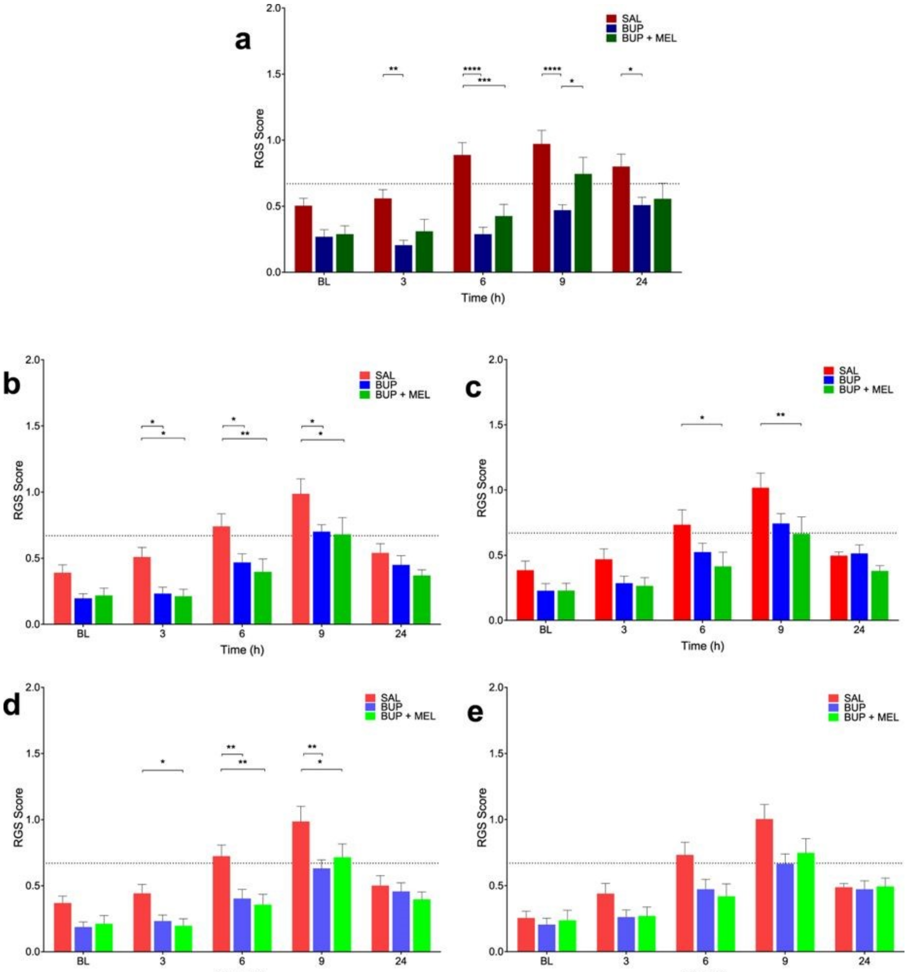
2.1.4.3. Observer effect

The presence of the observer did not significantly affect the RGS scores from the saline (F = 1.27, df 2, p = 0.30; Fig. 2.1.3.2.1a) and multimodal analgesia treatment groups (F = 1.37, df 2, p = 0.28, Fig. 2.1.3.2.1c). Unexpectedly, significant differences were observed at 24 h in the buprenorphine group between observation periods V1 and V2 (p < 0.0001, 95% CI: 0.17 to 0.56) and between IMG-O+ V and V2 (p = 0.01, 95% CI: 0.05 to 0.44, Fig. 2.1.3.2.1b).

2.1.4.4. Control groups

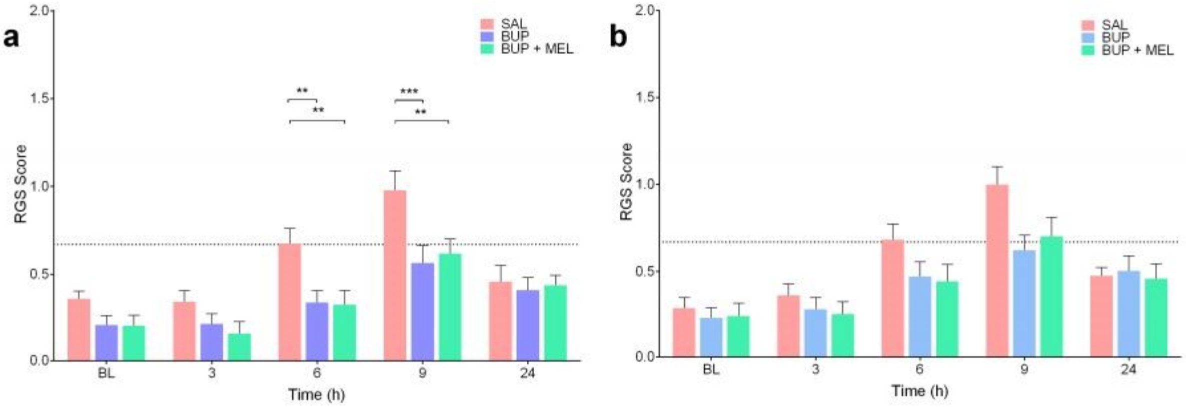
None of the control treatments resulted in significant changes to RGS scores compared with baseline values (Table 2.1.2.5.1).

Fig. 2.1.3.1.4: Both standard Rat Grimace Scale (RGS) and real-time interval RGS scoring were able to discriminate between saline and analgesia treatment groups.



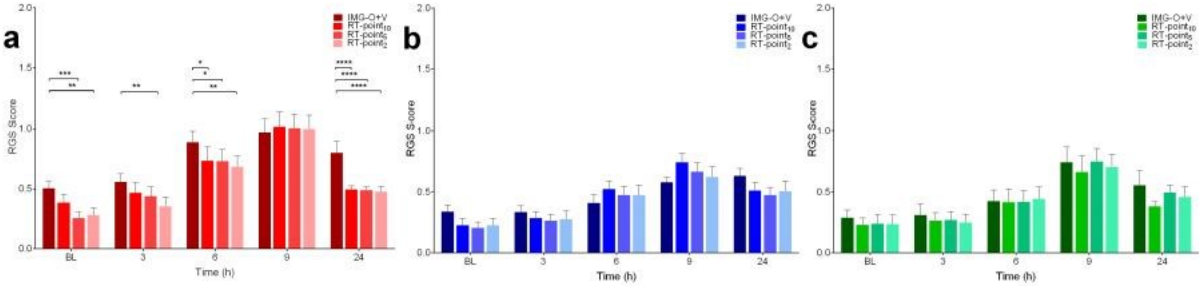
Legend: A) Standard (video-based) RGS scoring (IMG-O + V). Lower RGS scores were observed in the buprenorphine treatment group at 3 (p = 0.007), 6 (p < 0.0001), 9 (p < 0.0001) and 24 h (p = 0.03). RGS scores were reduced in the multimodal treatment group at 6 h (p = 0.0003) and a difference was observed between buprenorphine and multimodal treatment groups at 9 h (p = 0.04). (**B**) Real-time interval observation over 10 minutes (RT-interval10). RGS scores were lower in the buprenorphine group at 3 (p = 0.03), 6 (p = 0.03), and 9 h (p = 0.02). Similarly, multimodal analgesia (buprenorphine and meloxicam) resulted in a decrease in RGS scores at 3 (p = 0.02), 6 (p = 0.004) and 9 h (p = 0.01). (**C**) The real-time point observation over 10 minutes (RT-point10) identified a treatment effect in the multimodal treatment group at 6 h (p = 0.02) and 9 h (p = 0.007). (**D**) Real-time interval observation over 5 minutes (RT-interval5) showed that buprenorphine and multimodal analgesia were associated with a decrease in RGS scores at 6 h (buprenorphine, p = 0.005; multimodal, p = 0.001) and 9 h (buprenorphine, p = 0.002; multimodal, p = 0.02). RGS scores were also lower in the multimodal group at 3 hours (p = 0.04). (**E**) Realtime point observation over 5 minutes (RT-point5) did not identify analgesia treatment effects (p = 0.08). SAL = saline, BUP = buprenorphine, MEL = meloxicam. Data are mean ± SEM. Broken horizontal line represents a previously derived analgesic intervention threshold (Oliver *et al*., 2014).

Fig. 2.1.3.1.5: Treatment effects identified with RT-interval$_2$.



Legend: Treatment effects were identifiable with RT-interval$_2$ but not RT-point$_2$. (a) The real-time interval observation over 2 minutes (RT-interval2) identified treatment effect in the buprenorphine and multimodal treatment groups from saline treatment group at 6h (buprenorphine, p = 0.005, 95% CI: 0.09 to 0.60; multimodal, p = 0.003, 95% CI: 0.10 to 0.61) and 9h (buprenorphine, p = 0.0005, 95% CI: 0.16 to 0.67; multimodal, p = 0.003, 95% CI: 0.11 to 0.62). (b) Real-time point observation over 2 minutes (RT-point2) was not able to discriminate between analgesia and saline treatment groups (p = 0.19). SAL = saline, BUP = buprenorphine, MEL = meloxicam. Data are mean ± SEM. Broken horizontal line represents a previously derived analgesic intervention threshold (Oliver *et al*., 2014).

Fig. 2.1.3.1.6: Real-time (RT) point scoring methods compared to standard (video-based) Rat Grimace Scale (RGS) scoring (IMG O+V)



Legend: (a) Scoring method was found to differ significantly in the saline treatment groups: IMG-O+V vs. RT-point10 at 6 (p = 0.04, 95% CI: 0.01 to 0.31) and 24h (p < 0.0001; 95% CI: 0.16 to 0.46), IMG-O+T vs. RT-point2 at BL (p = 0.002; 95% CI: 0.07 to 0.37), 3 (p = 0.005; 95% CI: 0.05 to 0.36), 6 (p = 0.005; 95% CI: 0.05 to 0.36) and 24h (p<0.0001; 95% CI: 0.17 to 0.47). There was no effect of scoring method on RGS scores in (b) buprenorphine (F = 1.38, df 3, p = 0.27) and (c) multimodal treatment groups (F = 2.89, df 3, p = 0.05). BL = baseline. Data are mean ± SEM.

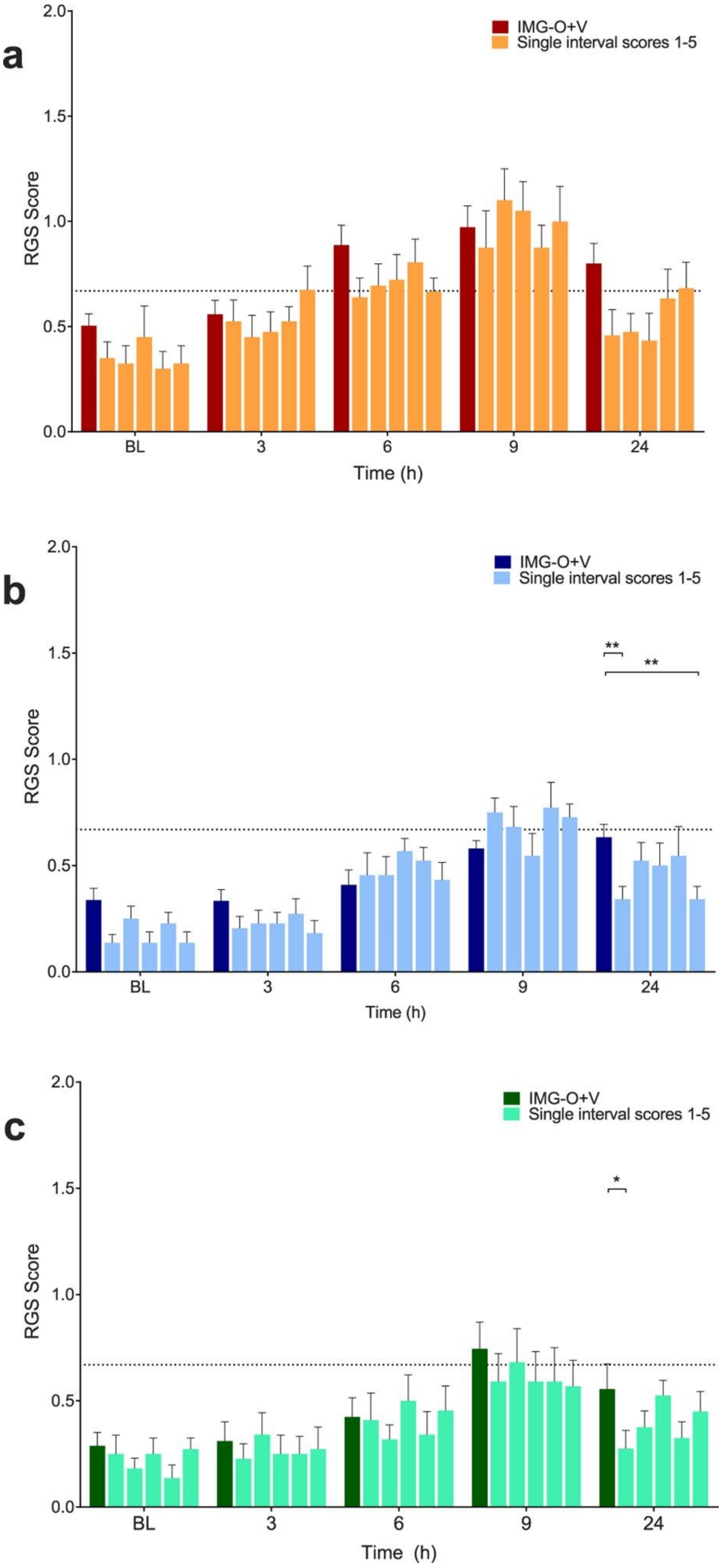
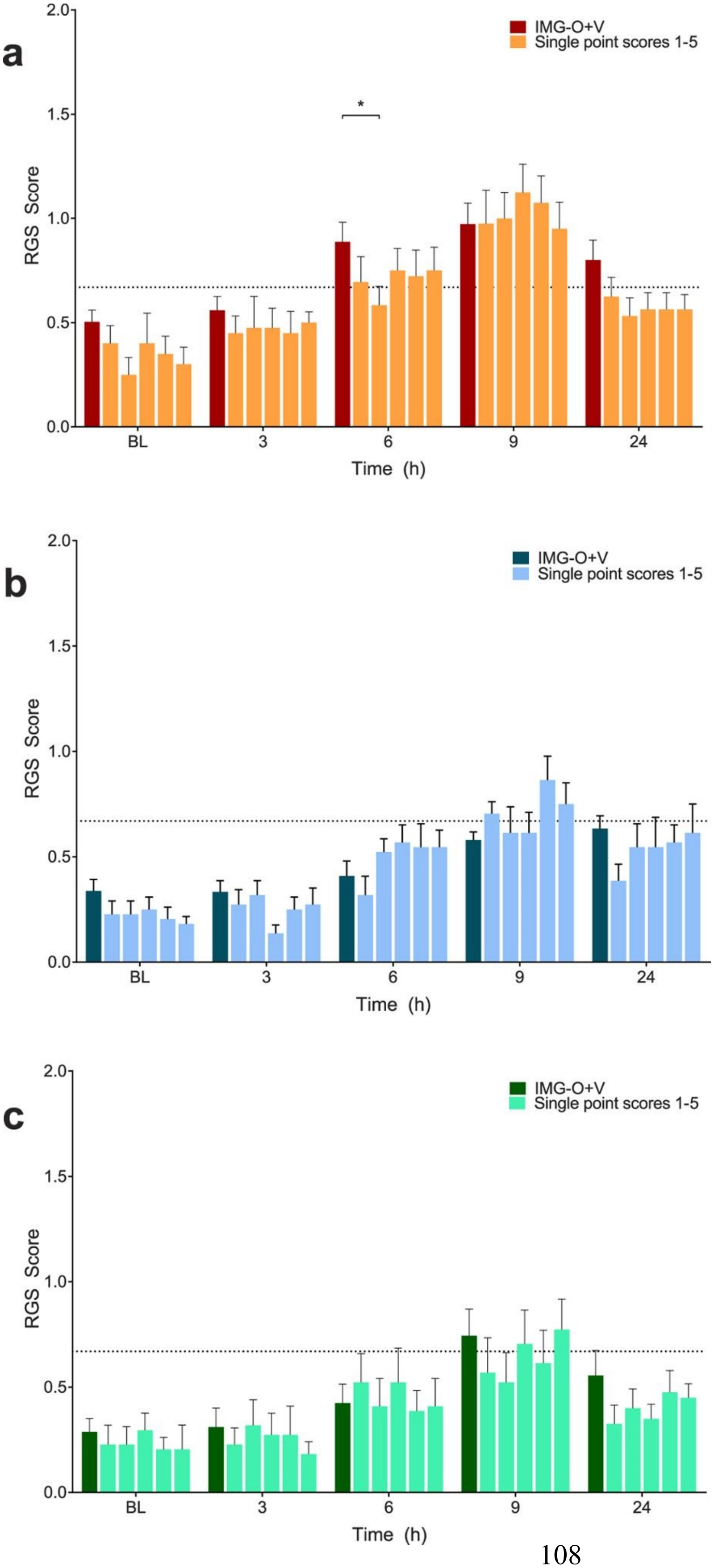


Fig. 2.1.3.1.7: Single rea-time interval scores.

Legend: Single real-time interval scores (scores 1-5) approximates the expected time course associated with each treatment, but visual inspection of the data reveals substantial variability between scores. (A) saline treatment group. There was no main effect of treatment (p=0.11). (B) Buprenorphine treatment group. A significant difference between scores was observed at 24 hours (p=0.003). (c) Multimodal treatment group. A significant difference was observed at 24 hours (p=0.03). Data are mean ± SEM. Broken horizontal line represents a previously derived analgesic intervention threshold (Oliver *et al*., 2014)

Fig. 2.1.3.1.8: Single real-time point scores.



Legend: Single real-time point scores (scores 1-5) approximates the expected time course associated with each treatment, but visual inspection of the data reveals substantial variability between scores. No man effects for scoring method were identified in the buprenorphine ((B) p = 0.13) and multimodal ((C) p = 0.16) treatment groups. A single difference was observed at 6 hours in the saline group ((A) p = 0.16). Data are mean ± SEM. Broken horizontal line represents a previously derived analgesic intervention threshold (Oliver *et al*., 2014).

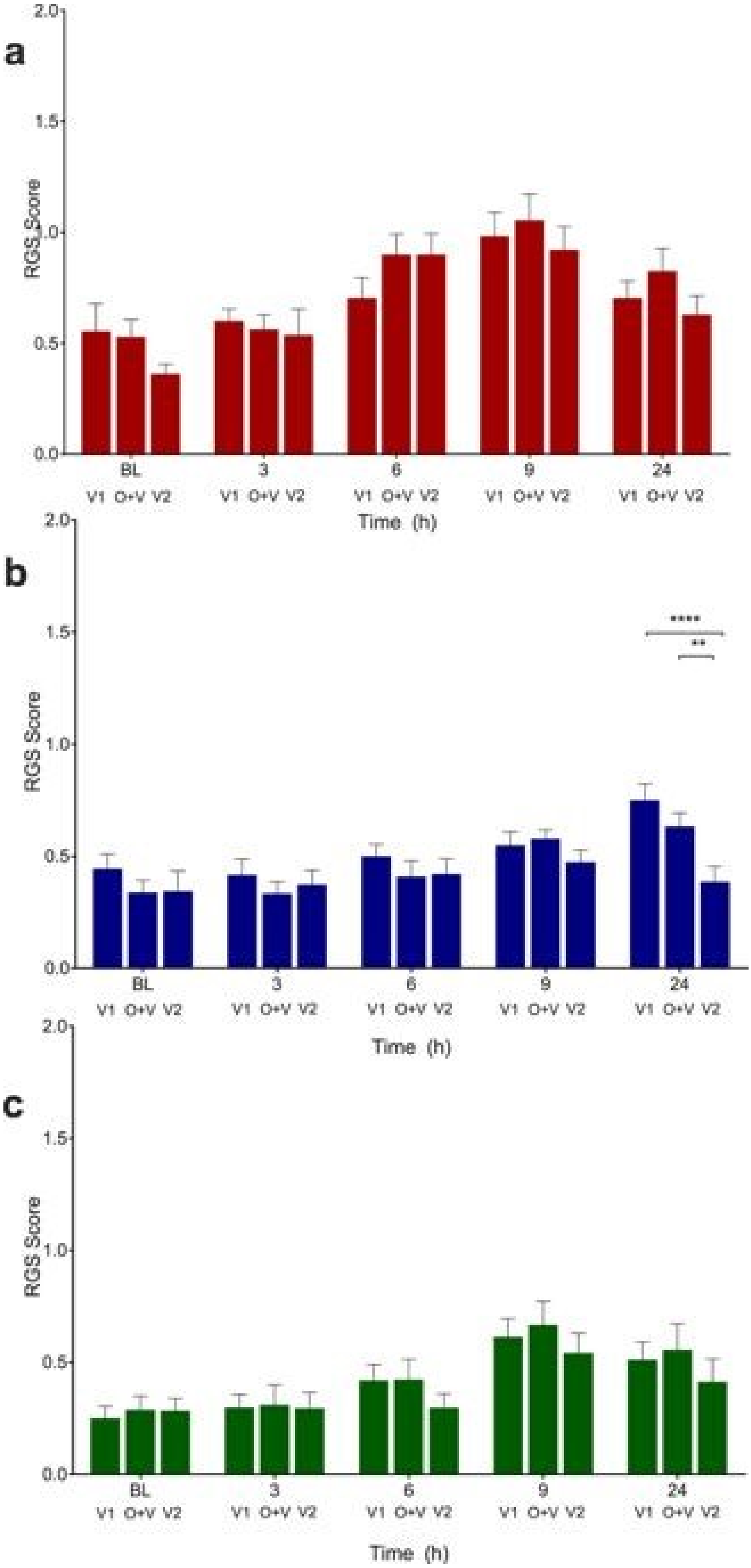


Figure 2.1.3.2.1: No observer effect was observed.

Legend: No observer effect was observed in the saline (**A**) p = 0.30) and multimodal treatment groups (**C**) p = 0.28). A significant difference between observation periods was present in the buprenorphine group (**B**) at 24 hours, between V1 and V2 (p < 0.0001) and between IMG-O+V and V2 (p = 0.01). V1 and V2 = video only, no observer present. O+V = video, with observer present. Data are mean ± SEM.

Table 2.1.2.5.1: Rat Grimace Scale scores from control group. Scores generated by standard (video-based) method.

	Treatment							
Treatment	BL Median (IQR)	3 Median (IQR)	p-value	6 Median (IQR)	p-value	9 Median (IQR)	p-value	24 Median (IQR)
A	0.40 (0.34, 0.48)	0.38 (0.33, 0.67)	> 0.99	0.63 (0.44, 0.75)	0.47	0.36 (0.27, 0.41)	< 0.99	0.40 (0.22, 0.67)
		0.25		0.31				
B	0.13 (0.08, 0.17)	0.25 (0.19, 0.31)	0.88	0.31 (0.19, 0.47)	0.29	0.28 (0.06,0.36)	0.88	0.57 (0.31, 0.91)
C	0.17 (0.04, 0.33)	0.50 (0.23, 0.71)	> 0.99	0.42 (0.25, 0.58)	0.23	0.54 (0.31, 0.83)	0.10	0.63 (0.13, 0.75)
D	0.25 (0.19, 0.31)	0.29 (0.19, 0.83)	> 0.99	0.21 (0.10, 0.44)	> 0.99	0.46 (0.25, 0.82)	0.72	0.31 (0.25, 0.66)
E	0.29 (0.02, 0.50)	0.50 (0.20, 0.65)	0.71	0.50 (0.23, 0.65)	0.47	0.42 (0.23, 0.57)	> 0.99	0.31 (0.25, 0.53)
F	0.19 (0.13, 0.31)	0.50 (0.25, 0.63)	0.29	0.46 (0.10, 0.81)	0.72	0.33 (0.27, 0.46)	> 0.99	0.38 (0.25, 0.78)

Legend: Scores at 3, 6, 9 and 24h were compared to their BL scores. A: saline (buprenorphine volume) + saline (meloxicam volume) + anesthesia + intra-plantar saline; B: buprenorphine + anesthesia + intra-plantar saline; C: buprenorphine + meloxicam + anesthesia + intra-plantar saline; D: buprenorphine + anesthesia; E: buprenorphine and meloxicam. There were no significant differences in RGS scores at the various time points compared to their baseline scores. BL = baseline. IQR = interquartile range.

2.1.4.5. Pica

There was no evidence of pica behaviour from necropsy examination or masses of petri dishes in the treatment groups (Table 2.1.3.3.1). The buprenorphine control groups exhibited a small amount of pica behaviour (petri dish weight changes of 0.1–0.6 g, Table 2.1.3.3.2).

Table 2.1.3.3.1: Weights of petri dishes in the three treatment groups

Treatment group	Cage #	Petri dish weight (g)		
		Before (BL)	After	difference
Saline	1	8.6	8.6	0
	2	2.7	2.7	0
	3	6.9	6.9	0
	4	8.5	8.5	0
	5	8.4	8.4	0
	6	7.2	7.2	0
Buprenorphine	1	8.6	8.6	0
	2	6.7	6.7	0
	3	-	-	0
	4	8.2	8.2	0
	5	8.5	8.4	0.1
	6	8.2	8.2	0
Buprenorphine + Meloxicam	1	8.6	8.6	0
	2	8.0	8.0	0
	3	8.7	8.7	0
	4	8.6	8.6	0
	5	6.0	6.0	0
	6	7.8	7.8	0

Legend: No non-food items were found in the stomach of all the rats and the weight of petri dishes did not changes from their baseline weights. Variations in weights of petri dishes "before" reflects chewing which occurred during habituation. BL = baseline. Difference = before (BL) – after.

Table 2.1.3.3.2: Weights of petri dishes in the control groups

Treatment group	Cage #	Crossover #	Petri dish weight (g)		
			Before (BL)	After	difference
A	1	1	7.8	7.8	0
	3	2	7.7	7.7	0
B	2	2	6.3	6.0	0.3
	4	2	8.1	-	-
C	3	1	7.7	7.7	0
	1	3	5.5	5.5	0
D	2	1	7.9	7.9	0
	4	1	7.9	7.9	0
E	2	3	5.1	4.5	0.6
	4	3	4.2	3.8	0.4
F	1	2	8.0	7.8	0.2
	3	3	6.3	6.2	0.1

Legend: (A) Saline (buprenorphine volume) + saline (meloxicam volume) + anesthesia + intraplantar saline. (B) buprenorphine + anesthesia + intraplantar saline (C) buprenorphine + meloxicam + anesthesia + intraplantar saline (D) buprenorphine + anesthesia (E) buprenorphine and (F) buprenorphine and meloxicam). No non-food items were found in the stomach of all the rats and unlike the rats given intra-plantar carrageenan (Table 2.1.3.3.1), weights of petri dishes did decrease from their baseline weight when the rats were given buprenorphine. Variation in weights of petri dishes "before" reflects chewing which occurred during habituation. BL = baseline. Difference = before (BL) – after.

2.1.5. Discussion

The appeal of real-time application of rodent grimace scales lies in expanding their current role as retrospective research instruments to one allowing early identification of pain, facilitating timely intervention and improving the welfare of laboratory rodents. The potential for rodent grimace scales to be applied as a real-time scoring system has been previously suggested (Langford *et al*., 2010; Oliver *et al*., 2014 and Roughan *et al*., 2016) and attempted with limited success in mice (Miller and Leach, 2015 and Faller *et al*., 2015). We have shown that real-time RGS scoring is an accurate and feasible alternative to the standard method described by Sotocinal *et al*. (2011), offering a refinement to the humane care of laboratory rats. The ability of a new method to reflect changes identified by the current (criterion) standard shows accuracy and construct validity. In evaluating different methods of real-time scoring, we identified multiple 15 s interval observations as more sensitive than multiple point observations. And we observed that single observations, both interval and point, approximated the predicted time course of pain, but exhibited substantial variability. Applying the Bland and Altman method to our data allowed assessment of systematic differences between observation methods and the variability around these differences. There was a small systematic underestimation by all the real-time methods, showing that on average, real-time scores are very close to image-generated scores. The similarity between 5 and 10-minute real-time observation periods indicates that 10-minute observation periods are unnecessary if the RGS is being applied as a tool to guide pain management (rather than as a research tool). Furthermore, the similarity between RT-interval$_5$ and RT-point$_5$ observations offers alternative means of scoring depending on user preference. The acceptability of a new (real-time) technique over a criterion standard (image-based) depends on a subjective assessment of the limits of agreement. For RT-interval$_5$ and RT-point$_5$ observations, the limits of agreement span a 0.5 score range either side of the bias. Therefore, there is the possibility of a single observation either over or underestimating the true score. Furthermore, the Bland and Altman plots show that data variability increases at RGS scores > 0.5. Interpreting these observations together, a practical approach could be a planned reassessment of any animal with an initial RGS score > 0.5 within a relatively short period (*e.g.* 1 hour), taking in to account the potential for suffering if providing analgesia is delayed against any side-effects associated with analgesic use. As RGS scores exceed a previously identified threshold for intervention (RGS score > 0.67; Oliver *et al*., 2014), the likelihood of an animal experiencing pain increases, in which case the reassessment interval should be kept short or analgesia provided immediately, and the animal reassessed for an improvement in RGS score. The agreement between RT scores and IMG scores was not reflected in their ability to discriminate treatment effects statistically as observations decreased to 2 minutes. Both interval and point observation methods (RT-interval$_{10}$ and RT-point$_{10}$) were able to discriminate between the saline and analgesic treatments at the 6- and 9-hour time points, when peak RGS scores are expected (Radhakrishnan *et al*., 2003 and De Rantere *et al*., 2016) and did not differ significantly from the standard RGS scoring method. Furthermore, the mean scores at these times exceeded a proposed analgesic intervention threshold (Oliver *et al*., 2014), providing evidence for the relevance of this decision-making tool. However, when the observation period was decreased to 5- or 2-minutes (RT-interval$_{5,2}$ and RT-point$_{5,2}$) only the interval scoring methods were able to reliably discriminate between saline and analgesia treatment groups, though the pattern of RGS scores did exhibit the expected time courses of the different treatment groups. This inability to discriminate was likely due to insufficient power when scoring with RT-point$_{5,2}$ as the Bland and Altman results showed similar agreement to the equivalent interval scoring methods. Our findings agree with those of Ballantyne *et al*. (1999), where a multidimensional 7 item pain scale, of which 3 items were facial action units, was evaluated in neonatal infants during painful and non-painful procedures (Ballantyne *et al*., 1999). The authors showed that real-time (bedside) observations (over a 45 s period) did not differ significantly from the standard video-based assessments and were able to discriminate between predicted painful and non-painful states. This assessment method is similar to the successful interval method we employed. Faller *et al*. (2015) successfully used the mode of observed scores (scored from 10 photographs taken over a 15–20-minute observation period) to identify a reduction in the MGS score following buprenorphine administration (Faller *et al*., 2015). This approach resembles our point observations, though the discriminatory ability identified differs from our findings with the RT-point$_{10}$ observation method, where 18 observations were recorded over a 10-minute period. However, a direct comparison between studies is limited by differences in the time allowed to perform the scoring (photograph *versus* live observation), species and grimace scales (the number of facial action units differs between the RGS and MGS).

The similarity in RGS scores we observed between RT-interval and standard RGS methods differs from the findings of Miller and Leach (2015) where they reported, using the MGS, that real-time scores were significantly lower than image scores in 6/7 comparisons (across strain and gender). Their real-time scoring was based on 3 × 5 s observations during a 10-minute observation period and image scores were derived from 3 randomly selected photographs taken during the same 10-minute period. Our RT-interval$_2$ and RT-point$_2$ observations at baseline provide the closest comparison to this study as the mice studied did not receive potentially painful interventions. While our results showed no significant differences between these observation types and the standard RGS method, only interval observations were capable of differentiating treatment effects. As suggested by the authors, the use of photographs to generate MGS scores may have resulted in an artificial elevation of scores by capturing behaviours interfering with scoring (such as blinking). A comparison with the standard RGS scoring method (Sotocinal *et al*., 2015) allowed evaluation of this possibility. Single observations with both the RT-interval and RT-point methods displayed the predicted time course for each treatment group, with RGS scores in the saline group exceeding a proposed threshold for analgesic intervention at 9 hours, in contrast to the buprenorphine and multimodal groups (Oliver *et al*., 2014). However, visual inspection of the data revealed substantial variability with both observation methods, indicating that reliance on a single observation for treatment decisions is insufficient, with the risk of failing to identify a painful state.

Buprenorphine was an effective analgesic, limiting the predicted increase in RGS scores at 6 and 9 hours after carrageenan administration (De Rantere *et al*., 2016 and Radhakrishnan *et al*., 2003). The timing of buprenorphine administration may have resulted in its analgesic effects waning around the 9-hour time point (Roughan and Flecknell, 2004), explaining the slight increases in RGS scores observed at this time in the buprenorphine and multimodal groups. The optimal dosing interval for buprenorphine in rats is unclear and is likely to vary according to procedure and strain, highlighting the importance of regular pain assessment with an appropriate instrument (Roughan and Flecknell, 2001; 2004 and Schaap *et al*., 2012). The choice of a 0.03 mg/kg dose was based on recent work showing its efficacy when evaluated with the RGS9. A dose of 0.05 mg/kg may have provided a longer duration of analgesia (Roughan and Flecknell, 2004) but has been associated with pica behaviour (Clark *et al*., 1997 and Schaap *et al*., 2012). Therefore, the lower dose was selected to minimise the possibility of pain from pica behaviour acting as a confounding factor.

Somewhat unexpectedly, the multimodal treatment group (buprenorphine and meloxicam) exhibited similar RGS scores to the buprenorphine treatment group at all time points, when it might be expected that a multimodal analgesic approach with a non-steroidal anti-inflammatory agent (NSAID) and opioid resulted in lower RGS scores (Ong *et al*., 2005; Rialland *et al*., 2012 and Ciuffreda *et al*., 2014). There are several interpretations of these findings. Firstly, the addition of meloxicam may not have conferred any additional benefit as the RGS scores were already low and below a level identified as painful (Oliver *et al*., 2014). Secondly, the relationship between inflammation and pain may be less clear than previously believed. Meloxicam may reduce inflammation without a concurrent decrease in pain (Bianchi *et al*., 2002 and Roughan *et al*., 2016). However, this contradicts a substantial body of evidence that NSAIDs are effective analgesics in rats (Engelhardt *et al*., 1995; Roughan and Flecknell, 2003; 2004 and Roughan *et al*., 2004) though the relationship between the behavioural (postural) pain scale used in those studies and the RGS is undefined. Finally, the RGS may not be sensitive enough to identify subtle variations in pain intensities. This is possible as original work validating the RGS used the potent opioid morphine to demonstrate analgesic sensitivity (construct validity) in several robust pain models (Sotocinal *et al*., 2011).

RGS scores were similar between observation periods (V1, O+ V, V2), indicating that the presence of an observer had negligible impact. The extent to which this lack of effect was related to the observer being female is unknown: a systematic effect of observer gender has been recently shown in mice, with a reduction in MGS scores in the presence of men as a result of stress-induced analgesia (Sorge *et al*., 2014). The exception to the general case was the difference observed between observation periods at 24 hours in the buprenorphine group. This is unlikely to be an 'observer effect' as this difference was limited to a single treatment group and time point. Furthermore, if an observer effect was present, RGS scores from V1 and V2 periods would be expected to be similar, and different from those generated during O + V.

Scoring by an observer involved with the study raised the possibility of observer bias as it was not possible to blind to time point. This may have affected the real-time RGS scores at baseline and 24 hours, when RGS scores would be predicted to be low for this model. This possibility was addressed by comparing real-time scores with those generated from randomised, blinded images. Without concurrent video-recording, observer bias cannot be accounted for unless the observer has no knowledge of the study design. This may reflect the situation encountered if real-time RGS scoring were to be used by technicians or veterinarians not involved with a study.

We have shown that the RGS can be successfully applied with real-time observations, lending itself to use as a rapid pain assessment tool to identify acute pain in rats. Interval observations over a 2-minute period was able to discriminate between treatment effects whereas point observations displayed lower sensitivity and were unable to discriminate between treatments. Single observations, interval or point, showed substantial variability and should not be used to determine analgesic administration without planned reassessment. The best balance between practicality and accuracy is achieved with 5-minute observation periods with either interval or point observations. When using real-time observations, we suggest implementing planned reassessments to account for score variability, particularly as RGS scores exceed 0.5. However, the decision to administer analgesia should be balanced against the welfare cost of delaying intervention for reassessment.

2.1.6. Acknowledgements

This study was supported by a Natural Sciences and Engineering Research Council (NSERC) Discovery Grant 424022-2013 (DSJP), VL receives stipend support from the NSERC Discovery Grant and a Queen Elizabeth II Graduate Scholarship (University of Calgary), EZ received stipend support from the Canadian Association of Laboratory Animal Medicine and Canadian Association of Laboratory Animal Science Research Fund. The authors wish to thank the technical staff of the Animal Resource Centre (University of Calgary) for study support, Cassandra Klune, Chelsea Schuster and Jessica Pang for critical reading of the manuscript, and Dr Grace Kwong for statistical support (Bland and Altman analysis).

2.2. 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

Vivian Leung[3], Marie-Odile Benoit-Biancamano[3], Daniel SJ Pang[3]

2.2.1. Abstract

Introduction: The Rat Grimace Scale (RGS), a facial expression scale, quantifies the affective component of pain in rats. The RGS was developed to identify acute and inflammatory pain and applicability in acute and chronic visceral pain is unknown. The dextran sulfate sodium (DSS) colitis model is commonly used in rats but pain is rarely assessed, instead, disease progression is monitored with the Disease Activity Index (DAI; assessing fecal blood, stool consistency and weight loss). The aim of this study was to assess if the RGS and two additional behavioral tools (Composite Behavior Score (CBS) and burrowing) could identify pain in an acute and chronic DSS colitis model.

Methods: Male and female Sprague Dawley rats were block randomized to: 1) acute colitis (four days DSS in drinking water); 2) chronic colitis (four days DSS, seven days water, three days DSS) or 3) control (14 days water). DAI, RGS, CBS and burrowing assessments were performed daily.

Results: RGS scores increased as DAI scores increased during both acute and chronic phases. Burrowing only decreased during the acute phase. In contrast, CBS scores did not increase significantly during either colitis phase.

Conclusions: These data show that the RGS and burrowing identify acute and chronic visceral pain and that variables assessed in the DAI are indicative of pain. This suggests that the RGS can be applied to a wider range of pain types and chronicity than originally suggested. These findings increase the application of the RGS as a pain scale and welfare improvement tool.

[3] Faculté de Médecine Vétérinaire, Université de Montréal, Saint-Hyacinthe, Québec, Canada

2.2.2. Introduction

In recent years, it has been proposed that spontaneous behaviors of animals should be used to assess pain in animals (Mogil *et al.*, 2010). In laboratory rats, one of these behavioral tools is the Rat Grimace Scale (RGS), a facial expression scale, which was developed with acute inflammatory pain models (Sotocinal *et al.*, 2011). Since its initial development, performance of the RGS in acute inflammatory pain models has been confirmed (De Rantere *et al.*, 2016 and Leung *et al.*, 2016) and its application in other acute and neuropathic pain models has been described (Liao *et al.*, 2014 and Akintola *et al*., 2017). Development of the Mouse Grimace Scale (MGS) identified a limited ability of this scale to identify pain in classic models of neuropathic pain (chronic constriction injury and spared nerve injury), but there has been little investigation of chronic pain using other grimace scales, including the RGS (Langford *et al*., 2011 and Akintola *et al*., 2017). Furthermore, a study of induced acute visceral mucositis failed to identify significant changes in the RGS (Whittaker *et al.*, 2016). Therefore, it is currently unclear what role the RGS may play in the evaluation of chronic or visceral pain. Potential alternative, or complementary, methods to the RGS include a Composite Behavior Score (CBS) and burrowing behavior (Roughan and Flecknell, 2003 and Andrews *et al*., 2012). The CBS, which uses an ethogram, including twitching, writhing and back-arching behaviors, has been employed successfully to assess visceral pain in laparotomy and mucositis models (Roughan and Flecknell 2003 and Whittaker *et al.*, 2016). Burrowing, as an expression of voluntary behavior, is performed by a high proportion of laboratory rats (Andrews *et al.*, 2016). It has been successfully applied in models of induced osteoarthritis and found to be robust in multi-center testing (Wodarski *et al*., 2016).

The dextran sulfate sodium (DSS) colitis model is well characterized and widely used to study colitis in mice and rats (Okayasu *et al*., 1990; Cooper *et al*., 1993 and Gaudio *et al*., 1999). With a focus on underlying disease mechanisms, the assessment of pain is performed infrequently in this model despite being a common symptom of clinical disease (Farrell *et al*., 2014). Where pain is evaluated, it is typically limited to non-specific behaviors or evoked hypersensitivity testing (Tobin *et al*., 2004; Larsson *et al*., 2006; Verma-Gandhu *et al*., 2006 and Jain *et al*., 2015), measures that may not capture the pain experience (Mogil and Crager, 2004 and Smith *et al*., 2016). Model severity and progression is commonly monitored using the Disease Activity Index (DAI), which scores the presence of fecal blood, stool consistency and weight loss (Cooper *et al.,* 1993). A relationship between similar clinical signs and pain is present in people but has not been established in rodent models of colitis (Bielefeldt *et al*., 2009).

The aim of this study was to assess the performance of the RGS, CBS and burrowing as measures of acute and chronic visceral pain in a DSS-colitis rat model. We hypothesized that the RGS and CBS would increase in parallel with the DAI, with a concurrent reduction in burrowing.

2.2.3. Methods and materials

2.2.3.1. Ethical statement

All experiments were approved by the institutional animal care and use committee (Comité d'Éthique de l'Utilisation des Animaux of Université de Montréal, #Rech-1876) and performed in accordance with the Canadian Council on Animal Care guidelines.

2.2.3.2. Animals

Thirty-eight male and female Sprague-Dawley rats of at least 6 weeks of age (females (n = 18): 182g [range: 144-289g]; males (n = 20): 217g [range: 183-293g]) were obtained from Charles River Laboratories (Sherbrooke, Canada). Animals were housed singly in polycarbonate rat cages (2154F, Tecniplast, Montreal, QC, Canada) in a conventional facility. Single housing was required to facilitate daily DAI assessments (stool consistency and presence of blood). Rats had hardwood laboratory bedding (Beta Chip, Charles River Laboratories, Sherbrooke, Canada), with a plastic tube (ABS tubing, Verdun, IPEX Inc., QC, Canada) and a nylon toy for enrichment (Bio-serv Inc., Flemington, NJ, USA). They were housed in a 14:10 hour light/dark cycle with lights on at 6 am and temperature and humidity settings of 22°C and 35-50% respectively. Rats were fed laboratory rat pellets (Charles River Rodent Diet #5075, Charles River Laboratories, Sherbrooke, QC, Canada) and tap water was provided *ad libitum* before the start of the study. Rats acclimatized to their new surroundings for at least three days before habituation procedures began.

2.2.3.3. Colitis model induction

Colitis was induced by adding dextran sulfate sodium (5% DSS, J63606, Alfa Aesar, Ward Hill, MA, USA, MW 40,000) in to distilled drinking water provided *ad libitum*. The DSS solution was prepared on the day of administration (day 0). Rats were block randomized with a list randomizer (random.org) with equal allocation of sexes to one of three treatment groups: 1) Group 1 (n = 12) were given one phase of DSS (acute phase); 2) Group 2 (n = 13) were given one phase of DSS (acute phase) followed by a water phase (distilled drinking water only), then a second phase of DSS (chronic phase); and 3) Group 3 controls (n = 13) were given distilled drinking water for the duration of the experiment (Figure 1). Randomization was performed after baseline (BL) assessments. DSS treatments were stopped when all rats within each block-randomized cohort displayed signs of colitis as indicated by the DAI (i.e. decrease in stool consistency, bloody stools and weight loss), with an average DAI score of 2/4. The water phase was terminated when all rats within the cohort had DAI scores of 0 for at least 24 hours before restarting DSS treatment. Following completion of the final assessments, rats were euthanized (induction of general anesthesia with isoflurane, followed by guillotine decapitation after confirming loss of righting and pedal withdrawal reflexes): on day 4 of the acute phase for group 1, on day 3 of the chronic phase for groups 2 and the equivalent day for group 3. All assessments (DAI, RGS, CBS, burrowing) were performed during the light phase. DAI and RGS were assessed in a room adjacent to the housing room. Burrowing was assessed in the housing room.

2.2.3.4. Habituation

Before the study, all rats were habituated to the observer (VL, Figure 2.2.3.4). On the day before habituation (day -5), two pieces of food reward (Honey Net Cheerios^{TM}, General Mills, Inc., Golden Valley, MN, USA) were introduced to each cage. For four days (day -4 to -1), rats were handled by the experimenter for a minimum of 10 minutes each while offering the food reward. Rats were also habituated to the Plexiglas observation box (28 cm length x 15 cm width x 21 cm height) daily, whereby they were placed inside for a maximum of 10 minutes with a food reward.

2.2.3.5. Disease Activity Index

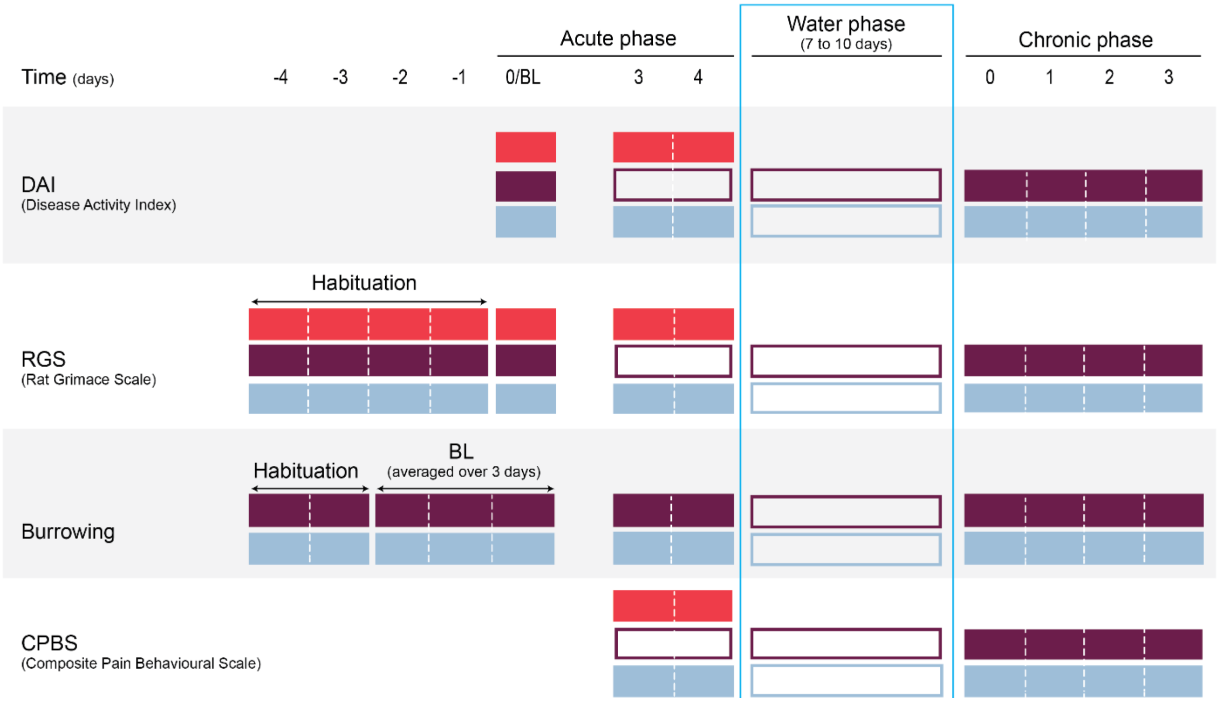
The DAI consists of three items, each scored from 0 to 4: weight loss, stool consistency and bloody stools (Table 2.2.3.5.1; Cooper *et al.*, 1993). Rats were weighed after completion of all assessments (RGS, CBS, burrowing). If gross bleeding was not evident, the presence of blood was assessed with a fecal blood slide test (Hemoccult II^{TM} Slides,60151A, Beckman Coulter, Inc., Brea, CA, USA).

Table: 2.2.3.5.1: Disease Activity Index scoring.

Score	Weight loss (%)	Stool consistency	Bloody stools
0	0	Normal	Normal
1	1-5		
2	5-10	Loose stool	Hemoccult positive
3	10-20		
4	>20	Diarrhea	Gross Bleeding

Legend: The average score is calculated from the sum of the three items: weight loss, stool consistency and bloody stools (Cooper *et al.*, 1993).

Fig. 2.2.3.3.1: Experimental timeline



Legend: Experimental timeline. (Disease Activity Index, Rat Grimace Scale, burrowing and Composite Behavior Scale). Each filled box indicates a habituation or assessment activity for each assessment method. Unfilled boxes indicate when no assessments were performed. During the acute phase, group 1 and group 2 rats were treated with 5% dextran sulfate sodium (DSS) administered in water. Group 1 rats were euthanized at the end of the acute colitis phase. During the water phase (Group 2 and controls) no assessments were made. During the chronic phase, group 2 rats were treated with 5% DSS for a second time before euthanasia on day 3. Tissue was harvested for microscopic and macroscopic analysis immediately following euthanasia. BL = baseline.

2.2.3.6. Rat Grimace Scale

The Rat Grimace Scale was scored as originally described by Sotocinal *et al*. (2011). Briefly, each of 4 action units (orbital tightening, nose/cheek flattening, ear changes and whisker changes) was assigned a score of 0, 1 or 2 based on degree of presentation.

The RGS was scored two ways: 1) in real time and 2) with video-based analysis. During real-time scoring, observations began 3 minutes after introducing the rat to the observation chamber. Facial expression was scored based on a 15s observation period repeated every 30 seconds which generated a total of 18 scores for each time point over a 9-minute timer period (Leung *et al*., 2016). Scores were averaged every three-minute interval and the resultant three scores were averaged again for a final score. For video-based analysis, video-recording took place at the same time as real-time observations. Blinded video-based scoring was performed in 'real-time' while the video was playing to assess observer bias because it was not possible to blind the observer from treatment groups and time points (Leung *et al.*, 2016). Video-based data were used for analysis. Both real-time and video scores were performed by the same observer (VL). The observer was previously trained in RGS scoring by an experienced rater (Zhang *et al.*, 2019). Real-time scoring was performed between 8 am to 12 pm and the order in which the rats were assessed was randomized each day with a list randomizer (random.org).

2.2.3.7. Burrowing

The technique described by Andrews *et al*. (2012) was followed. During two days of habituation (days -4 to -3; Figure 2.2.3.3.1), rats were placed in same sex pairs in a 53 L box (burrowing box; 58.4 cm length x 41.3 cm width x 31.4 cm height; Sterilite Corporation, Townsend, MA, USA) with the empty burrowing tube (32 cm in length x 10 cm in diameter, elevated by 6 cm at the open end of the tube with two metal legs) for 30 minutes. After 30 minutes, the burrowing tube was filled with 2.5 kg of gravel (2-5mm, Premium Aquarium Gravel, Clifford W. Estes Company, Fairfield, NJ, USA) and placed with the rats for 60 minutes. If a pair of rats did not burrow sufficiently (< 100g of gravel displaced) on the first day of habituation (day -4), a new pair was created including a burrowing rat (identified on day -4) and the protocol repeated. BL assessments were made over the next three days (days -2 to 0) with rats placed individually in the burrowing box with the gravel-filled burrowing tube for 60 minutes daily. The amounts of gravel displaced over these three days were averaged to produce a BL score for each rat. It was predetermined that rats that had a BL of less than 100g of gravel displaced would be excluded. Burrowing assessments were always performed after RGS scoring. Burrowing was assessed in group 2 and control animals during both acute and chronic phases.

2.2.3.8. Composite Behavioural Score

The composite behavior score (CBS) consisted of recording the frequency of five behaviours (writhing, vertical back arching, stagger/fall, twitch and belly pressing) as described by Roughan and Flecknell (2003) and Thomas *et al*. (2016) Writhing behavior was defined as the contraction of the abdominal muscles. Back arching was defined as a vertical stretch upwards that resembled a cat stretching. Stagger/fall behavior was defined as a rat falling over or losing its balance while moving. Twitch behavior was defined as a fleeting contraction of flank muscles. These behaviours were observed from the same video recordings used for the Rat Grimace Scale (observer blinded to treatment). The total frequency of each behavior was summed to produce a total score.

2.2.3.9. Macroscopic

Following euthanasia, abdomens were opened via a midline incision and colons removed. Macroscopic scoring consisted of body weight loss from BL, changes in colon length compared with controls, adhesion of the colon to the mesentery, length of any ulcer present, percentage of colon inflamed, presence of erythema, fecal blood, diarrhea and bowel thickness (Cluny *et al*., 2010). Ulcer length and bowel thickness were measured with digital calipers after fixation in formalin for 48 hours. The score for each item was summed to provide a total macroscopic score (Table 2.2.3.9).

2.2.3.10. Microscopic

Colons were collected and fixed in neutral buffered 10% formalin for approximately 48 hours before four samples (7mm transverse sections) were collected from the distal colon of each rat. Any ulcers identified were transected and both halves examined. Tissues were routinely processed, and slides were cut at 4 um and stained with hematoxylin-eosin-phloxine-saffron (HEPS). The microscopic assessments consisted of three items (severity of inflammation, mucosal damage and crypt damage) and the highest score used for analysis (Table 2.2.3.10; Vowinkel *et al*., 2004). Each item was then multiplied by the factor of the pathological change rate, taking into account the total surface of the affected area.

Table 2.2.3.9.1: Macroscopic scoring of colon samples

Score	0	1	2	3	4
Body weight	0%	0-5%	5-10%	10-20%	>20%
Length (% shrunk from controls)	<15%	15-25%	25-35%	>35%	
Adhesion	No adhesion	Some adhesion	Extensive adhesion		
Erythema	Absent	Present			
Fecal blood	Absent	Present			
Diarrhea	Absent	Present			
Ulcer length (cm)	Measurements of ulcers in colon				
% total length inflamed	inflamed length of colon (cm)/total length of colon (cm)				
Bowel thickness (mm)	Measured with calipers at thickest point				

Legend: Scores for individual items are summed to produce a total score (Cluny *et al*., 2010).

Table 2.2.3.10.1: Microscopic scoring of colon samples

Score	Inflammation	Mucosal damage	Crypt damage	Pathological change rate
0	None	None	None	None
1	Mild	Mucous layer	1/3	0-25%
2	Moderate	Submucosa	2/3	26-50%
3	Severe	Muscularis and serosa	100%	51-75%
4			100% with epithelium loss	76-100%

Legend: Each item is assigned a score which is then multiplied by the pathological change rate (extent of colon section affected). Resultant score from each item is summed to produce a total score. (Vowinkel *et al*., 2004).

2.2.3.11. Humane Endpoints

Humane endpoints were assessed daily and consisted of: 1) more than 15% weight loss, 2) an RGS score of 2/2 for more than 4 hours, 3) a DAI score ≥3/4 and 4) obvious lethargy. Any rat that reached an endpoint was euthanized.

2.2.3.12. Statistical analysis

Data were analyzed, and sample size estimated with commercial statistical software (Prism 6.07, GraphPad Software, La Jolla, CA, USA; MedCalc Software 18.5, Ostend, Belgium and G*Power 3.1.9.2, Germany). All data, except the CBS and pathology data (macro- and microscopic scores), approximated a normal distribution according to the D'Agostino-Pearson omnibus normality test. Comparisons between DSS-treated groups and controls were performed with a two-way ANOVA followed by a post-hoc Bonferroni test. Comparisons within groups (from baseline) were performed with a two-way ANOVA followed by a post-hoc Dunnett test (RGS, CBS, DAI, burrowing) and a Kruskall-Wallis test for microscopic and macroscopic scores (Dunn's post-hoc test). A Bland-Altman analysis of repeated measures was used to assess if RGS real-time and video scores were similar. Sample sizes were estimated for the primary outcomes of interest; the RGS, CBS and burrowing. For the RGS: a sample size of 12 animals per group was estimated based on an alpha of 0.05, power of 0.8, SD of 0.25 and a mean difference of 0.3 (Leung *et al.*, 2016). For the DAI: a sample size of 12 animals per group was estimated based on an alpha of 0.05, beta of 0.8, SD of 0.9 and a mean difference of 1.0 (Kullmann *et al.*, 2001). A p-value of < 0.05 was considered statistically significant for all comparisons. Data are presented as mean ± SD (text) or SEM (figures) with the 95%CI for the mean difference. Data supporting the results are available in an electronic repository: https://doi.org/10.7910/DVN/MLJTCV.

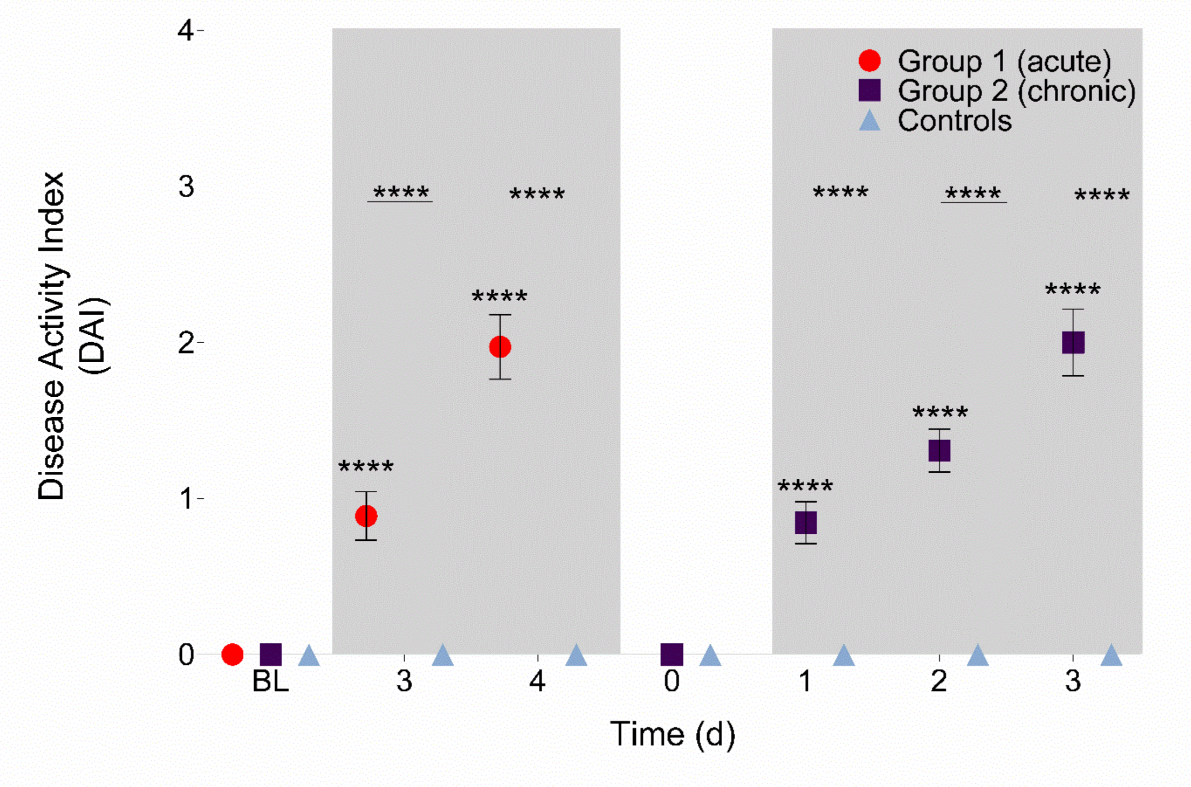
2.2.4. Results

During the first cohort of rats tested, four rats assigned to group 2 were euthanized for reaching humane endpoints on the fifth day of the acute phase. Data (RGS, DAI, CBS and burrowing) collected from these rats up to day 4 were included in the group 1 data set (except for necropsy data, which were not used). The final sample sizes were unchanged as block randomization was maintained. Due to these animals reaching their humane endpoints, the remaining rats that had no yet been treated with DSS (group 1: n = 8, group 2: n = 10, controls: n = 10) received an acute phase that lasted 4 days, the water phase lasted 7-10 days and the chronic phase lasted 3 days (group 2 rats displayed similar DAI scores as day 4 of the acute phase, average DAI 2/4). The burrowing data of one rat from the control group was excluded as it burrowed an average of 2g during BL.

2.2.4.1. Disease Activity Index

During the acute phase, there were significant main effects for treatment and time (F (1, 23) = 95, $p < 0.0001$) and F (2, 46) = 59, $p < 0.0001$ respectively) and the interaction effect was significant (F (2, 46) = 59, $p < 0.0001$). Post-hoc tests revealed that group 1 had increased DAI scores from BL and from the control group on days 3 ($p < 0.0001$, 95% CI [0.59 to 1.2]; $p < 0.0001$, 95% CI [01.2 to -0.54] respectively) and 4 ($p < 0.0001$, 95% CI [1.7 to 2.3]; $p < 0.0001$, 95% CI [-2.3 to -1.6] respectively; Fig. 2.2.4.1). During the chronic phase, there were significant main effects for treatment and time (F (1, 24) = 97, $p < 0.0001$) and F (4, 96) = 63, $p < 0.0001$, respectively) and the interaction effect was significant (F (4, 96) = 63, $p < 0.0001$). Post-hoc tests revealed that the DAI scores of group 2 animals returned to the BL score of 0 before increasing significantly from BL and from controls on chronic phase days 1 ($p < 0.0001$, 95% CI [0.58 to 1.1]; $p < 0.0001$, 95% CI [-1.2 to -0.51] respectively), 2 ($p < 0.0001$, 95% CI [1.0 to 1.6]; $p < 0.0001$, 95% CI [-1.6 to -0.97] respectively) and 3 ($p < 0.0001$, 95% CI [1.7 to 2.3], $p < 0.0001$, 95% CI [-2.3 to -1.7] respectively). Animals from the control group maintained DAI scores of zero throughout.

Fig. 2.2.4.1.1: Disease activity index scores during the acute and chronic phases



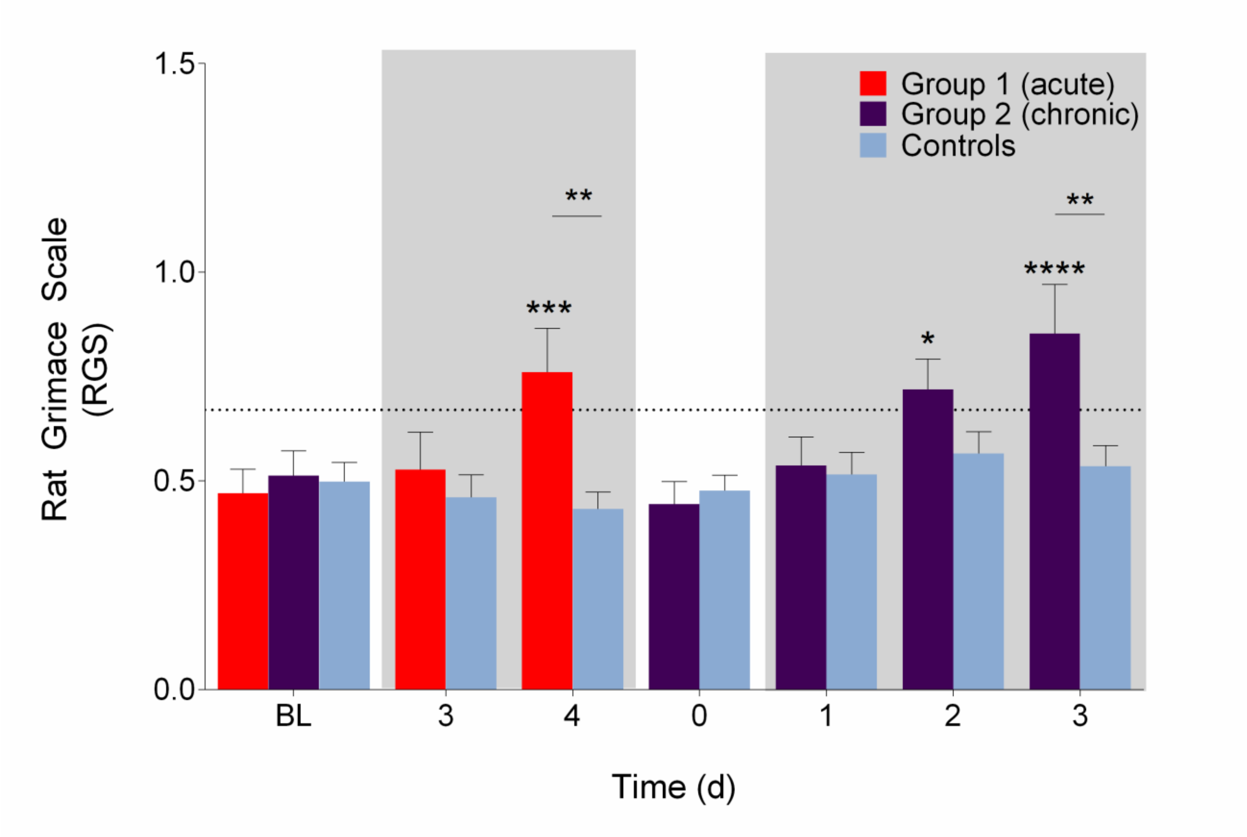
Legend: DAI scores increased significantly during acute DSS exposure compared to baseline and controls on days 3 and 4 (p < 0.0001). DAI scores increased significantly during chronic DSS exposure compared to baseline on day 0 (before DSS treatment began again) and controls on days 1, 2 and 3 (p < 0.0001). Shaded boxes represent when DSS treatment was given. **** p < 0.0001. Data presented as mean ± SEM.

2.2.4.2. Rat Grimace Scale

With video-scoring during the acute phase, there were significant main effects for treatment and time (F (1, 23) = 2.3, p = 0.14 and F (2, 46) = 3.6, p = 0.034, respectively) and a significant interaction effect (F (2, 46) = 7.8, p = 0.0012). Post-hoc tests revealed that group 1 showed increased RGS scores from BL (p = 0.0002, 95% CI [0.14 to 0.44]) and controls (p = 0.003, 95% CI [-0.56 to -0.09]) on day 4 (Fig. 2.2.4.2). During the chronic phase, there were significant main effects for treatment and time (F (1, 24) = 2.4, p = 0.14 and F (4, 96) = 6.8, p < 0.001, respectively) and a significant interaction effect (F (4, 96) = 3.6, p = 0.0092). Post-hoc tests revealed that the RGS scores of group 2 decreased to their baseline and control levels before increasing significantly from baseline on chronic phase days 2 (p = 0.03, 95% CI [0.02 to 0.39]) and 3 (p < 0.0001, 95% CI [0.15 to 0.53]), crossing a previously established intervention threshold of 0.67 (Oliver *et al.*, 2014). A significant increase compared to controls was visible on day 3 (p = 0.004, 95% CI [-0.56 to -0.08]).

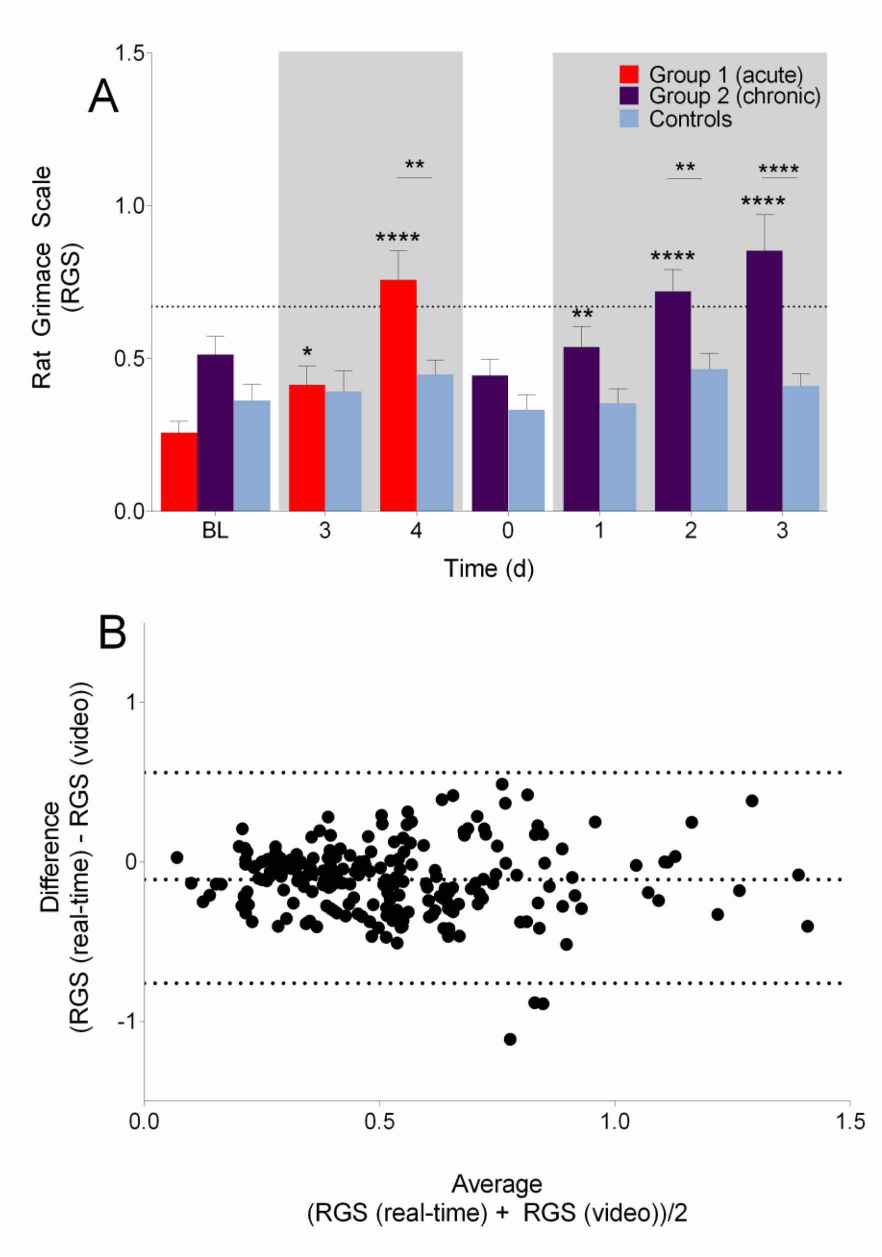
Similar increases from baseline and controls were observed from DSS-treated animals during the acute and chronic phases, when analyzed with real-time observations: there were significant main effects for treatment and time (F (1, 24) = 13, p = 0.0016 and F (4, 96) = 28, p < 0.0001, respectively) and a significant interaction effect (F (4, 96) = 16, p < 0.0001) (Suppl. Fig. 1A). The similarities between RGS real-time and video scores were also evident with a Bland-Altman of repeated measures, real-time scores had a bias of -0.11 when compared to video scores with limits of agreement ranging from -0.76 to -0.56 (Suppl. Fig. 2.2.4.2.2).

Fig. 2.2.4.2.1: Rat Grimace Scale (video) scores during the acute and chronic phases



Legend: Significant increases from baseline were observed on day 4 of the acute phase in group 1 and on days 2 and 3 of the chronic phase in group 2 ($p < 0.05$). Significant increases from controls were observed on day 4 during the acute phase and on day 2 and 3 during the chronic phase ($p < 0.01$). Broken horizontal line represents a derived analgesic intervention threshold.[24] Shaded boxes represent DSS treatment phases. *$p < 0.05$. **$p < 0.01$. ***$p < 0.001$. ****$p < 0.0001$. Data presented as mean ± SEM.

Fig. 2.2.4.2.2: The Rat Grimace Scale scores (real-time observations) during the acute and chronic phases (A, shaded boxes) and B) a comparison between real-time and video scores with Bland-Altman analysis for repeated measures.

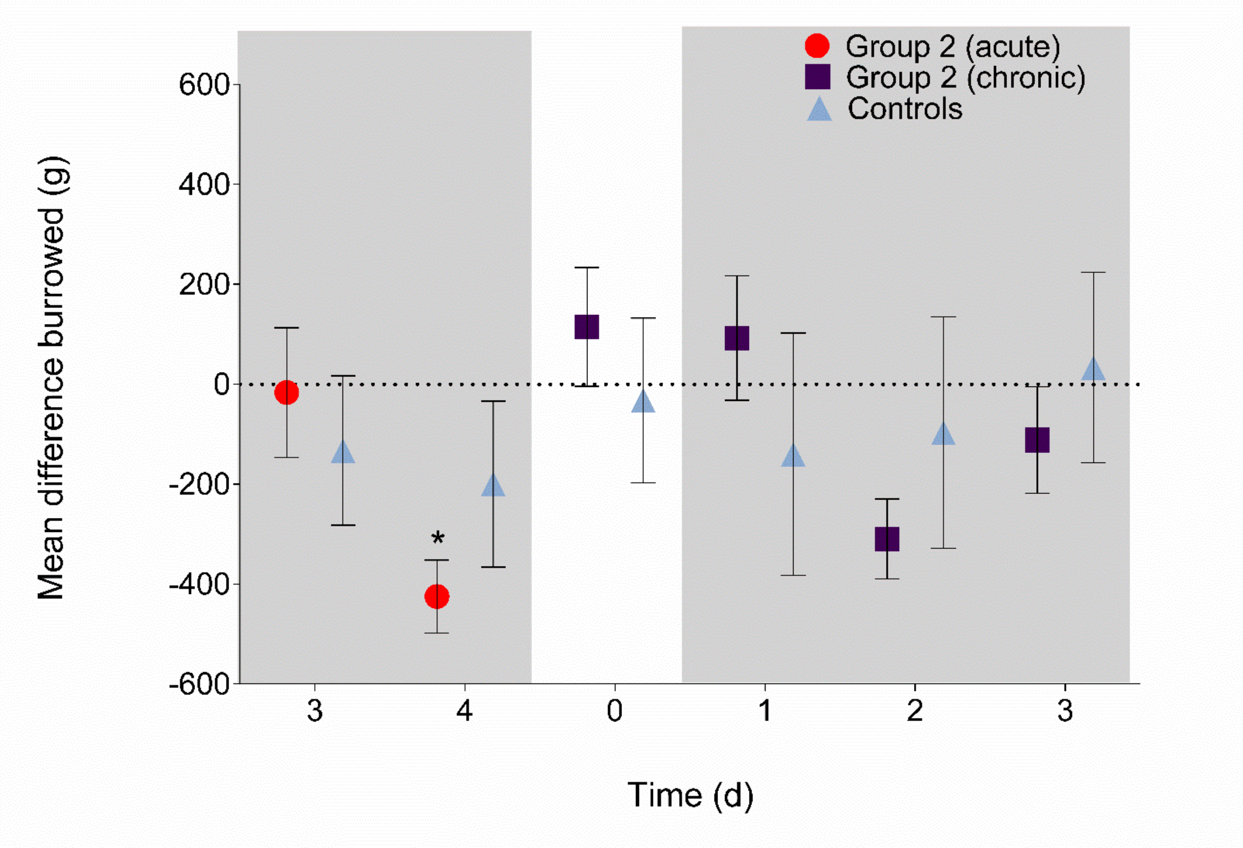


Legend: A) Significant increases from baseline were evident on days 3 and 4 of the acute phase ($p < 0.05$) and on day 1, 2 and 3 during the chronic phase ($p < 0.01$). Significant increases from controls were evident on day 4 of the acute phase ($p < 0.01$) and on days 2 and 3 of the chronic phase. Horizontal dotted line represents a previously derived intervention threshold of 0.67 (Oliver *et al.*, 2014). Data presented as mean ± SEM. B) Bias (-0.11, central broken horizontal line) reflects underestimation of video-based scores by real-time scores. Limits of agreement (broken horizontal lines) range from -0.76 to 0.56. *$p < 0.05$, **$p < 0.01$, ****$p < 0.0001$.

2.2.4.3. Burrowing

All rats burrowed to a similar degree at baseline (group 2: 1404.2 ± 566.5g; controls: 1330.0 ± 559.1g). During the acute phase, there was a significant main effect of time (F (1, 23) = 5.9, p = 0.023) but not treatment (F (1, 23) = 0.12, p = 0.74) and a non-significant interaction effect (F (1, 23) = 3.0, p = 0.095). Post-hoc tests revealed that there were no differences between the mean difference of gravel burrowed between group 2 and controls in both the acute and chronic phases (p > 0.99, all comparisons; Fig 4). During the acute phase, group 2 rats burrowed significantly less than baseline on day 4 (p = 0.03, 95% CI [36.2 to 813.7]). During the chronic phase, there were no significant differences observed (p > 0.05, all comparisons. 95%CI ranged from approximately -300 to 400).

Fig. 2.2.4.3.1: Mean difference in gravel displacement during acute and chronic colitis phases

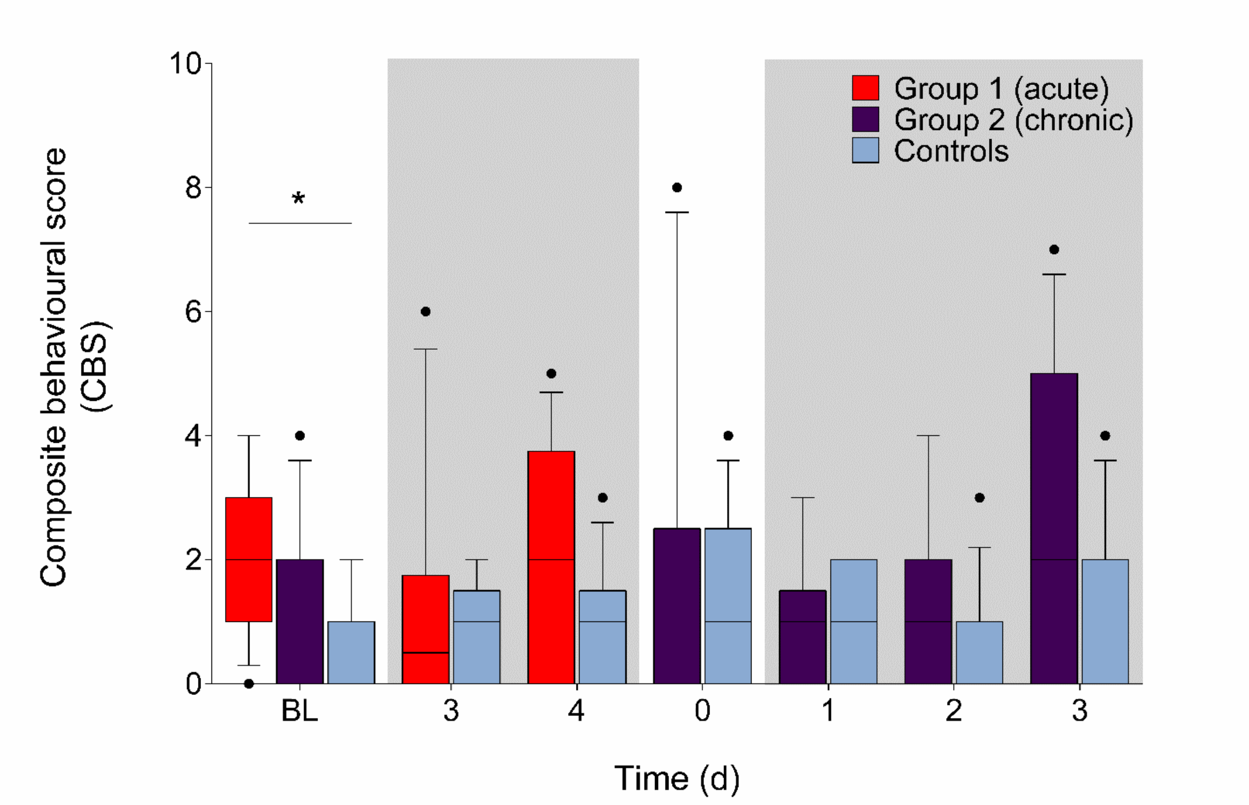


Legend: During both phases, no significant differences were observed between DSS treated and control rats (p > 0.99). A significant decrease from baseline was observed on day 4 (acute phase; p < 0.05). *p < 0.05. Data presented as mean ± SEM.

2.2.4.4. Composite Behavioural Score

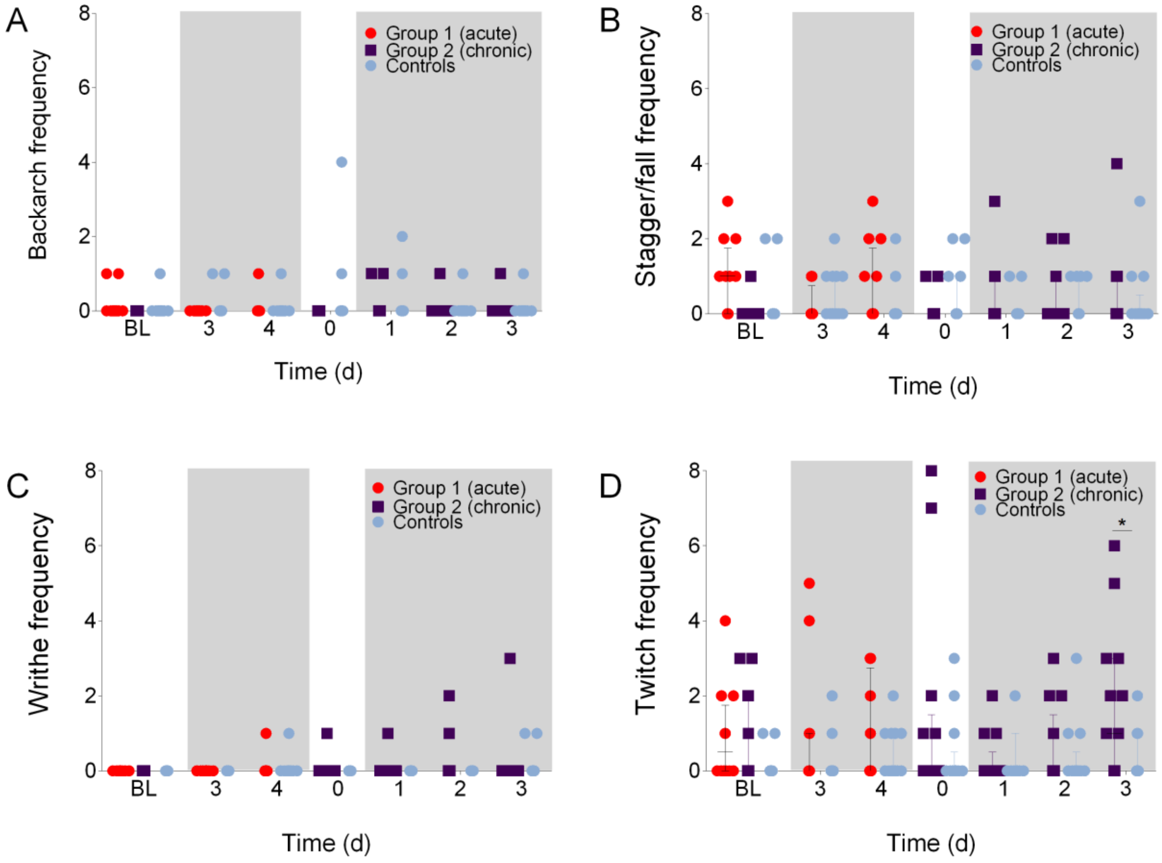
All behaviors except belly pressing were observed. During the acute phase, there was a significant main effect of treatment (F (1, 23) = 5.8, p = 0.024) but not time (F (2, 46) = 0.67, p = 0.52) and a non-significant interaction effect (F (2, 46) = 1.7, p = 0.20). During the chronic phase, there was a significant main effect of treatment (F (1, 24) = 5.6, p = 0.027) but not time (F (4, 96) = 1.7, p = 0.15) and a non-significant interaction (F (4, 96) = 0.79, p = 0.53). Post-hoc tests revealed that the only difference was between group 1 and controls at baseline (p = 0.02, 95% CI [-2.8 to -0.17], Fig. 5). No differences were observed between group 2 and controls or baseline (Figure 5, p > 0.05). Twitch frequency was the only behavior that identified treatment effects between group 2 and control rats during the third day of the chronic phase (p = 0.04, 95% CI [-2.7 to -0.021]; Suppl. Fig. 2D).

Fig. 2.2.4.4.1: Summed frequency of four behaviours (back arch, stagger/fall, writhe and twitch) evaluated during the acute phase and the chronic phase



Legend: Differences between groups were identified at baseline between group 1 and controls (p < 0.05). Differences within groups (from baseline) were not observed. Shaded boxes represent when DSS treatment was given. Data presented as median (10-90 percentile). *p < 0.05

Fig. 2.2.4.4.2: Breakdown of the frequency of all behaviours from the Composite Behaviour Score (CBS)



Legend: Shaded boxes represent DSS treatment durations. Belly pressing was never observed. No significant differences were observed within groups (from baseline). A significant difference between groups (control and group 2) was only observed for twitch behaviour ($p < 0.05$). Data presented as median ± IQR.

2.2.4.5. Microscopic score

After both acute and chronic phases, the microscopic score increased significantly from controls (p = 0.001, p < 0.0001, respectively; Table 4).

2.2.4.6. Macroscopic score

After both acute and chronic phases, significant increases from controls were evident (p = 0.003, p < 0.0001, respectively, Table 4).

Table 2.2.4.5.1: Microscopic and macroscopic scores of colon samples

	Controls (n=13)	Group 1 (acute phase, n=8)		Group 2 (chronic phase, n=13)	
	Mean [SD]	Mean ± SD	p-value [95% CI]	Mean [SD]	p-value [95% CI]
Microscopic	0.08 ± 0.28	7.3 ± 7.0	0.007 [-12 to -1.9]	4.3 ± 6.1	0.07 [-8.8 to 0.36]
Macroscopic	2.3 ± 1.7	8.0 ± 3.2	0.002 [-9.3 to -2.1]	10 ± 4.8	<0.0001 [-11 to -4.6]

Legend: Group 1: Significant differences from controls were evident from the microscopic and macroscopic scores (p < 0.01). Group 2: Significant differences from controls were evident from the macroscopic scores (p < 0.0001). p-values are comparisons between each treatment group and controls

2.2.5. Discussion

These results show: 1. Clinical signs of increasing disease severity (measured by the DAI) are reflected by an increase in RGS scores, but not by the CBS. 2. During acute colitis, as the DAI score increases, burrowing decreases. These data demonstrate that pain is likely to be present in DSS colitis models and increases concurrently with the presence of the clinical signs of the model (bleeding, loose stools and weight loss). This is in line with previous studies showing that visceral nociception (assessed with a colorectal balloon pressure measurement) and referred hypersensitivity (assessed with the von Frey filaments) occurred (Jain *et al*., 2015 and Tobin *et al*., 2004). Previous work has described the temporal relationship between hypersensitivity and ongoing pain, showing that pain presents over a shorter time course than hypersensitivity (in a peripheral models of inflammation), a situation that may better model the human experience (Gould., 2000 and De Rantere *et al*., 2016).

The changes in RGS scores coincide with model severity as assessed with the DAI, confirming that pain is present when clinical signs of colitis are apparent. Furthermore, the mean RGS scores exceeded an established analgesic interventional threshold (0.67; Oliver *et al.*, 2014). This observation may be helpful in guiding manipulations in this model (decision to provide pain relief, response to treatment, humane endpoints). The similar pattern of increase in RGS and DAI scores suggest that the DAI can be used as a proxy measure of pain. At the times when RGS scores crossed the analgesic intervention threshold DAI scores were >1, suggesting that this could be used as a proxy to trigger intervention. The successful application of facial expressions (Mouse Grimace Scale) has been previously applied to a murine colitis model (intrarectal allyl isothiocyanate), though no comparison was made with the DAI (Hassan *et al.*, 2017).

With real-time RGS scoring, the same pattern of change upon exposure to DSS was also observed, providing further support for the notion that real-time RGS scoring is a useful and feasible method of rapid pain assessment (Leung *et al.*, 2016). Furthermore, the closeness in RGS scores generated by real-time and standard scoring techniques supports the use of real-time scoring by a trained observer to routinely assess pain and welfare in this model. This means of rapid assessment could serve to identify humane endpoints or facilitate decisions regarding analgesia.

Unexpectedly, differences between DSS-treated and control rats were not identified with the CBS after DSS treatment. The behaviors evaluated (writhe, twitch, back arch, belly pressing and fall/stagger) were previously validated in rats that underwent a laparotomy and were suggested as a potential tool to assess visceral pain (Roughan and Flecknell, 2003 and Thomas *et al.*, 2016). The incidence of some of these same behaviors has also been observed to increase in ureteral calculi and intestinal mucositis models (Giamberardino *et al.*, 1995 and Whittaker *et al.*, 2016). However, a slightly different combination of behaviors was observed in each model. For example, back-arching behavior was the only behavior observed in all three models (laparotomy, ureteral calculi and intestinal mucositis models) while writhing was only observed after a laparotomy and intestinal model. This suggests that rats display a different combination of behaviors in different types of visceral pain models. Additional work is required to assess if the addition of different behaviors will allow discrimination between treatment groups in a DSS-colitis model.

Rats burrowed less on the same days where increases in DAI and RGS scores were observed during the acute phase (of group 2). This agrees with a previous mouse study that also observed reduced burrowing when mice were exposed to an acute dose of 2% DSS (Jirkof *et al.*, 2013). However, this decrease was not sustained during the chronic phase and no differences were observed compared to controls or baseline. The absence of changes in burrowing behavior from baseline during the chronic phase may reflect a lack of study power (reflected in wide 95%CI). Furthermore, the effect of chronic pain on burrowing behavior is currently unknown.

A limitation of this study is that a more comprehensive set of behaviors was not used as part of the CBS. Inclusion of additional behaviors may have better reflected the pain in this model. These behaviours could include abdominal licking and horizontal stretching, which were observed in mice following an allyl isothiocyanate induced colitis model (Hassan *et al.*, 2017).

In conclusion, the RGS was able to identify both acute and chronic phases of a colitis model, with changes occurring in tandem with clinical signs (reflected by the DAI). Additionally, burrowing activity reflects ongoing acute visceral pain in this colitis model and may be changed in the presence of chronic pain. The concurrent changes observed in the DAI and RGS suggest that the DAI may be a proxy measure for pain that is simple to apply. Pain assessments with the real-time RGS or DAI is recommended to assess the efficacy of treatment or analgesics for colitis-related pain, to study visceral pain mechanisms or to ensure the well-being of rats with colitis.

2.2.6. Disclosures

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. Funding for this study was from the Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (ID: 424022-2013, awarded to DSJP). VL receives a stipend from the Fondation J-Louis Lévesque. Funders had no role in study design, data collection and analysis, or decision to publish.

2.2.7. Acknowledgements

The authors would like to thank Colin Laferrière, Dr. Geneviève Fortin Simard, Dr. Frédérik Rousseau-Blass and Dr. Maxime Rufiange for help with data collection.

2.3. The influence of rater training on inter- and intra-rater reliability when using the Rat Grimace Scale

Emily Zhang[4]*, Vivian Leung[5]*, Daniel SJ Pang[4]

2.3.1. Abstract

Rodent grimace scales facilitate assessment of spontaneous pain and can identify a range of acute pain levels. Reported rater training in using these scales varies considerably and may contribute to observed variability in inter-rater reliability. This study evaluated the effect of training on inter-rater reliability with the Rat Grimace Scale (RGS). Two training sets, of 42 and 150 images, were prepared from several acute pain models. Four trainee raters progressed through 2 rounds of training, first scoring 42 images (set 1) followed by 150 images (set 2a). After each round, trainees reviewed the RGS and any problematic images with an experienced rater. The 150 images were then re-scored (set 2b). Four years after training, all trainees re-scored the 150 images (set 2c). A 'no training' group was also recruited and scored image sets 1 to 2b without reviewing with an experienced rater. Inter- and intra-rater reliability was evaluated using the intra-class correlation coefficient (ICC) and ICCs compared with a Feldt test. In the trainee group, inter-rater reliability increased from moderate (0.58 [95%CI: 0.43-0.72]) to very good (0.85 [0.81-0.88]) between set 1 and set 2b ($p < 0.01$) and also increased between set 2a and set 2b ($p < 0.01$). The action units with the highest and lowest ICCs at set 2b were orbital tightening (0.84 [0.80-0.87]) and whiskers (0.63 [0.57-0.70]), respectively. In comparison to an experienced rater the ICCs for all trainees improved, ranging from 0.88 to 0.91 at set 2b. Four years later, very good inter-rater reliability was retained (0.80 [0.76-0.84]) and intra-rater reliability was good or very good (0.78-0.86). In the 'no training' group, inter-rater reliability was moderate and did not improve from set 1 (0.43 [0.30-0.58]) to set 2a (0.41 [0.26-0.54]) or to set 2b (0.55 [0.44-0.64]; $p>0.05$). Training improves inter-rater reliability between trainees, with an associated reduction in 95%CI. Additionally, training resulted in improved inter-rater reliability alongside an experienced rater.

[4] Western College of Veterinary Medicine, University of Saskatchewan, Saskatoon, Saskatchewan, Canada

[5] Faculté de Médecine Vétérinaire, Université de Montréal, Saint-Hyacinthe, Québec, Canada

*These authors contributed equally to this work

Performance was retained after several years. The beneficial effects of training potentially reduce data variability and improve experimental animal welfare.

2.3.2. Introduction

The effectiveness of a pain assessment scale lies in its validity (does a scale measure what is intended) and reliability (measurement error). Rodent grimace scales have renewed interest in measuring the affective component of pain and have been promoted as a means of overcoming the shortfalls of nociceptive threshold testing (Langford *et al*., 2010; Sotocinal *et al*., 2011; Oliver *et al*., 2014 De Rantere *et al*., 2016 and Leung *et al*., 2016). There is increasing evidence that grimace scales discriminate painful and non-painful states in a range of acute pain models and interventions (Langford *et al*., 2010; Sotocinal *et al*., 2011; Leach *et al*., 2012; Oliver *et al*., 2014 and De Rantere *et al.,* 2016). However, there are conflicting reports regarding reliability when multiple raters score images (Langford *et al.,* 2010; Sotocinal *et al*., 2011; Oliver *et al*., 2014; Faller *et al*., 2015; and Mittal *et al*., 2016). Factors contributing to this variability may include a lack of structured training and variation in individual learning curves (de Oliveria, 2002 and Campbell *et al*., 2014).

It is unclear what level of training is required to attain proficiency in using grimace scales. Most studies include minimal, non-specific descriptions of training (Langford *et al*., 2010; Sotocinal *et al*., 2011; Leach et al., 2012; Oliver et al., 2014; Faller *et al*., 2015; Mittal *et al*., 2016 and Philips *et al*., 2017) and few reports any measure of reliability (Roughan and Flecknell, 2006; Langford *et al*., 2010; Oliver *et al*., 2014 and Mittal *et al*., 2016). Trainees progress at different rates during training to achieve proficiency in a task (Roughan and Flecknell, 2006; Campbell *et al*., 2014 and Mittal *et al*., 2016); therefore, in addition to training, some assessment of score reliability is necessary. The impact of training on scoring reliability with the Rat Grimace Scale (RGS) has not been formally evaluated. The objective of this study was to assess the effect of training on inter-and intra-rater reliability when scoring was performed with single and multiple raters applying the RGS. We hypothesised that training would improve inter-rater reliability.

2.3.3. Methods and materials

2.3.3.1. Animals and image selection

Two sets of training images were created from images collected during an unrelated project that had received institutional animal care and use committee approval from the University of Calgary Health Sciences Animal Care Committee (protocol IDs: AC13-0161 and AC13-0124; De Rantere *et al*., 2016). This project used the following acute pain models: intra-plantar carrageenan, intra-plantar Complete Freund's adjuvant or plantar incision. The RGS scores from the three models display the RGS' spectrum of possible scores (scores 0-2; De Rantere *et al.,* 2016). Animals were adult (> 10 weeks old) male Wistar (n = 34) rats, from a commercial source (Charles River Laboratories, Québec, Canada).

The methodology used to generate images was as previously described (Sotocinal *et al*., 2011). Briefly, still images were captured from high-definition video-recordings and cropped so that only the face was visible. Each image was presented on a single slide in presentation software (Microsoft PowerPoint, version 14.0, Microsoft Corporation, Redmond, WA, USA). Slide order was randomised and identifying information (animal ID, time point, model) removed.

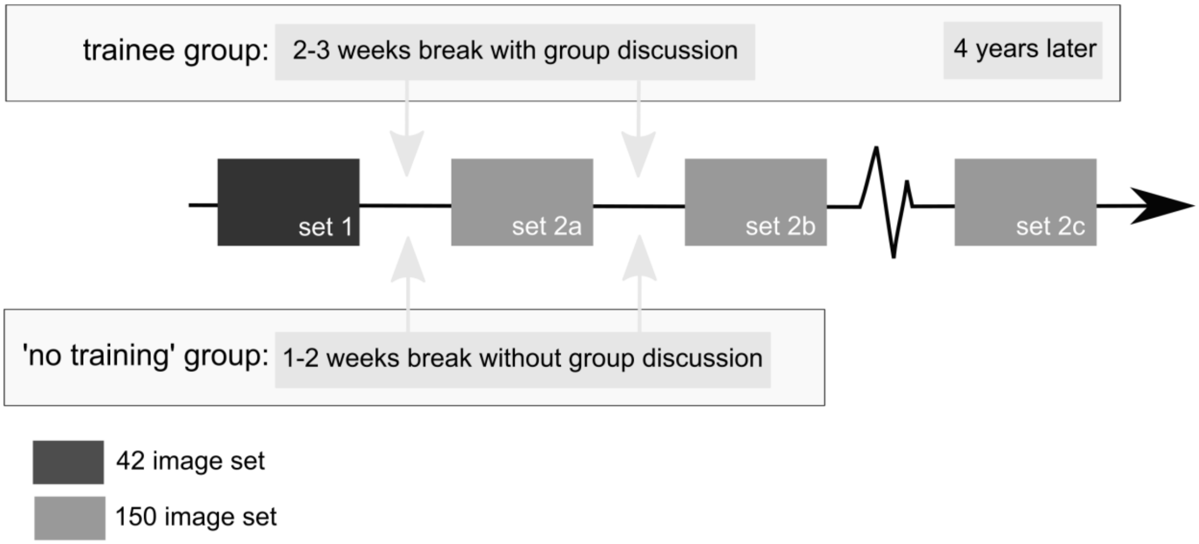
Images were selected based on image quality alone, by an individual not involved with the study. Two unique sets of training images were created, of 42 (set 1) and 150 (set 2) images. Images were scored using the RGS (scale range 0-2 for each action unit) and the average score calculated from four action units: orbital tightening, nose/cheek flattening, ear changes, and whisker change.

2.3.3.2. Training protocol

None of the 4 trainee raters recruited had previous experience with the RGS. All trainee raters (rater 1, 2, 3 and 4) were female undergraduate and graduate students (age range 20-25 years), studying veterinary medicine, biology (n = 2) and health sciences and were recruited when joining the research group as project students. No trainee raters had previous experience with rats, as experimental animal or pets, before beginning training. The experienced rater (DP) had applied the RGS for several years, successfully identifying painful interventions using established models (a form of construct validity; known-group discrimination; Calvo *et al*., 2014; Oliver *et al.,* 2014 and De Rantere *et al*., 2016) and adoption of the RGS method within the research group of the experienced rater was supported with the assistance of the Mogil laboratory (McGill University), developers of the mouse and rat grimace scales (Langford *et al.,* 2010 and Sotocinal *et al*., 2011), through informal evaluation of scoring performance.

All trainee raters followed the same scoring protocol (Figure 2.1.2.3.1): set 1 image was scored independently by each individual, using the training manual provided by Sotocinal *et al* (2011) alongside a training manual from our laboratory (Appendix B). Trainee raters were encouraged to record comments for any images they found difficult to score. Following set 1 scoring, trainee raters reviewed their scores as a group with an experienced rater, discussing recorded comments and areas of inconsistency. Images with the most variation between raters were selected for review. The primary goal of the discussion was to improve standardisation of scoring images assigned a score of 0 or 2. Disagreement in scores was tolerated provided differences between raters did not exceed 1 point on the scale. The standard of scoring was set by the experienced rater, following establishment of the technique within the laboratory with the support of the Mogil laboratory (McGill University). Once review of set 1 scoring was complete, set 2 images were scored independently by each individual and comments recorded as before (set 2a). The set 2 image set was added when more images were available. The S2 image set was then scored independently a second time (set 2b) after a facilitated group discussion with the experienced rater (as per the set 1 image set discussion). Approximately 15-30 images were reviewed during group discussions, with 2-3 weeks between reviews. Intra-rater reliability was assessed by asking the trainee raters to independently re-score the set 2 image set (set 2c) with access to the training manual. Scoring set 2c took place 4 years after initial training. The order of the images was randomised from set 2b. At the time of set 2c scoring, trainee rater 1 had not used the RGS in 10 months and trainee raters 3 and 4 had not used it in three years. Trainee rater 2 was still in the research group and actively using the RGS. All trainee raters were asked if they remembered any previous scores or images from the data set.

Fig. 2.1.2.3.1: Timeline of training protocol



Legend: Two image sets of 42 and 150 images (set 1 and set 2 respectively) were scored independently by all trainee raters and 'no training' raters. For trainee raters, set 1, set 2a and set 2b were scored with 2-3 weeks break in between. During the break, a group discussion with the experienced rater took place to discuss inconsistencies. After each scoring session, the scores from each individual trainee rater was compared to the experienced rater to assess inter-rater reliability. Four years later, the 150 image set was randomized and re-scored (set 2c) by all trainee raters. Their scores were compared to the experienced rater's and their own scores from set 2b to assess inter- and intra-rater reliability respectively. For 'no training' raters, set 1, set 2a and set 2b were scored with 1 week break in between. These raters never participated in any discussion and their scores were also compared to the experienced rater to assess inter-rater reliability.

2.3.3.3. 'No training' group

A second group of raters ('no training') were later recruited to assess if repeated scoring of images with no group discussions (only the training manual utilised) would be enough to attain scoring proficiency. Eight raters were recruited but only six completed the entire process (rater 5, 6, 7, 8, 9 and 10). Five raters were female, and one was male (rater 7; age range 24-26 years). Rater 7 was the only one that had prior knowledge of the RGS, but he had not used or been trained to use the RGS. All other raters had no previous experiences with rats, as experimental animals or pets, nor any previous knowledge regarding the RGS. All raters are from a science background having graduated with an undergraduate degree in zoology (n = 3), possess (n = 1) or are candidates for a Doctor of Veterinary Medicine (n = 1), or pursuing a Masters in integrative biology (n = 1).

These raters were asked to score the same image sets raters 1-4 had previously scored (set 1, set 2a, set 2b) following the same training protocol, with the exception being that these raters did not discuss or consult with the experienced rater (or each other) during the entire scoring process and there was a week in between scoring each image set (Figure 2.1.2.3.1).

2.3.3.4. Statistics

Intraclass correlation coefficients (ICCs, MedCalc version 12.6.1.0, MedCalc Software, Ostend, Belgium) were calculated to measure the reliability of RGS scoring between and within raters for the individual action unit scores and average RGS scores. An absolute model was used for the ICC calculation and single measure reported. This was done for each dataset (set 1, set 2a, set 2b and set 2c) for both groups of raters (trainee or 'no training'). ICCs were also calculated for the comparison between each individual trainee or 'no training' rater's scores and those of the experienced rater (DP) to determine reliability of each individual rater. Planned comparisons were pre-established: calculated ICCs were compared with a Feldt test for set 1 *versus* set 2b, set 1 *versus* set 2a, set 2a *versus* set 2b and set 2b *versus* set 2c (critical F set at alpha = 0.01 and differences considered significant if the observed F value was greater than the critical F value; Feldt *et al*., 1987 and Kuzmic, 2015). ICCs were also calculated between each trainee and 'no training' rater's own scores (set 2b and set 2c) to assess intra-rater reliability over time. Interpretation of the ICC followed the same divisions as used previously: ‘‘very good’’ (0.81–1.0), ‘‘good’’ (0.61–0.80), ‘‘moderate’’ (0.41–0.60), ‘‘fair’’ (0.21–0.40), ‘‘poor’’ (< 0.20; Oliver *et al*., 2014). During the training process, trainee raters were said to be proficient when calculated ICCs ± 95%CI overlapped with those published in a study reporting inter-rater reliability (Oliver *et al*., 2014) and obtained an ICC of at least 0.80 (Haidet *et al*., 2009). To assess the potential impact of scores memorised during group discussion between set 2a and set 2b introducing bias in to the ICC calculation for set 2b, images with the greatest scoring variability at set 2a (those with a difference of 2 points between any 2 raters and therefore the most likely to have been discussed) were removed and the ICCs for set 2b recalculated. Data are presented as ICC (± 95%CI) and a corrected p value for multiple comparisons of ≤ 0.017 was considered significant. Scoring accuracy was assessed by comparing the experienced rater's scores for images collected at baseline and 6-9 hours after treatment (when a peak in RGS scores could be expected for the models studied (De Rantere *et al.*, 2016); paired t test with alpha set at 0.05) from the set 2 images. The datasets generated from this study and training manual are available in the Harvard Dataverse repository (Pang, 2018).

2.3.4. Results

Four trainee raters and six 'no training' raters completed the study. All training images were scored by every rater, and all scores included in the subsequent analysis.

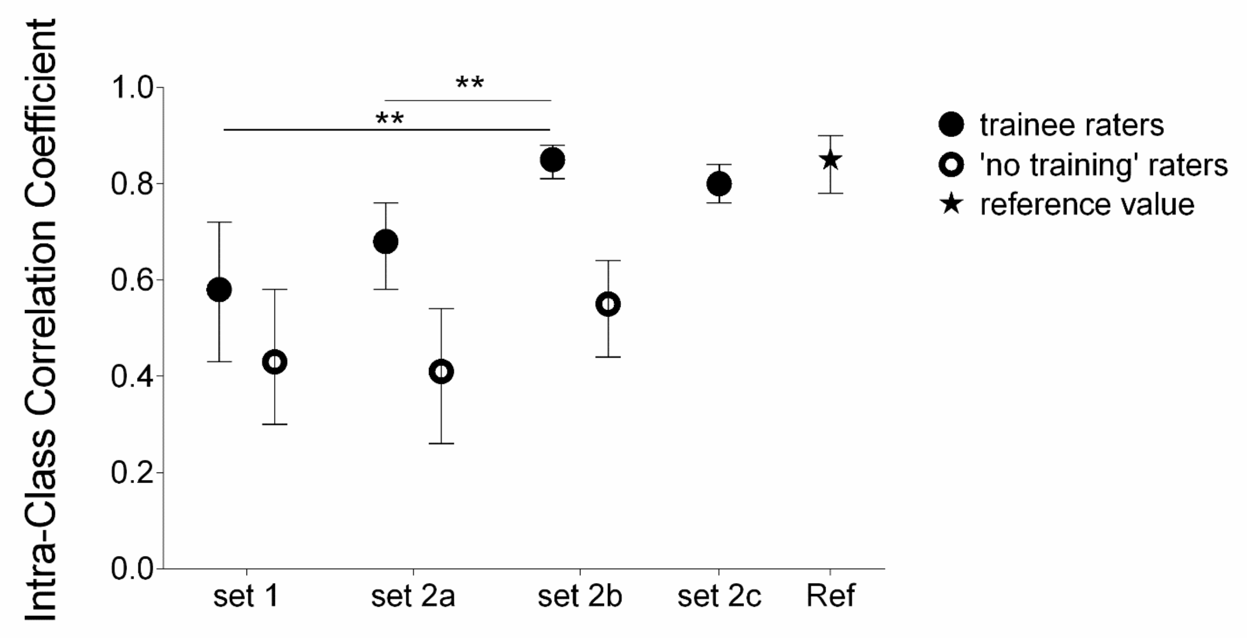
2.3.4.1. Inter-rater reliability of trainee raters

Training was associated with a progressive improvement in inter-rater reliability and narrowing 95%CI (Figure 2.3.4.1.1). The first training round (set 1) resulted in a moderate ICC for the average RGS scores, with wide 95%CI (0.58 [0.43-0.72]). The increase in average RGS ICC between set 1 and set 2a (0.68 [0.58-0.76]) was not statistically significant ($F_{0.01;149,41} = 1.88$, observed F = 1.31, $p > 0.05$). A significant improvement was observed at set 2b (0.85 [0.81-0.88]) compared with set 1 (observed F = 2.8) and set 2a ($F_{0.01;149,149} = 1.47$, observed F = 2.13, $p < 0.01$ for both comparisons). The resultant set 2b ICC was classified as very good and comparable with published values (Figure 2.3.4.1.1; Oliver *et al*., 2014).

A similar pattern of improvement was observed in the scores of individual action units (Table 2.3.4.1.1). Significant increases in ICCs were observed between set 1 and set 2b for orbital tightening (observed F = 1.94), ear changes (observed F = 2.14) and nose/cheek flattening (observed F = 2.21, $p < 0.01$ all comparisons), but not whisker changes (observed F = 1.65, $p > 0.05$). And between set 2a and set 2b: orbital tightening (observed F = 1.81), ear changes (observed F = 1.96) and nose/cheek flattening (observed F = 1.72, $p < 0.01$ all comparisons), but not whisker changes (observed F = 1.35, $p > 0.05$). At all stages, orbital tightening had the highest ICC, improving from 0.69 to 0.84. Following training, ICCs for individual action units fell within the good or very good range (Table 2.3.4.1.1).

Comparing individual trainee rater performance against the experienced rater showed considerable variation following the first training round with ICCs ranging from fair to good. All trainee raters showed improvement with training (Table 2.3.4.1.2).

Fig. 2.3.4.1.1: Average group intra-class correlation coefficients



Legend: Average group intra-class correlation coefficients for each of the four datasets (mean and 95%CI) with reference values. Ref = reference (Oliver *et al.*, 2014).

Table: 2.3.4.1.1 Group Intra-class Correlation Coefficients (ICC) for each of the datasets.

Action Unit	set 1	set 2a	set 2b	set 2c	Reference values
Orbital tightening	0.69 [0.56-0.80]^{a}	0.71 [0.63-0.78]^{b}	0.84 [0.80-0.87]^{a,b,c}	0.76 [0.70-0.81]^{c}	0.92 [0.89-0.95]
Ear changes	0.40 [0.25-0.56]^{a}	0.45 [0.35-0.54]^{b}	0.72 [0.66-0.77]^{a,b,c}	0.60 [0.51-0.68]^{c}	0.62 [0.51-0.72]
Nose/Cheek flattening	0.36 [0.21-0.52]^{a}	0.50 [0.41-0.58]^{b}	0.71 [0.65-0.76]^{a,b}	0.64 [0.57-0.70]	0.62 [0.51-0.72]
Whisker change	0.39 [0.26-0.55]	0.50 [0.42-0.58]	0.63 [0.57-0.70]	0.54 [0.45-0.62]	0.52 [0.39-0.63]

Legend: Set 1, set 2a and set 2b are the first, second and third training round, respectively. Set 2c was scored 4 years after initial training. ICC scores are divided as ‘‘very good’’ (0.81–1.0), ‘‘good’’ (0.61–0.80), ‘‘moderate’’ (0.41–0.60), ‘‘fair’’ (0.21–0.40), ‘‘poor’’ (< 0.20). Data are ICCsingle [95%CI]. Within a row, identical superscript letters indicate significant differences between the different training rounds, $p < 0.01$. Reference values and ICC score divisions are from Oliver *et al*. (2014).

Tables: 2.3.4.1.2: Agreement of each individual trainee rater when compared to an experienced rater (DP).

Image set	Rater 1 vs DP	Rater 2 vs DP	Rater 3 vs DP	Rater 4 vs DP
set 1	0.41 [0.06-0.66]^{a,b}	0.70 [0.50-0.83]^{a}	0.62 [0.36-0.79]^{a}	0.42 [0.13-0.64]^{a}
set 2a	0.84 [0.79-0.88]^{a}	0.75 [0.68-0.82]^{b}	0.68 [0.25-0.84]^{b}	0.65 [0.38-0.79]^{b}
set 2b	0.89 [0.85-0.92]^{b}	0.88 [0.84-0.91]^{a,b}	0.91 [0.88-0.94]^{a,b}	0.90 [0.87-0.93]^{a,b,c}
set 2c	0.87 [0.82-0.90]	0.86 [0.82-0.90]	0.86 [0.80-0.90]	0.78 [0.71-0.83]^{c}

Legend: ICC scores are divided as ‘‘very good’’ (0.81–1.0), ‘‘good’’ (0.61–0.80), ‘‘moderate’’ (0.41–0.60), ‘‘fair’’ (0.21–0.40), ‘‘poor’’ (< 0.20). Data are ICCsingle [95%CI]. Within a column, matching superscript letters indicate significant differences ($p < 0.01$). ICC score divisions are from Oliver *et al*. (2014).

There were 28 images (19%) with score differences between raters of 2 points at set 2a. Removing these scores had a minimal effect on the recalculated ICCs for set 2b (average RGS scores were 0.85 [0.81-0.88] and 0.86 [0.83-0.89] for 150 and 122 images, respectively).

There was a significant increase in RGS scores between baseline (n = 41, 0.45 ± 0.07) and 6-9 hours after treatment (n = 29, 0.92 ± 0.08, $p < 0.001$, 95%CI of mean difference 0.27 to 0.68), at which time the mean RGS score exceeded a published analgesic intervention threshold (Oliver *et al.*, 2014).

When the images were re-scored four years after initial training (set 2c), the ICC was good for the averaged RGS scores (0.80 [0.76-0.84]) and proficiency was maintained from set 2b (observed F = 1.33, $p > 0.01$). Between set 2b and set 2c there were no significant differences for nose/cheek flattening (observed F = 1.24, $p > 0.05$), whisker changes (observed F = 1.24, $p > 0.05$) and ear changes (observed F = 1.42, $p > 0.01$), Table 2.3.4.1.1. However, inter-rater reliability from set 2b was not maintained and decreased significantly for orbital tightening (observed F = 1.50, $p < 0.01$). All trainee raters maintained similar proficiency with the experienced rater (observed F < 1.31, $p > 0.05$) except for trainee rater 4 (observed F = 2.20, $p < 0.01$; Table 2.3.4.1.2).

2.3.4.2. Intra-rater reliability of trainee raters

The ability of a trainee rater to score reliably over time was good or very good with ICCs ranging from 0.78 to 0.86 for the average RGS (Table 2.3.4.2.1). The intra-rater reliability of individual action units ranged from moderate to very good depending on the action unit and trainee rater. Two trainee raters (2 and 4) reported that they did not recognise any images or remember previous scores while the remaining trainee raters (1 and 3) reported recognizing a few images but did not remember scores.

2.3.4.3. Inter-rater reliability of 'no training' raters

In the 'no training' group, repeated scoring of images did not result in improvement of inter-rater reliability (Figure 2.3.4.1.1). The agreement between raters was moderate during all three stages of scoring with no significant improvement observed from set 1 (0.43 [0.30-0.58]) to set 2a (0.41 [0.26-0.54]; $F_{0.01;\ 149;\ 41}$ = 1.88, observed F = 1.04, p > 0.05), or from set 1 to set 2b (0.55 [0.44-0.64]; $F_{0.01;\ 149;\ 41}$ = 1.88, observed F =1.27, p > 0.05) or from set 2a to set 2b ($F_{0.01;\ 149;\ 149}$ = 1.47, observed F =1.31, p > 0.05).

This pattern of no improvement was also observed with individual action units (Table 2.3.4.3.1). Some improvements were observed from individual raters when comparing their scores to the experienced rater, however, none of the raters had very good agreement with the experienced rater (Table 2.3.4.3.2). Rater 6 improved from set 1 to set 2a (observed F = 1.97, p < 0.01) and raters 7 and 8 improved from set 2a to set 2b (observed F = 1.58, p <0.01; observed F = 1.57, p < 0.01 respectively).

Table 2.3.4.2.1: Intra-class Correlation Coefficients (ICC) for intra-rater reliability for each individual trainee rater four years after initial training

Action Unit	Rater 1	Rater 2	Rater 3	Rater 4
Average	0.85 [0.78-0.90]	0.86 [0.82-0.90]	0.86 [0.79-0.90]	0.78 [0.71-0.84]
Orbital tightening	0.72 [0.53-0.82]	0.86 [0.82-0.90]	0.85 [0.78-0.89]	0.75 [0.63-0.83]
Ear changes	0.68 [0.48-0.80]	0.49 [0.11-0.70]	0.74 [0.66-0.81]	0.71 [0.61-0.79]
Nose/Cheek flattening	0.64 [0.53-0.73]	0.68 [0.56-0.77]	0.74 [0.60-0.82]	0.63 [0.53-0.72]
Whisker change	0.77 [0.70-0.83]	0.69 [0.55-0.78]	0.53 [0.27-0.69]	0.47 [0.34-0.59]

Legend: ICC scores are divided as ''very good'' (0.81–1.0), ''good'' (0.61–0.80), ''moderate'' (0.41–0.60), ''fair'' (0.21–0.40), ''poor'' (< 0.20). Data are ICC single [95% CI]. ICC score divisions are from Oliver *et al.* (2014).

Table 2.3.4.3.1: Group Intra-class Correlation Coefficients (ICC) for each of the datasets for the "no training group"

Action Unit	set 1	set 2a	set 2b	Reference values
Orbital tightening	0.48 [0.35-0.62]	0.65 [0.58-0.71]	0.71 [0.65-0.76]	0.92 [0.89-0.95]
Ear changes	0.24 [0.14-0.38]	0.35 [0.25-0.46]	0.35 [0.24-0.46]	0.62 [0.51-0.72]
Nose/Cheek flattening	0.35 [0.23-0.50]	0.17 [0.09- 0.26]	0.35 [0.27-0.43]	0.62 [0.51-0.72]
Whisker change	0.19 [0.09 -0.32]	0.23 [0.16-0.32]	0.25 [0.18-0.33]	0.52 [0.39-0.63]

Legend: Set 1, set 2a and set 2b are the first, second and third rounds of scoring, respectively. ICC scores are divided as ''very good'' (0.81–1.0), ''good'' (0.61–0.80), ''moderate'' (0.41–0.60), ''fair'' (0.21–0.40), ''poor'' (< 0.20). Data are ICCsingle [95%CI]. Reference values and ICC score divisions are from Oliver *et al*. (2014).

Table 2.3.4.3.2: Agreement of each individual "no training" rater when compared to an experienced rater (DP).

Image set	Rater 5 vs DP	Rater 6 vs DP	Rater 7 vs DP	Rater 8 vs DP	Rater 9 vs DP	Rater 10 vs DP
set 1	0.63 [0.40-0.78]	0.37 [0.07-0.60]^{a}	0.57 [0.33-0.74]	0.33 [0.04-0.57]	0.56 [0.24-0.75]	0.57 [0.25-0.76]
set 2a	0.72 [0.60-0.81]	0.68 [0.58-0.76]^{a}	0.51 [0.06-0.73]^{a}	0.12 [-0.06-0.30]^{a}	0.63 [0.28-0.80]	0.67 [0.45-0.79]
set 2b	0.68 [0.57-0.77]	0.65 [0.54-0.74]	0.69 [0.41-0.82]^{a}	0.41 [0.05-0.64]^{a}	0.73 [0.56-0.82]	0.68 [0.51-0.78]

Legend: Agreement of each individual "no training" rater when compared to an experienced rater (DP). ICC scores are divided as ''very good'' (0.81–1.0), ''good'' (0.61–0.80), ''moderate'' (0.41–0.60), ''fair'' (0.21–0.40), ''poor'' (< 0.20). Data are ICCsingle [95%CI]. Within a column, matching superscript letters indicate significant differences ($p < 0.01$). ICC score divisions are from Oliver *et al.* (2014).

2.3.5. Discussion

Our results suggest that reliability is limited when raters only review the training manual and score images repeatedly. Improvement is observed when feedback and discussion with an experienced rater are included. The high level of reliability and proficiency achieved from training can be maintained for several years.

Little is known regarding the need for, or role of, rater training in the use of rodent grimace scales. Where training has been described, it ranges from reviewing the grimace scale training manuals (Leach *et al*., 2012 and Faller *et al*., 2015) to a single training session of variable length (Langford *et al*., 2010; Sotocinal *et al*., 2011; Oliver *et al*., 2014; De Rantere *et al*., 2016 and Philips *et al*., 2017) or multiple training sessions (Mittal *et al*., 2016). Few studies describe an assessment of reliability (Langford *et al*., 2010; Sotocinal *et al*., 2011; Oliver *et al*., 2014 and Mittal *et al*., 2016). The results of this study show that an assessment of reliability is necessary to confirm that training will lead to proficiency as well as standardised scoring. This study also demonstrated that inclusion of a group discussion as part of training is beneficial. While repeated exposure without discussion does have some benefits as observed by the increased reliability amongst the individual raters from the 'no training' group, this improvement is variable between raters and none of them achieved the same reliability as those in the trainee group.

The rate at which individuals achieve proficiency in a task is highly variable and, as such, it is erroneous to assume that participating in training guarantees proficiency. Neither a single training session nor repeated attempts at a task ensure proficiency (de Oliveira, 2002; Roughan and Flecknell, 2006; Campbell *et al*., 2014). The length and intensity of training should depend on the difficulty of mastering the tool and the proficiency of the trainee (Haidet *et al*., 2009). Additionally, proficiency should not be assumed just because a rater feels confident using a scale following training (Björn *et al*., 2017). Instead, it is important to test the actual proficiency of raters, and a simple approach is to assess inter-rater reliability (Streiner and Norman, 2008). This provides assurance that scoring has reached the desired standard, that variability is at an acceptable level and enables rogue raters to be identified (Brondani *et al*., 2013 and Mittal *et al*., 2016). Identification of rogue raters during training allows for further testing and assessment or removal from participation in scoring (Mittal *et al*., 2016 and Mullard *et al*., 2017). Ensuring reliability and standardizing scoring will reduce data variability and consequently, animal use. An alternative approach is to use a single rater; however, it is still useful to compare the performance of a single rater against that of an experienced rater, or a standard set of scores, to confirm reliability and consistency over time (Oliver *et al*., 2014). The presence of systematic bias may negatively affect data interpretation and pain management (Faller *et al*., 2015).

Orbital tightening had the highest associated ICC following the initial round of scoring, which was maintained throughout training. In contrast, the reliability of whisker scoring remained relatively low throughout training. These results support previous findings that assessing the whisker change action unit is more difficult for raters than orbital tightening (Oliver *et al*., 2014).

Four years after training, with variable use of the RGS during this time, the inter- and intra-rater reliability of the average RGS was maintained. This indicates that raters can retain scoring proficiency and score consistently with each other, with themselves and achieve the standard set by the experienced rater. This agrees with a previous study showing that a single rater maintained scoring reliability after a break of six months (Oliver *et al*., 2014). Nevertheless, the observed reductions in ICC for one of the action units indicate that some degree of re-training may be beneficial.

A recent description of a successful machine learning approach to the MGS highlights the potential for simplifying the standard method of facial image acquisition and scoring (Streiner and Norman, 2008). This advance could greatly shorten what is currently a relatively slow process and allow for the scoring of large numbers of animals in a short period of time, an advance over real-time scoring (Leung *et al*., 2016). However, the need for proficient human raters remains necessary to classify those images that cannot currently be scored by machine with a high degree of confidence (Tuttle *et al*., 2018).

A limitation of this study was re-scoring the 150-image set in the final training round, with the potential for memorised scores assigned during the group discussion following the second training round being applied rather than a rater scoring independently. We feel this is unlikely due to the large number of images scored, the similar appearance of rodent faces from similar strains, the time elapsed between review rounds, the small number of images reviewed during group discussion and the nature of the group discussion, where disagreement between raters was acceptable. The minimal difference in ICCs after removal of the 28 image scores supports this assertion as well as the maintained quality of scores after 4 years. A further limitation is the generalisability of these findings, based on 4 trainee raters and 6 ‘no training’ raters, to a larger population. These results highlight the risk of assuming that some form of training in the use of the RGS (and perhaps other facial expression scales) is unnecessary and should serve to encourage users to regularly evaluate scoring reliability and accuracy. In more general terms, scale performance is specific to the population and context studied, so that performance when applied by different raters or in a different context should be formally evaluated (Streiner and Norman, 2008).

Images for training were selected on the basis of quality rather than to allow comparison between treatment groups. This limits any assessment of construct validity but the comparison of baseline and predicted peak pain periods indicates that accuracy was preserved.

In conclusion, these data show that reliance on access to the available manuals for rater training may be insufficient. Formal training that includes group discussion with an experienced rater improves inter-rater reliability and is likely to reduce data variability if rater proficiency is assessed before embarking on data collection. Collaborative training between research groups would ensure similar levels of rater proficiency and improve the reproducibility of research. Inclusion of clear descriptions of rater training and assessment would help in evaluating study results. Lastly, once raters achieve proficiency, this may be maintained over several years even without scoring during the intervening period.

2.3.6. Acknowledgements

The authors thank Susana Sotocinal of the Mogil Laboratory, McGill University, for invaluable assistance in establishing the Rat Grimace Scale in our laboratory and reviewing the selection of images in our training manual, Kent Hecker and Grace Kwong (University of Calgary) for statistical advice and Audrey Pang, Chelsea Schuster, Elfleda Chik, Jesse Tong, Julie Reimer, Shi Jie Zhou, Vimanda Chow and Winnie Yang for taking part in the study.

2.4. ARRIVE has not ARRIVEd: Support for the ARRIVE (Animal Research: Reporting of in vivo Experiments) guidelines does not improve the reporting quality of papers in animal welfare, analgesia or anesthesia

Vivian Leung[6]*, Frédérik Rousseau-Blass[5*], Guy Beauchamp[5], Daniel SJ Pang[5]

2.4.1. Abstract

Poor research reporting is a major contributing factor to low study reproducibility, financial and animal waste. The ARRIVE (Animal Research: Reporting of *In Vivo* Experiments) guidelines were developed to improve reporting quality and many journals support these guidelines. The influence of this support is unknown. We hypothesised that papers published in journals supporting the ARRIVE guidelines would show improved reporting compared with those in non-supporting journals. In a retrospective, observational cohort study, papers from 5 ARRIVE supporting (SUPP) and 2 non-supporting (nonSUPP) journals, published before (2009) and 5 years after (2015) the ARRIVE guidelines, were selected. Adherence to the ARRIVE checklist of 20 items was independently evaluated by two reviewers and items assessed as fully, partially or not reported. Mean percentages of items reported were compared between journal types and years with an unequal variance t-test. Individual items and sub-items were compared with a chi-square test. From an initial cohort of 956, 236 papers were included: 120 from 2009 (SUPP; n = 52, non- SUPP; n = 68), 116 from 2015 (SUPP; n = 61, nonSUPP; n = 55). The percentage of fully reported items was similar between journal types in 2009 (SUPP: 55.3 ± 11.5% [SD]; non- SUPP: 51.8 ± 9.0%; p = 0.07, 95% CI of mean difference -0.3±7.3%) and 2015 (SUPP: 60.5 ± 11.2%; nonSUPP; 60.2 ± 10.0%; p = 0.89, 95%CI -3.6±4.2%). The small increase in fully reported items between years was similar for both journal types (p = 0.09, 95% CI -0.5±4.3%). No paper fully reported 100% of items on the ARRIVE checklist and measures associated with bias were poorly reported. These

[6] Faculté de Médecine Vétérinaire, Université de Montréal, Saint-Hyacinthe, Québec, Canada

*These authors contributed equally to this work results suggest that journal support for the ARRIVE guidelines has not resulted in a meaningful improvement in reporting quality, contributing to ongoing waste in animal research.

2.4.2. Introduction

Accurate and complete reporting of animal experiments is central to supporting valid, reproducible research and to allow readers to critically evaluate published work. Poor or absent reporting is associated with deficiencies in experimental design that introduce bias and exaggerated effect sizes in to the literature (Macleod *et al.*, 2008 and Vesterinen *et al.*, 2010). As a result, irreproducible animal research has significant ethical and financial costs (Freedman *et al.*, 2015). The use of animals in poorly designed studies and in efforts to reproduce such studies represents a failure to uphold the 3Rs (refine, reduce, replace) of animal research (Russell and Burch, 1992). Incomplete reporting of research contributes to a waste of funding, with a conservative estimate for preclinical research, of US$28 billion annually (Freedman *et al.*, 2015).

To address low standards of reporting, the ARRIVE (Animals in Research: Reporting *In Vivo* Experiments) guidelines for reporting were published in 2010 (Kilkenny *et al.*, 2009 and 2010). The ARRIVE guidelines are summarised by a 20-item checklist that includes reporting of measures associated with bias (randomisation, blinding, sample size calculation, data handling) (Landis *et al.*, 2012 and Macleod *et al.*, 2015). Over 1000 journals have responded to publication of the guidelines by linking to it on their websites and in their instructions to authors (www.nc3rs.org). The effect of these endorsements is unknown. For the majority of existing health research guidelines, the impact of journal support for other reporting guidelines on guideline adherence in published papers is unclear (Stevens *et al.*, 2014). The impact of the CONSORT guidelines for the reporting of randomised controlled trials have been evaluated more than other reporting guidelines, and current evidence suggests that though reporting of some items has improved, overall standards of reporting remain low (Turner *et al.*, 2012).

To our knowledge, there have been no studies comparing reporting standards between journals classified as ARRIVE guideline supporters and non-supporters. Furthermore, no studies examining adherence to the ARRIVE guidelines have been conducted in the veterinary literature. We hypothesised that papers published in supporting journals would have greater adherence to the guidelines, and therefore higher reporting standards, than those published in non-supporting journals. Additionally, we hypothesised that papers published in supporting journals would show a greater improvement in reporting standards since the guidelines became available. To test these hypotheses the related subjects of anesthetic and analgesic efficacy and animal welfare were selected for study.

2.4.3. Methods and materials

2.4.3.1. Journal and paper selection

Journals were categorised as ARRIVE supporters (SUPP) or non-supporters (nonSUPP) based on whether the ARRIVE guidelines were mentioned in their instructions to authors when beginning the study (November 2016). Editorial offices of SUPP journals confirmed by email that the ARRIVE guidelines were included in the instructions to authors before December 2014. Papers were selected from a selection of journals from these two categories (SUPP and nonSUPP) from two years: 2009 (pre-ARRIVE) and 2015 (post-ARRIVE). SUPP journals were: Journal of the American Association for Laboratory Animal Science, Comparative Medicine, Animal Welfare, Laboratory Animals and Alternatives to Animal Experimentation. NonSUPP journals were: Applied Animal Behaviour Science and Experimental Animals. Journals were selected based on an initial search for those publishing papers on the predetermined subjects of interest (welfare, analgesic and anesthetic efficacy). Additionally, none of the selected journals had previously been included in a study assessing adherence to the ARRIVE guidelines.

An initial screening of all papers was performed by a single author (VL) by manual search of tables of contents, using titles, abstracts and keywords to identify relevant papers. Papers were selected based on subject and study type. A second screening was performed by two authors (VL and FRB) during the full text evaluation of the selected papers. Anesthesia or analgesia papers described studies assessing the efficacy of anesthetics or analgesics as a primary objective. Animal welfare papers described studies where the objective was to improve the well-being of animals used in research. Only prospective *in vivo* studies were included. Case studies were excluded.

2.4.3.2. Evaluation

Evaluation of adherence to the ARRIVE guidelines was performed independently by two authors (VL and FRB). The ARRIVE checklist (Kilkenny *et al*., 2010) of 20 items and 46 associated sub-items was operationalised and used as the basis for evaluation (Table 2.4.3.2.1). Descriptors were developed by consensus to promote consistency during evaluation (Table 2.4.3.2.1). Items without associated sub-items were categorised as either not reported, partially reported or fully reported. Items with sub-items were categorised as not reported if no sub-items were reported, partially reported if only some sub-items were reported and fully reported if all sub-items were reported. For example, for Item 6 (Study design, Table 2.4.3.2.1), the item would only be classified as fully reported if all sub-items (6a-d) were reported, otherwise it would be classified as partially (3 or fewer sub-items reported) or not reported (none of the 4 sub-items reported).

A sub-item was added to the original ARRIVE checklist to clarify drug use (sub-item 7e, Table 2.4.3.2.1). Where items or sub-items were considered not applicable, no score was entered. For example, a paper on zebra fish would have the sub-items bedding materials, access to water and humidity classed as not applicable.

Item and sub-item scores were compared between authors and differences resolved by consensus (with DP).

Table 2.4.3.2.1. The ARRIVE guidelines checklist: operationalised items and sub-items to facilitate assessment of reporting (Kilkenny *et al*., 2010).

Item/sub-item	ARRIVE items and sub-items	Possible categories	Descriptor
1	**Title**	**not reported; partially reported; fully reported**	Accurate and concise description of article content
2	**Abstract**	**not reported; partially reported; fully reported**	Accurate summary of background, research objectives, species or strain of animal used, key methods, principle findings, and conclusions
	Introduction		
3	**Background**	**depends on sub-items**	**-**
3a	Motivation for and context of study	not reported; reported	Sufficient scientific background (with references) on motivation and context of study, with explanation of experimental approach and rationale
3b	Animal species and models justified	not reported; reported	Explain how and why animal species and models were chosen
4	**Objectives**	**not reported; partially reported; fully reported**	Objectives or hypotheses of study are clearly described
	Methods		
5	**Ethical Statement**	**not reported; fully reported**	Statement to indicate ethical review permissions, relevant licenses and national or institutional guidelines for care and use of animals
6	**Study design**	**depends on sub-items**	**-**
6a	Number of groups	not reported; reported; N/A	Number of experimental and control groups clearly stated; N/A if single group study
6b	Randomisation	not reported; reported; N/A	Statement that randomisation was used or justification for no randomisation; N/A if single group study
6c	Blinding	not reported; reported; N/A	Statement that blinding was used or justification for no blinding; N/A if single group study. Classified as "reported" if blinding was mentioned for any step (*e.g.* blinding to allocation, blinding to outcome assessment, treatment administration *etc.*).

6d	Experimental unit	not reported; reported	Reader is able to understand if comparisons were between a single animal or a group of animals
7	**Experimental procedures**	**depends on sub-items**	**-**
7a	How	not reported; reported	Description of experiment performed, and details of specialised equipment used can be replicated with the information present
7b	When	not reported; reported; N/A	Statement of when during the day the procedures took place and when according to the experimental timeline; N/A if paper was assessing continuous assessment or if light cycle unlikely to affect assessment (*e.g.* lameness)
7c	Where	not reported; reported	Some indication of where each procedure took place
7d	Why	not reported; reported	Rationale for why chosen experimental procedures were performed
7e	Drugs used	not reported; reported	Statement of the name, dose, route, and frequency of the analgesics or anesthetics used; N/A if procedures can be obviously performed without analgesic or anesthetics
8	**Experimental animals**	**depends on sub-items**	**-**
8a	Species	not reported; reported	Statement of species used
8b	Strain	not reported; reported	Statement of strain used
8c	Sex	not reported; reported	Statement of sex used
8d	Developmental stage	not reported; reported	Statement of age of animals used
8e	Weight	not reported; reported; N/A	Statement of the animals' weight; N/A for zoo animals
8f	Source	not reported; reported; N/A	Statement of animals' source; N/A for zoo animals
8g	Health/immune status	not reported; reported	Statement of animals' heath (*i.e.* screening of tested animals or sentinel animals for lab animals) or general statement that animals were healthy for farm, companion, and zoo animals
9	**Housing and husbandry**	**depends on sub-items**	**-**

9a	Type of cage/housing	not reported; reported; N/A	Statement of cage dimensions and product source for lab animals and a general description for companion and zoo animals; N/A if paper was on animals being process for slaughter (*e.g*. study at abattoir)
9b	Bedding material	not reported; reported; N/A	Statement of bedding type and source for lab animals and a general description for non-lab animals; N/A for fish species or animals being processed for slaughter
9c	Type of facility	not reported; reported; N/A	Statement of facility type and a general description for non-lab animal; N/A if paper was on animals being process for slaughter
9d	Number of cage companions	not reported; reported; N/A	Statement of number of animals housed together or individually; N/A if paper was on animals being processed for slaughter
9e	Light/dark cycle	not reported; reported; N/A	Statement of time lights were on/off for lab animals; information of place of facility and time of experiment is accepted as an alternative for farm and zoo animals*; N/A if paper was on animals being process for slaughter
9f	Temperature	not reported; reported; N/A	Statement of temperature animals were housed in; information of place of facility and time of experiment is acceptable as an alternative for farm and zoo animals*; N/A if paper was on animals being process for slaughter
9g	Type of food	not reported; reported; N/A	Statement of food type and sources for lab animals; general description (*e.g*. hay for cattle) acceptable for non-lab animals; N/A if paper was on animals being process for slaughter
9h	Water access	not reported; reported; N/A	Statement that water was provided; N/A for fish species or animals being processed for slaughter
9i	Environmental enrichment	not reported; reported; N/A	Statement that a form of enrichment was provided; N/A if paper was on animals being processed for slaughter

9j	Humidity	not reported; reported; N/A	Statement of humidity for lab animals; information of place and time of experiments is acceptable as an alternative for farm and zoo animals*; N/A for fish species or animals being processed for slaughter
9k	Welfare assessment	not reported; reported; N/A	Statement that a form of welfare assessment was in place; point was awarded by default if the paper was a welfare paper; N/A if the intervention performed was not for the benefit of the animals involved
9l	Welfare interventions	not reported; reported; N/A	Statement of what type of welfare intervention prepared; intervention must be in response to animals' well-being and not from an outcome of the experiment *e.g.* Eye issues from eye procedure vs. Weight loss; N/A if no adverse event is expected (*i.e.* animal assessed after death)
9m	Time of welfare assessment or intervention	not reported; reported; N/A	Statement of when welfare assessment or intervention occurred; N/A if no adverse event expected (*e.g.* study was assessing a new enrichment)
10	**Sample size**	**depends on sub-items**	**-**
10a	Total number of animals used	not reported; reported	Statement specifying in absolute numbers of the total number of animals used in each experiment and treatment groups
10b	Sample size calculation	not reported; reported; N/A	Statement that sample size calculation was performed; N/A if pilot study
10c	Number of independent replications**	reported; N/A	Statement of the number of independent replications performed
11	**Allocating animals**	**depends on sub-items**	**-**
11a	Allocation method	not reported; reported; N/A	Statement of how animals were allocated to groups, including randomisation or matching if done; N/A if single treatment group
11b	Treatment and assessment of animals	not reported; reported	Describe the order in which the animals in the different experimental groups were treated and assessed

12	**Experimental outcomes**	**not reported; partially reported; fully reported**	Define the primary and secondary experimental outcomes assessed
13	**Statistical methods**	**depends on sub-items**	**-**
13a	Details of statistical methods used	not reported; reported	Statistical tests performed for each analysis was clear
13b	Specify unit of analysis	not reported; reported	Unit of analysis was clear for each data set
13c	Assess normality	not reported; reported	Statement that assessment of normality was performed
	Results		
14	**Baseline data**	**not reported; fully reported**	Statement to report relevant characteristics and health status of animals were collected
15	**Numbers analysed**	**depends on sub-items**	**-**
15a	Animals included	not reported; reported	Statement of the number of animals included/excluded in absolute numbers
15b	Reasons for animal exclusion	not reported; reported; N/A	Statement detailing why animals were excluded; N/A if no animals excluded
16	**Outcomes and estimation**	**not reported; partially reported; fully reported**	Results for each analysis was clear with a measure of precision (*e.g*. standard error or confidence interval)
17	**Adverse events**	**depends on sub-items**	**-**
17a	Details of adverse events	not reported; reported; N/A	Reported details of adverse events that occurred or a statement to report no adverse events occurred; N/A if no adverse events expected
17b	Modifications to reduce adverse events	not reported; reported; N/A	Modifications to experimental procedures made to reduce adverse events were described; N/A if no adverse event expected
	Discussion		
18	**Interpretation/scientific implications**	**depends on sub-items**	**-**
18a	Interpretation	not reported; reported	Interpret results, taking into account study objectives and hypotheses, current theory and other relevant studies in literature

18b	Study limitations	not reported; reported	Commented on the study limitations including potential sources of bias, any limitations of the animal model and the imprecision associated with results
18c	Implications for 3Rs of animal use	not reported; reported; N/A	Described any implications of experimental methods or findings for the replacement, refinement or reduction (3Rs) of the use of animals in research; point was awarded if it was a welfare paper; N/A if assessing anatomic response to an analgesic or anesthetic (*e.g.* buprenorphine effects on limb volume)
19	**Generalisability/translation**	**not reported; fully reported; N/A**	Commented on whether the findings of this study are likely to translate to other species or systems, including any relevance to human biology; N/A for welfare paper unless specified in discussion
20	**Funding**	**not reported; fully reported**	List all funding sources and the role of the funder(s) in the study

Legend: Items are bolded and listed with a number. Sub-items are listed with a number and letter. *Acceptable to report only place and time of year for 9e) light/dark cycle; 9f) temperature; 9j) humidity as this information can be inferred if animals (production and zoo types) are housed outdoors ** Number of independent replications was scored as not applicable (N/A) when not reported as this sub-item was not required for a complete study.

2.4.3.3. Statistics

Each paper was assessed against the 20 items of the ARRIVE guidelines, generating percentages of fully reported items. From this, mean percentages of items were calculated for each journal type during each publication year. Following Levene's test revealing heterogeneity of variances, an unequal variance t-test was used to compare these mean percentages between journal types (SUPP 2009 vs nonSUPP 2009; SUPP 2015 vs nonSUPP 2015) and between years (SUPP 2009 vs. SUPP 2015; nonSUPP 2009 vs. nonSUPP 2015). Correction for multiple comparisons was not applied as comparisons between identical items were viewed as independent from other items. The overall quality of item reporting was classified as well-reported (> 80%), average (50-80%) or poor (< 50%; Delgado-Ruiz *et al.*, 2014). For each journal type, the percentages of individual items and sub-items that were fully, partially or not reported were compared between years with a chi-square test. Additionally, to provide an overall impression of reporting standards in 2015 data from both journal types were pooled.

2.4.4. Results

After initial screening, 271 papers were identified. Thirty-five papers were excluded following full text evaluation, leaving 236 papers included in the final analysis (SUPP 2009: n = 52; SUPP 2015: n = 61; nonSUPP 2009: n = 68; nonSUPP 2015: n = 55, Fig 2.4.4.1). One item and one sub item (generalisability/translation (item 19), number of independent replication (sub-item 10c)) were removed before analysis as they were only applicable in a small number of papers (4/236 and 10/236, respectively).

The percentages of fully reported items between journal types were similar in 2009 (p = 0.07) and 2015 (p = 0.89; Table 2.4.4.1). The percentage of fully reported items increased significantly from 2009 to 2015 for both SUPP (p = 0.02) and nonSUPP (p = 0.0001; Table 2.4.4.1) journals. Although both journal types showed improvements from 2009 to 2015, neither improved significantly more than the other (absolute difference in change between nonSUPP – SUPP = 3.3%, p = 0.09 [95% CI -0.5 – 4.3%]).

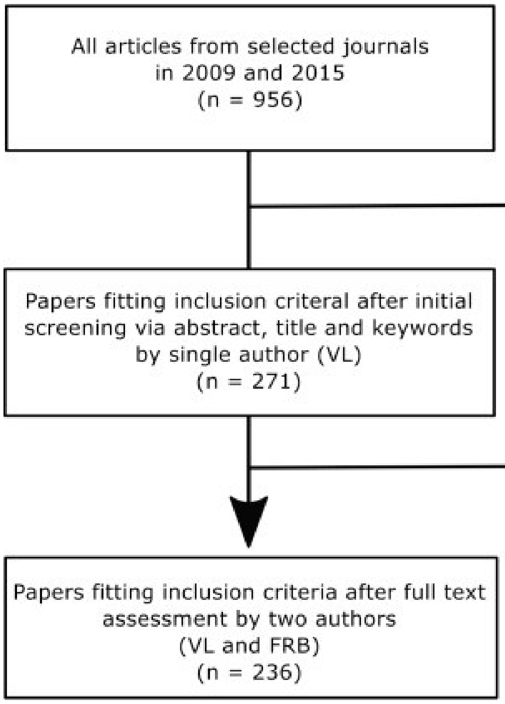

All articles from selected journals
in 2009 and 2015
(n = 956)
Exclude: reviews, commentaries, in vitro studies
and studies that did not fit criteria
(n = 685)
Papers fitting inclusion criteral after initial
screening via abstract, title and keywords
by single author (VL)
(n = 271)
Exclude: case studies, non-prospective studies
and studies that did not fit criteria.
(n = 35)
Papers fitting inclusion criteria after full text
assessment by two authors
(VL and FRB)
(n = 236)

Fig 2.4.4.1 Flow diagram of paper selection process.

Legend: Papers were selected from studies reporting research in anesthesia, analgesia and animal welfare from 5 veterinary journals.

Table 2.4.4.1: Overall reporting quality in journals supporting (SUPP) and not supporting (nonSUPP) the ARRIVE guidelines for 2009 and 2015.

	2009 (%)	2015 (%)	[b]p-value [95% CI]
SUPP	55.3 ± 11.5	60.5 ± 11.2	0.02 [1.0 – 9.4]
Non-SUPP	51.8 ± 9.0	60.2 ± 10.0	0.0001 [5.0 – 11.8]
[a]p value [95% CI]	0.07 [-0.3 – 7.3]	0.89 [-3.6 – 4.2]	

Legend: Values ae mean percentages of fully reported items. The numbers of papers examined were: SUPP 2009; n = 52, SUPP 2015; n = 61, nonSUPP 2009; n = 68, nonSUPP 2015; n = 55. [a]p values of differences between journal types within the same year. [b]p-values of differences between years for the same journal type. 95% confidence interval (95% CI) is for the mean difference.

2.4.4.1. Items

Despite minimal improvements in overall reporting standards between 2009 and 2015, several individual items showed significant improvement in full reporting. For SUPP journals, these items were the abstract (from 69.2 to 91.8%, p = 0.003), housing and husbandry (from 3.9 to 21.3%, p = 0.01) and sample size (from 3.8 to 21.3%, p = 0.01; Table 2.4.4.1.1). For nonSUPP journals, the following items were increasingly fully reported from 2009 to 2015: ethical statement (from 36.8 to 81.8%, p < 0.0001); experimental animals (from 1.5 to 10.9%, p = 0.04) and interpretation/scientific implications (from 10.3 to 38.2%, p = 0.0004; Table 2.4.4.1.1).

In SUPP journals, sample size was reported at least partially by all papers in 2009 but was not reported in 9.8% of papers in 2015 (p = 0.03, Table 2.4.1.1.1 and Table 2.4.1.1.2. In both SUPP and nonSUPP journals, items that were frequently not reported in both 2009 and 2015 were baseline data, numbers analysed and funding.

Pooling the percentage of fully reported items in 2015 from both journal types revealed that items with excellent (> 80%), average (50-80%) and poor (< 50%) reporting was distributed in to thirds (Fig 2.4.4.1.1). Title, abstract, background, objectives, ethical statement, experimental outcomes, and outcomes and estimation were well reported. In contrast, ethical statement, baseline data, numbers analysed, adverse events and funding were poorly reported.

2.4.4.2. Sub-items

There were significant improvements in percentages of papers reporting a small number of sub-items between years for each journal type though overall levels of reporting remained low (Table 2.4.4.2.1). Notably amongst these were sub-items associated with bias: blinding (sub-item 6c), sample size calculation (sub-item 10b), allocation method (sub-item 11a) and data handling (sub-item 15b) (Fig 2.4.4.2.1) Randomisation (sub-item 6b) was alone in being reported more than 50% of the time (Fig 2.4.4.2.1).

Table 2.4.4.1.1: Papers fully reporting ARRIVE checklist items in supporting (SUPP) and non-supporting (nonSUPP) journals in 2009 and 2015.

Item		SUPP			NonSUPP		
		2009 (N = 52)	2015 (N = 61)		2009 (N = 68)	2015 (N = 55)	
		n/N (% reported)	n/N (% reported)	p-value	n/N (% reported)	n/N (% reported)	p-value
1	Title	52/52 (100)	61/61 (100)	1	68/68 (100)	55/55 (100)	1
2	Abstract	36/52 (69.2)	56/61 (91.8)	**0.003**	45/68 (66.2)	44/55 (80.0)	0.11
3	Background	52/52 (100)	60/61 (98.4)	1	68/68 (100)	55/55 (100)	1
4	Objectives	47/52 (90.2)	60/61 (98.4)	0.09	68/68 (100)	55/55 (100)	1
5	Ethical statement	39/52 (75.0)	52/61 (85.2)	0.23	25/68 (36.8)	45/55 (81.8)	**<0.0001**
6	Study design	10/52 (19.2)	19/61 (31.1)	0.20	10/68 (14.7)	15/55 (27.3)	0.12
7	Experimental procedure	34/52 (65.4)	30/61 (49.2)	0.09	45/68 (66.2)	42/55 (76.4)	0.24
8	Experimental animals	8/52 (15.4)	18/61 (29.5)	0.12	1/68 (1.5)	6/55 (10.9)	**0.04**
9	Housing and husbandry	2/51 (3.9)	13/61 (21.3)	**0.01**	3/67 (4.5)	8/54 (14.8)	0.06
10	Sample size	2/52 (3.8)	13/61 (21.3)	**0.01**	1/68 (1.5)	3/55 (5.5)	0.32
11	Allocating animals	11/52 (21.2)	16/61 (26.2)	0.66	14/68 (20.6)	17/55 (30.9)	0.22
12	Experimental outcomes	52/52 (100)	61/61 (100)	1	66/67 (98.5)	55/55 (100)	1
13	Statistical methods	23/52 (44.2)	29/61 (47.5)	0.85	38/68 (55.9)	32/55 (58.2)	0.86
14	Baseline data	24/41 (58.5)	27/50 (54.0)	0.68	20/30 (66.7)	18/35 (51.4)	0.31
15	Numbers analysed	29/52 (55.8)	39/61 (63.9)	0.44	37/68 (54.4)	25/55 (45.5)	0.37
16	Outcomes and estimation	45/52 (86.5)	49/61 (80.3)	0.45	55/68 (80.9)	49/55 (89.1)	0.32
17	Adverse events	18/29 (62.1)	17/41 (41.5)	0.15	4/18 (22.2)	8/23 (34.8)	0.50
17a	Details of adverse events	25/29 (86.2)	25/41 (61.0)	**0.03**	8/18 (44.4)	20/24 (83.3)	**0.02**
18	Interpretation/scientific implications	15/52 (28.8)	20/61 (32.8)	0.69	7/68 (10.3)	21/55 (38.2)	**0.0004**
19	Generalisability/translation	-	-	-	-	-	-
20	Funding	29/52 (55.8)	43/61 (70.5)	0.12	48/68 (70.6)	44/55 (80)	0.30

Legend: N = total number of papers where the item was applicable. n = total number of papers reporting the item. p values are for comparisons between years for each journal type.

Table 2.4.1.1.2. Papers partially reporting ARRIVE checklist items in supporting (SUPP) and non-supporting (nonSUPP) journals in 2009 and 2015.

Item		SUPP			NonSUPP		
		2009 (N = 52)	2015 (N = 61)		2009 (N = 68)	2015 (N = 55)	
		n/N (% reported)	n/N (% reported)	p-value	n/N (%	n/N (% reported)	p-value
1	Title	0/52 (0)	0/61 (0)	1	0/68 (0)	0/55 (0)	1
2	Abstract	16/52 (30.8)	5/61 (8.2)	0.003	23/68 (33.8)	11/55 (20.0)	0.11
3	Background	0/52 (0)	1/61 (1.6)	1	0/68 (0)	0/55 (0)	1
4	Objectives	1/52 (2.0)	0/61 (0)	0.46	0/68 (0)	0/55 (0)	1
5	Ethical statement	0/52 (0)	0/61 (0)	1	0/68 (0)	0/55 (0)	1
6	Study design	42/52 (80.8)	42/61 (68.9)	0.20	58/68 (85.3)	40/55 (72.7)	0.12
7	Experimental procedure	28/52 (34.6)	31/61 (50.8)	0.09	23/68 (33.8)	13/55 (23.6)	0.24
8	Experimental animals	44/52 (84.6)	43/61 (70.5)	0.12	67/68 (98.5)	49/55 (89.1)	0.04
9	Housing and husbandry	49/51 (96.1)	48/61 (78.7)	0.01	64/67 (95.5)	46/54 (85.2)	0.06
10	Sample size	50/52 (96.2)	42/61 (68.9)	0.0002	67/68 (98.5)	51/55 (92.7)	0.17
11	Allocation animals	41/52 (78.8)	45/61 (73.8)	0.66	54/68 (79.4)	38/55 (69.1)	0.22
12	Experimental outcomes	0/52 (0)	0/61 (0)	1	1/67 (1.5)	0/55 (0)	1
13	Statistical methods	27/52 (51.9)	30/61 (49.2)	0.85	25/68 (36.8)	23/55 (41.8)	0.58
14	Baseline data	1/41 (2.4)	0/50 (0)	0.45	0/30 (0)	0/35 (0)	1
15	Numbers analysed	4/52 (7.7)	2/61 (3.3)	0.31	3/68 (4.4)	2/55 (3.6)	1
16	Outcomes and estimation	6/52 (11.5)	11/61 (18.0)	0.43	13/68 (19.1)	6/55 (10.9)	0.32
17	Adverse events	7/29 (24.1)	9/41 (2.0)	1	4/18 (22.2)	10/23 (43.5)	0.2
18	Interpretation/scientific implications	37/52 (71.2)	41/61 (67.2)	0.69	61/68 (87.7)	34/55 (61.8)	0.0004
19	Generalisability/translation						
20	Funding	0/52 (52.0)	0/61 (0)	1	0/68 (0)	0/55 (0)	1

Legend: N = total number of papers where the item was applicable. n = total number of papers partially reporting the item. p values are for comparisons between years for each journal type.

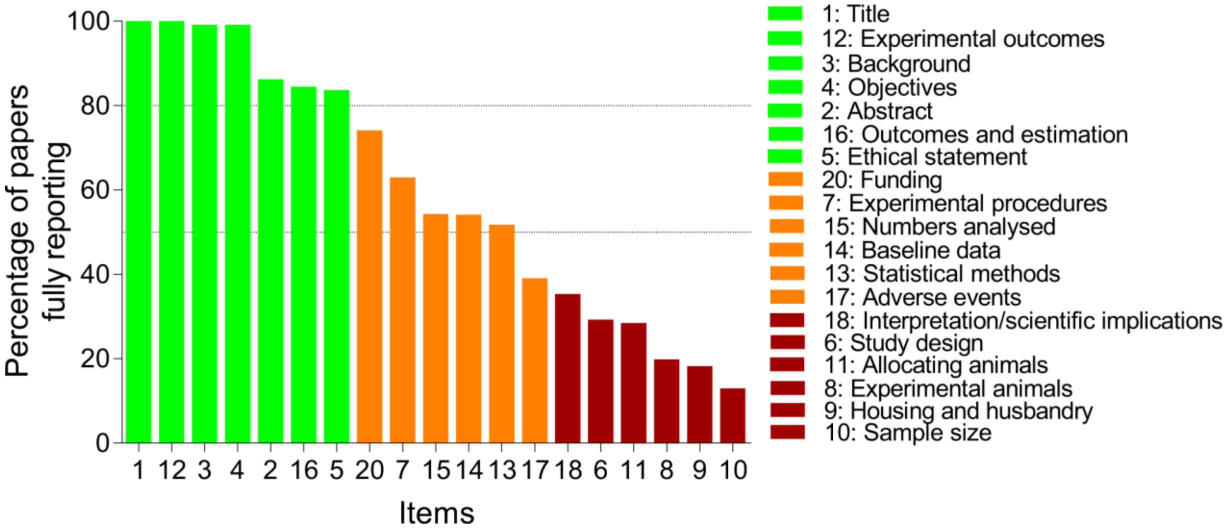
Table 2.4.4.2.1 Papers fully reporting ARRIVE checklist sub-items in supporting (SUPP) and non-supporting (nonSUPP) journals in 2009 and 2015.

Items	Sub-items				SUPP			NonSUPP
			2009 (N = 52)	2015 (N = 61)		2009 (N = 68)	2015 (N = 55)	
			N (% reported)	n (% reported)	p-value	n (% reported)	n (% reported)	p-value
Background	3a	Motivation and context of study	52/52 (100)	61/61 (100)	1	68/68 (100)	55/55 (100)	1
	3b	Animal species and model justification	52/52 (100)	60/61 (98.4)	1	68/68 (100)	55/55 (100)	1
Study design	6a	Number of groups	52/52 (100)	61/61 (100)	1	68/68 (100)	55/55 (100)	1
	6b	Randomisation	29/46 (63.0)	37/52 (71.2)	0.52	38/67 (56.7)	34/48 (70.8)	0.17
	6c	Blinding	13/52 (25.0)	24/60 (40.0)	0.11	10/68 (14.7)	15/53 (28.3)	0.08
	6d	Experimental units	52/52 (100)	61/61 (100)	1	68/68 (100)	55/55 (100)	1
Experimental procedure	7a	How	52/52 (100)	61/61 (100)	1	68/68 (100)	55/55 (100)	1
	7b	When	31/44 (70.5)	36/60 (60.0)	0.31	47/68 (69.1)	45/54 (83.3)	0.09
	7c	Where	45/52 (86.5)	55/61 (90.2)	0.57	67/68 (98.5)	52/55 (94.5)	0.33
	7d	Why	52/52 (100)	61/61 (100)	1	68/68 (100)	55/55 (100)	1
	7e	Drugs used	20/20 (100)	29/33 (87.9)	0.17	5/6 (83.3)	9/10 (90)	1
Experimental animals	8a	Species	52/52 (100)	61/61 (100)	1	68/68 (100)	55/55 (100)	1
	8b	Strain	48/52 (92.3)	61/61 (100)	0.04	62/68 (91.2)	49/54 (90.7)	1
	8c	Sex	46/52 (88.5)	60/61 (98.4)	0.05	55/68 (80.9)	50/55 (90.9)	0.13
	8d	Developmental stage	38/52 (73.1)	49/61 (80.3)	0.38	49/68 (72.1)	45/55 (81.8)	0.29
	8e	Weight	28/49 (57.1)	32/58 (55.2)	0.85	26/63 (41.3)	23/50 (46)	0.70
	8f	Source	34/52 (65.4)	52/61 (85.2)	0.02	34/68 (50)	32/55 (58.2)	0.47
	8g	Health/immune Status	19/52 (36.5)	34/61 (55.7)	0.06	7/68 (10.3)	7/55 (12.7)	0.78
Housing and husbandry	9a	Type of cage/housing	42/50 (84.0)	49/55 (89.1)	0.57	59/66 (89.4)	51/53 (96.2)	0.19
	9b	Bedding material	34/48 (70.8)	36/52 (69.2)	1	39/61 (63.9)	42/51 (82.4)	0.04
	9c	Type of facility	23/50 (46.0)	33/54 (61.1)	0.17	39/66 (59.1)	33/53 (62.3)	0.85
	9d	Number of cage companions	47/50 (94.0)	51/55 (92.7)	1	62/66 (93.9)	49/53 (92.5)	1
	9e	Light/dark cycle	36/50 (72.0)	46/55 (83.6)	0.17	22/66 (33.3)	29/53 (54.7)	0.025
	9f	Temperature	34/50 (68.0)	35/55 (63.6)	0.68	21/66 (31.8)	30/53 (56.6)	0.009

	9g	Type of food	45/50 (90.0)	47/54 (87.0)	0.76	63/67 (94.0)	44/53 (83.0)	0.076
	9h	Water access	42/49 (85.7)	43/52 (82.7)	0.79	42/61 (68.9)	41/52 (78.8)	0.29
	9i	Environmental enrichment	25/50 (50.0)	29/54 (53.7)	0.84	35/66 (53.0)	23/53 (43.4)	0.36
	9j	Humidity	25/48 (52.1)	25/53 (47.2)	0.69	12/61 (19.7)	20/52 (38.5)	0.04
	9k	Welfare assessment	45/51 (88.2)	51/59 (86.4)	1	66/68 (97.1)	52/55 (94.5)	0.66
	9l	Welfare interventions	12/27 (44.4)	11/27 (40.7)	1	1/7 (14.3)	8/22 (36.4)	0.38
	9m	Time of welfare assessment or intervention	43/50 (86.0)	52/60 (86.7)	1	66/68 (97.1)	51/55 (92.7)	0.41
Sample size	10a	Total number of animals used	52/52 (100)	56/61 (91.8)	0.06	67/68 (98.5)	53/55 (96.4)	0.59
	10b	Sample size calculation	1/52 (1.9)	8/59 (13.6)	0.04	2/68 (2.9)	3/55 (5.5)	0.66
	10c	Sample size: Number of independent replications	-	-	-	-	-	-
Animal allocation	11a	Allocation method	5/49 (10.2)	7/61 (13.2)	0.76	10/64 (15.6)	10/50 (20)	0.62
	11b	Treatment and assessment of animals	52/52 (100)	61/61 (100)	1	68/68 (100)	55/55 (100)	1
Statistical methods	13a	Details of statistical methods used	50/52 (96.2)	59/61 (96.7)	1	64/68 (94.1)	55/55 (100)	0.13
	13b	Specify unit of analysis	48/52 (92.3)	57/61 (93.4)	1	63/68 (92.6)	55/55 (100)	0.06
	13c	Assess normality	27/52 (51.9)	29/61 (47.5)	0.71	38/68 (55.9)	34/55 (61.8)	0.58
Numbers analysed	15a	Animals included	31/52 (59.6)	42/61 (68.9)	0.33	40/68 (58.8)	27/55 (49.1)	0.36
	15b	Reasons for animal exclusion	13/33 (39.4)	17/37 (45.9)	0.63	18/45 (40)	13/41 (31.7)	0.50
Adverse Events	17a	Details of adverse events	25/29 (86.2)	25/41 (61.0)	0.03	8/18 (44.4)	20/24 (83.3)	0.02
	17b	Modifications to reduce adverse events	8/19 (42.1)	8/30 (26. 7)	0.35	1/15 (6.7)	5/20 (25)	0.21
Interpretation	18a	Interpretation	52/52 (100)	61/61 (100)	1	68/68 (100)	55/55 (100)	1
	18b	Study limitations	16/52 (30.8)	22/61 (36.1)	0.69	7/68 (10.3)	20/55 (36.4)	0.0008
	18c	Implications for 3Rs of animal use	51/51 (100)	61/61 (100)	1	68/68 (100)	55/55 (100)	1

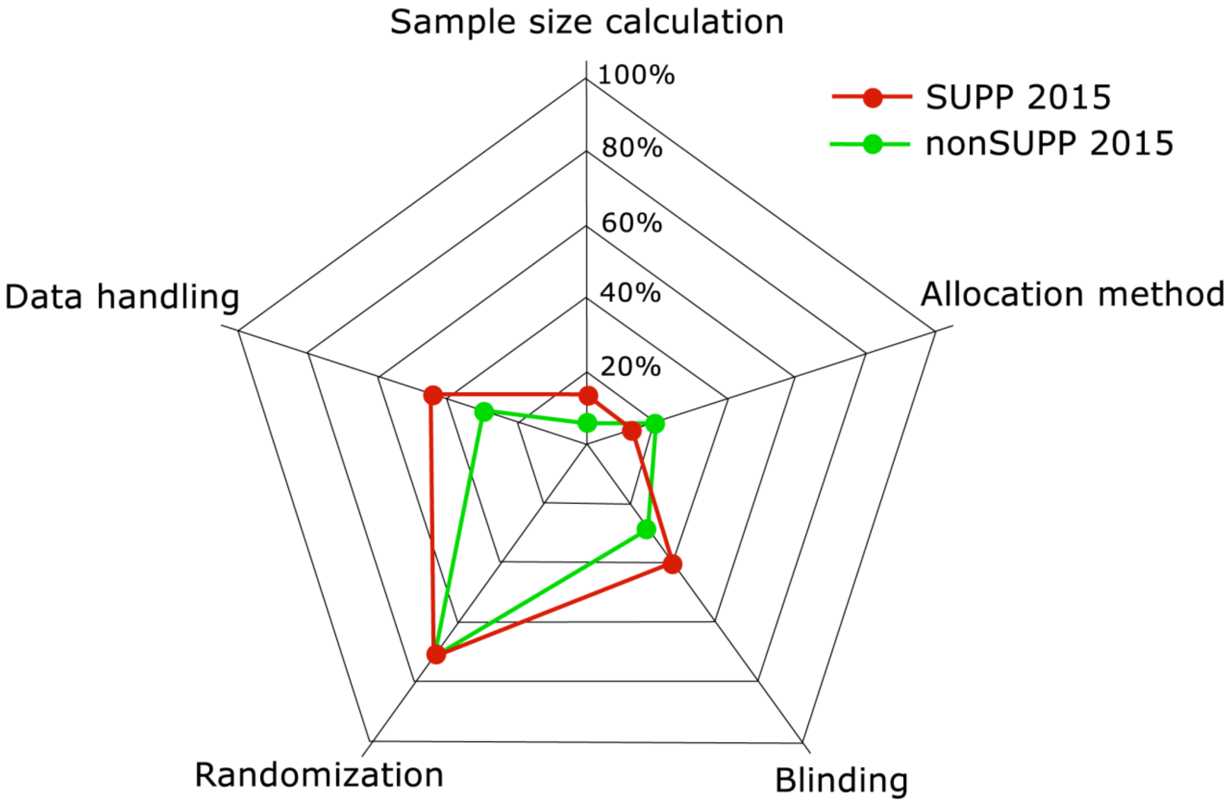
Legend: N = total number of journal articles where the sub-item was applicable; n = total number of journal articles reporting the sub-item. p values are for comparisons between years for each journal type.

Fig 2.4.4.1.1 Bar graph of papers fully reporting individual items from the ARRIVE checklist.



Legend: Data from papers published in 2015 were pooled from ARRIVE supporting (SUPP, n = 61 papers) and non-supporting (nonSUPP, n = 55 papers) journals. Broken horizontal lines indicate reporting quality thresholds: excellent (> 80%), average (50-80%) and poor (< 50%; Delgado-Ruiz *et al*., 2015).

Fig 2.4.4.2.1 Radar plot of ARRIVE checklist sub-items associated with bias reported in ARRIVE supporting (SUPP) and non-supporting (nonSUPP) journals in 2015.



Legend: Sub-items associated with bias reported in ARRIVE supporting (SUPP) and non-supporting (nonSUPP) journals in 2015.

2.4.5. Discussion

Numerous studies across different research fields have shown that reporting quality has remained low since the publication of the ARRIVE guidelines (Schwarz *et al*., 2012; *Baker et al*., 2014; Delgado-Ruiz *et la*., 2015; Ting *et al*., 2015 Avey *et al*., 2016 and Lui *et al*., 2016). This is in spite of large-scale support for the guidelines by biomedical journals and increasing awareness of the financial and ethical cost of irreproducible research (Ioannidis *et la*., 2005; Kilkenny *et al*., 2009; Landis *et al*., 2012; Freedman *et al*., 2015). The results of our study confirm that reporting quality remains low and that journal support for the ARRIVE guidelines has not resulted in meaningful improvements in reporting standards.

2.4.5.1. Adherence to reporting guidelines remains low despite journal support

Reporting standards in this sample of anesthesia, analgesia and animal welfare papers was low, with little indication that the ARRIVE guidelines have made an impact in improving reporting standards. These findings echo those of others (Baker *et al*., 2014; Macleod *et al.,* 2015 and Liu *et al.,* 2016). The data presented here, published 5 years after introduction of the ARRIVE guidelines, reflect the low reporting rates identified by Kilkenny *et al.* (2009) that served as the catalyst for creation of the guidelines. As in those findings, reporting of important indicators of study design quality (randomisation, blinding, sample size calculation and data handling) remain low.

A recent study of the veterinary literature that focused on reporting of randomisation in randomised controlled trials found a higher percentage of papers (49%, n = 106) reporting the allocation method than reported here (13-20% for SUPP and nonSUPP, respectively; Di Girolamo *et al.,* 2017). This difference is likely to have resulted from selecting papers self-describing as randomised clinical trials.

With the small observed increase in reported items in both SUPP and nonSUPP journals, an increased awareness of reporting standards, such as the ARRIVE guidelines, cannot be ruled out. However, these increases were limited, with no significant differences in fully reported items between journal types in 2015 and, perhaps most importantly, the reporting of key sub-items indicating bias (randomisation; sub-items 6b and 11a, blinding; sub-item 6c, animals excluded; sub-item 15b and sample size calculation; sub-item 10b) remained low (Landis *et al*., 2012 and Macleod *et al*., 2015). Similar findings have been reported in surveys of experimental animal models, including acute lung injury, peri-odontology, autoimmunity and neoplasia (Schwarz *et al*., 2012; Baker *et al*., 2014; Avey *et al*., 2016; Liu *et al*., 2016 and Ting *et al*., 2015). Sample size justification, in particular, is consistently poorly reported, with reporting percentages ranging from 0 – 7% (Schwarz *et al*., 2012; Baker *et al*., 2014; Avey *et al*., 2016; Liu *et al*., 2016 and Ting *et al*., 2015). This is an alarming figure given the impact it has on interpretation of findings and animal use (Button *et al*., 2013).

A common feature in this and other studies of ARRIVE guideline adherence has been a lack of enforcement of reporting standards. In contrast, when reporting is mandatory, important improvements have been achieved (Han *et al*., 2017 and Macleod *et al*., 2017). Following a change in editorial policy in 2013, the Nature research journals now require that authors accompany accepted manuscripts with a completed checklist identifying inclusion of key items associated with quality of reporting and study design (Anon, 2013). This checklist has numerous items in common with those of the ARRIVE guidelines. In reviewing approximately 440 papers in each of two groups (those published in the Nature publishing journals and those from other publishers, before and after checklist implementation), the positive effect of the checklist was evident in that reporting of bias criteria (randomisation, blinding, sample size calculation and data handling; Landis *et al*., 2012) improved significantly from 0 to 16.4% (Macleod *et al*., 2017). While this number remains low, the percentage of papers from other publishers reporting these items was < 1% over the same time period. In striking contrast with the findings presented here and elsewhere (Schwarz *et al*., 2012; Baker *et al*., 2014; Avey *et al*., 2016; Liu *et al*., 2016 and Ting *et al*., 2015), introduction of the checklist was associated with a mention of sample size calculation in 58% (90/154) of papers, increasing from < 2% (3/192).

2.4.5.2. Suggestions to improved guideline adherence

To date, a change in editorial policy accompanied by mandatory submission of a reporting checklist is the only method shown to have resulted in an increase in reporting quality (Macleod *et al*., 2017). This clearly indicates that enforcement is required to generate a change in behavior. As others have suggested, achieving change in a well-established process, such as peer-review, is difficult (McGrath and Lilley, 2015). Furthermore, placing the responsibility of policing guideline adherence on reviewers is unrealistic, when they are volunteering their time, usually busy and may share the same view of an unenforced request to complete a checklist (Landis *et al*., 2012 and McGrath and Lilley, 2015).

Other, albeit untested, suggestions to improve reporting standards include: 1. using a template of the methods section to require completion of desired items (McGrath and Lilley, 2015), 2. standardizing reporting of common outcomes by learned societies and research communities (Fisher *et al*., 2009; Ludolph *et al*., 2010 and Shineman *et al*., 2011; Baker and Amor, 2012 and Baker *et al*., 2014) and 3. mandating adherence to reporting standards at the stage of applying for federal authority to conduct research (in countries where this applies), perhaps in the form of study registration (Vogt *et al*., 2016). These suggestions, along with the checklist used by the Nature research journals, represent a shift away from the current format of the ARRIVE guidelines towards a shorter checklist. Irrespective of scope and format, it is clear reporting standards will remain low without some form of enforced adherence (Baker *et al.,* 2014 and McGrath and Lilley, 2015). An important consequence of enforced compliance, which must be considered when selecting a method to improve reporting, is the associated cost (time and financial resources) to publishers and authors and striking an acceptable balance between an ideal and that which is feasible, practical and achievable.

2.4.5.3. Limitations

Our data may have been skewed by the small number of journals in the nonSUPP group and any policies of individual journals on how compliance with the ARRIVE reporting guidelines were assessed. The choice of journals was limited due to the large number that have registered support for the ARRIVE guidelines and our choice of subject matter. While this reflects the success of the ARRIVE guidelines in being widely adopted, our data highlight that the relationship between guideline support and adherence merits investigation (Baker *et al.,* 2014 and Cressey, 2016). Despite the low number of journals included, the risk of systematic journal bias is likely to be low given similar standards of reporting have been documented across a wide range of biomedical journals (Schwarz *et al*., 2012; Baker *et al*., 2014; Delgado-Ruiz *et al*., 2015; Ting *et al.,* 2015; Avey *et al.,* 2016 and Liu *et al.,* 2016).

2.4.5.4. Conclusion

Journal support for the ARRIVE guidelines has not resulted in improved reporting standards, with the lowest levels of reporting associated with factors reflecting potential study bias. To achieve meaningful improvements in reporting standards, as a means to improve study reproducibility and reduce financial and animal waste, enforcement of reporting is necessary.

2.4.6. Acknowledgements

Dr Michael E Mullins (Washington University School of Medicine) for inspiring the title of this paper.

3. Discussion

3.1 Overview

At the beginning of this *PhD*, researchers were beginning to reconsider how animal pain research should be conducted, shifting from an overreliance on nociceptive tests to utilising non-evoked spontaneous behaviours to assess pain (Mogil and Crager, 2004; Rice *et al.,* 2008 and Mogil *et al.*, 2010). One of these behaviours were facial expression scales or "grimace scales", such as the RGS, which has been demonstrated as a viable pain assessment method applicable for a variety of pain types (Sotocinal *et al*., 2011; Liao *et al.,* 2014 and Akintola *et al*., 2017). It is evident that a validated pain scale is required to ensure animal welfare of laboratory rodents and to perform pain research studies in rodents. It is also evident that these pain research studies must be well performed and reported. However, it was also recognised that current reporting standards in preclinical research are poor (Kilkenny *et al.,* 2009). This prompted the creation and publication of the ARRIVE guidelines (Kilkenny *et al*., 2010) which has since garnered tremendous support and endorsement by various journals and related stakeholders.

The MGS and RGS were developed in 2010 and 2011 respectively and were proposed as assessments of ongoing pain in rodents (Langford *et al*., 2010 and Sotocinal *et al.,* 2011). The MGS demonstrated robustness at assessing acute inflammatory pain that lasted less than a day but not for neuropathic pain or pain that lasted for more than a day (Langford *et al.*, 2010). It was assumed that the MGS and RGS may only be able to assess acute pain types. However, it was later demonstrated that rodent grimace scales could be used to assess neuropathic and inflammatory pain of that lasted for more than a day (Akintola *et al.,* 2017 and Leung *et al.,* in press). In general, the rodent grimace scales seem to be viable tools for the evaluation of a wide array of pain types and duration. However, the RGS was still a recent development and the possible applications and limitations of the RGS was still unknown. Therefore, this thesis focused on assessing the various applications and possible limitations of the RGS as a research and clinical tool. In addition, this thesis also assessed the impact of the ARRIVE guidelines, a reporting guideline to guide authors on what should be reported in *in vivo* experimental studies, on reporting standards of papers published five years after the ARRIVE guidelines were published.

Subsequently, the studies described in this thesis demonstrated that the RGS can be a practical tool for both research and clinical use with application of real-time scoring. The application for real-time scoring and an intervention threshold (Oliver *et al.,* 2014) allows both researches and animal care personnel the ability to quickly assess if a rat is in pain and determine if analgesic intervention is required. It was also demonstrated that the RGS is an applicable tool for a wider array of pain types, as the RGS can also be used to assess acute and chronic visceral pain in a DSS colitis model. This suggests that the RGS can be used as a pain assessment method in other visceral pain models, thereby expanding its usefulness in a research and clinical setting. Additionally, it was highlighted that training and assessment of proficiency in scoring with the RGS is important before a trainee rater begins scoring. This will ensure that trainee raters score similarly and reliably in comparison to an experienced rater. It was also demonstrated that proficiency and reliability is maintained four years after training. Lastly, it was demonstrated that the publication of the ARRIVE guidelines have not improved reporting standards in a meaningful way. This implies that journal support for the ARRIVE guidelines is not sufficient and enforcement of the ARRIVE guidelines may be required for improvements.

3.2 Contributions to the field of pain, study limitations and future work

3.2.1 Real-time application of the RGS

This study has demonstrated that real-time RGS scoring is possible. This means that an observer can simply observe the animal in real-time and assess its pain. Researchers are no required to go through the tedious steps of the standard method: video-recording of the animals and manual image extraction and cropping (Sotocinal *et al.,* 2011). Real-time application of the RGS therefore drastically decreases the time and labour required, increasing the practicality of this tool in a research setting. Furthermore, when observations can be performed in minutes (10 minutes or less), the RGS can now be applied in a clinical setting and lends itself as a welfare tool as rats can be assessed quickly and if it is required, an intervention can be performed.

As observed in this project, a blinded and trained observer could differentiate between rats treated with analgesics or saline after a painful intra-plantar injection of carrageenan, thus demonstrating construct validity. Furthermore, the observer scored similarly with both real-time and standard methods and demonstrates criterion validity. This differs from a previous study that reported significantly lower scores with real-time assessment in comparison to standard method MGS scores (Miller and Leach, 2015). This difference in results may be because of species differences or because the authors only performed three 5s observations over a 10-minute period in their assessment of real-time scoring. In contrast, this study observed each rat with 18 separate time point or 15s interval observations. In this study, a single point or 15s interval observation was the most direct comparison to the study performed by Miller and Leach (2015). When a single point or 15s interval observation was randomly selected in this study, it was observed that the scores were very variable. Therefore, this suggests that a single or a short observation period is inadequate and may explain the differences in results between the Miller and Leach study and the one in this thesis.

There have also been other developments that have refined the RGS and made it more practical, these include the RFF and the *a*MGS (Sotocinal *et al.,* 2011 and Tuttle *et al.,* 2018). The RFF and *a*MGS are computer programs that automatically collect images from videos and automatically scores images respectively. When the two programs are used in conjunction with each other, they can greatly reduce the labour and time required by a researcher to obtain and score images. However, these programs in their current state still need time to run before images are collected and scored. This means that hours or days will still pass before a score for an animal is obtained. Furthermore, the current version of the *a*MGS only scores images binarily as 'pain' or 'no pain' (Tuttle *et al*., 2018). It is also unable to score images with intermediate scores with high confidence and performs poorer than human raters in this regard. This means that the aMGS cannot be used to identify animals that are beginning to experience pain or are currently experiencing low levels of pain. Therefore, early pain identification and analgesic intervention is not possible. While these two programs reduce the manual work required to obtain scores and therefore increases the RGS' practicality as a research tool, they cannot be used in a clinical setting because they do not allow for early identification of animals in pain. It is likely that with time, these two computer programs can be improved to allow immediate and real-time identification of rats with low levels of pain. However, trained observers will still be required to assess the rats in real-time and decide if an intervention is appropriate. Lastly, another advantage of real-time scoring over the RFF and *a*MGS is that video recording equipment and computer programs are not required to perform a pain assessment.

This study (Leung *et al*., 2016) implies that researchers, animal technicians or veterinarians can assess pain cage-side with real-time RGS scoring. This refinement to the standard method of RGS scoring drastically reduces the labour and time required. This will hopefully encourage use of the RGS as a research tool and a clinical tool that will improve animal welfare of laboratory rats by allowing a quick and practical way to identify and assess pain.

A limitation of this study is that rats utilised in this study were habituated to the observer and the observation box and were assessed in a quiet room. It is unknown if rats will still show similar facial expressions when in the presence of an unknown observer and in a conventional housing environment. It has been reported that mice display lower MGS scores if the olfactory cues of a male observer were present (Sorge *et al*., 2014). It is unknown if this is a species-specific limitation or if rats will react similarly. Rat were also assessed individually in the observation box, and it is unknown if RGS scoring would be affected by the presence of cage mates nearby. It has been observed that mice will display changes in facial expressions which are similar to a painful grimace during an aggressive encounter with an intruder conspecific (Defensor *et al*., 2012). This demonstrates that changes in facial expressions in rodents are not always pain-specific and observers need to consider the environmental context the animal is in when scoring. Furthermore, it has been observed that rodents understand pain behaviours and have empathy for their conspecifics: rats are highly motivated to press a lever to help a cage mate out of an uncomfortable situation (*i.e.* stuck in a small box, Ben-Ami Bartal *et al*., 2011), mice are more likely to display pain behaviours in the presence of conspecifics (*i.e.* writhing, Langford *et al*., 2006) and rats will avoid an area if shown a picture of a rat with a pain face (Nakashima *et al*., 2015). Additionally, real-time scoring of the RGS requires the observer to be competent at scoring quickly and accurately and consultation of a manual is not possible. Therefore, it is unknown how much training is required for an observer to be proficient at scoring in real-time. Lastly, with real-time scoring of the RGS, it is not possible to check the accuracy and reliability of the observer. Therefore, video recordings should always be performed concurrently so that they may be used to assess inter- and intra-rater reliability later.

The future work for real-time scoring should be to investigate the perceived limitations. Namely, it should be assessed if rats will still display the same facial expressions if observed by an unknown observer and if the presence of male olfactory cues will have an effect (Sorge *et al.*, 2014). Therefore, a study examining the effects of an observer's gender and familiarity to the rats is warranted. Unfamiliar male and female observers should enter a room separately with the rat during baseline testing and after an acute pain model. This should then be repeated, after a washout and habituation period and the RGS scores from both scenarios compared. This will answer if the observers' gender affects the grimace score displayed and if habituation is required before real-time observations are performed. Additionally, it should also be assessed if pain assessment is still possible when performed in a conventional housing environment and in the presence of conspecifics. It would be interesting to assess if observers can walk up to a rat's home cage and score facial expressions and if the approach of an observer will influence facial expression and behaviour. The current use of polycarbonate cages is likely to be a limiting factor as they are prone to scratching in use, impeding view of the interior. The use of a camera inside of the cage may be an alternative to allow remote monitoring as well as avoid any observer effects. Furthermore, the presence of conspecifics in the same housing room or cage may affect the displayed facial expression. Lastly, the effect of training on real-time RGS scoring should be considered to ensure that new observers are able to score proficiently and quickly in real-time. This may be evaluated by first having the trainee observer go through the training protocol described by Zhang *et a/.* (2019). Then both trainee and experienced observer should score a video while it is playing at a normal speed and assess if both observers score similarly. This will evaluate if proficiency of scoring still images is enough to translate to proficiency of scoring in real-time or if additional training is required for real-time scoring.

3.1.1 Applying the RGS to a secondary model: acute and chronic visceral pain in a colitis model

This study (Leung *et al*., in press) has demonstrated that the use of RGS scoring allows an observer to assess and characterise acute and chronic visceral pain from a DSS-colitis model. The RGS scores increase with exposure to DSS during both the acute and chronic phases. This increase of RGS scores mimicked the increase observed with DAI scores, demonstrating that pain is present when DAI scores increase or when clinical signs are present. This association between RGS and DAI scores also suggest that the DAI can be used as a proxy pain measure. Burrowing in DSS-treated rats also decreased from baseline when the DAI scores increased, however, this decrease was not sustained during the chronic phase. In addition, it was no different from control rats. This questions whether it can discriminate between control rats and rats experiencing pain. It is unknown why the decrease in burrowing did not continue during the chronic phase since the RGS and DAI scores remained high. It is possible that because burrowing is a highly motivated and self-rewarding behaviour in rats, the rats attenuated to the pain and burrowed. The CBS scores were too variable to allow differences between treatment and from controls to be identified. Therefore, this study has demonstrated that the RGS and burrowing can be used to assess the visceral pain present in an acute and chronic DSS-colitis model. This also implies that these behavioural assessment methods may be applied in other visceral pain models.

Overall, this study and other studies have demonstrated that the RGS is a useful tool to assess ongoing pain from various different types of pain, ranging from acute inflammatory pain to acute and chronic neuropathic pain, orofacial pain, muscle pain, chronic migraine, pain from an intracerebral hemorrhage and sepsis (Sotocinal *et al*., 2011; Akintola *et al*., 2017; Liao *et al*., 2014; Asgar *et al*., 2015; Harris *et al*., 2017; Saine *et al*., 2016 and Jeger *et al*., 2017). Consequently, since the RGS can be used to assess ongoing pain present in wide variety of pain models it can and has been used to search for and assess different pain biomarkers and processes and the painfulness of novel surgical devices and analgesics (Asgar *et al*., 2015; Long *et al*., 2015; Yousef *et al*., 2015; Gao *et al*., 2016; Long *et al*., 2017 and Fujita *et al*., 2018). Furthermore, this also supports the use of the RGS as an animal welfare tool as the presence of pain from different experimental models can be identified and the appropriate interventions can be provided.

A limitation of this study was that the behaviour items in the CBS, validated for use in a laparotomy model, were applied directly to the DSS colitis model. This may be the reason why the CBS was unable to identify treatment effects and the inclusion of more behaviours may allow treatment effects to be identified with this method. Furthermore, analgesia was not administered at the end of the study to assess if the RGS scores would decrease with analgesic administration.

Future work should assess if the RGS scores will decrease with analgesic or colitis treatment, thus displaying construct validity of the RGS to assess pain from this model. For example, the use of morphine, a potent opioid, has been observed to reduce RGS scores following a laparotomy procedure (Sotocinal *et al*., 2011) and reduced nociception after DSS colitis in mice (Jain *et al*., 2015). Therefore, the DSS colitis model of this study could be repeated and assess if RGS scores reduce after morphine administration. Additionally, a closer observation of the actual behaviour items that DSS-treated rats displayed may have provided the information for improving the CBS scale to be utilised in this model. Some behaviours that may be useful would be the abdominal licking and horizontal stretching, which have been observed in a mouse colitis model (Hassan *et al*., 2017). Lastly, it may be worthwhile to reassess burrowing during the chronic phase with a lower dose of DSS and allow for a longer exposure to the DSS and a prolonged time interval to assess the effect on burrowing behaviour. It would be interesting to assess if the reduction in burrowing would be sustained with a longer exposure to DSS. A suggested dose would be 2% DSS as this was the dose used by a study that observed a reduction in burrowing after mice were exposed to DSS (Jirkof *et al*., 2013).

3.1.2 The effects of training on the reliability of RGS scoring

This study (Zhang *et al*., 2019) demonstrated that multiple training sessions which included discussion with an experienced rater was beneficial in improving reliability between the trainee raters and the experienced rater. It was also demonstrated that a single training session or scoring of multiple images did not produce a meaningful improvement in reliability.

It was also observed that high inter- and intra-rater reliability was maintained four years later. This study agrees with previous studies that a single training session is insufficient for a trainee rater to score similarly to an experienced rater (Solomon *et al.,* 1997; Mich *et al*., 2010 and Lim *et al*., 2014). Furthermore, training does not guarantee proficiency and hence, it is important that proficiency should be assessed to ascertain that training was effective (Campbell *et al*., 2014).

A limitation of this study was that during the third session of scoring, raters were given the same 150 images to score. These images were also not randomised from the previous scoring session. Therefore, there was the potential for the observed improvement to be due to the trainee raters memorizing the rat faces and the associated scores. However, this is unlikely as there were over a hundred images of similar looking albino rats, additionally when images that were most likely to result in discrepancies in the scores were removed, it did not significantly change the ICC scores. Furthermore, the ability of the trainee raters to still score reliably with one other and the experienced rater four years after training demonstrates that these trainee raters scored reliably due to proficiency and not from the memorisation of the images. Another limitation of the study was the small sample size of participants: 4 trainee raters and 6 'no training' raters, this limits the generalisability of these findings.

Future work should assess the effects of training in a different population of trainee raters. This may be completed in a similar manner as Roughan and Flecknell (2006) who introduced the CBS method to students during a lecture and during an advanced training program. During this study, trainees were asked to score a 5-minute long video with a visual analog scale (*i.e.* estimation of pain severity by marking along a line). Trainees were then taught about the various behaviours in the CBS and asked to reassess the same video by noting down the behaviours observed. This study demonstrated that trainees improved in their ability to identify painful rats from rats that received analgesic treatment after training. Therefore, a follow up study that includes a larger population of trainee raters will increase the overall sample size and generalisability of the findings. Future work should also assess the inter-rater variability between different labs to discern the variability between labs as it seems that each employs a different training or learning method. This could provide insight on how comparable the data from studies performed by different research groups.

3.1.3 Assessment of adherence to the ARRIVE guidelines five years later

This study has demonstrated that five years after the publication of the ARRIVE guidelines, reporting standards have not improved meaningfully in studies of animal welfare, anesthesia and analgesia in veterinary journals. This result echoes many other studies that assessed reporting adherence to the ARRIVE guidelines in other subjects (Schwarz *et al*., 2012; Bara and Joffe, 2014; Delgado-Ruiz *et al*., 2014; Ting *et al*., 2014; Avey *et al*., 2016; Gulin *et al*., 2016; Lin *et al*., 2016; Nam *et al.*, 2018). This study also demonstrated that papers published in journals that support the ARRIVE guidelines did not perform better. It is also discouraging to discover that not one paper fully adhered to all items on the ARRIVE guidelines and that on average papers only fully report 50-60% of the items on the ARRIVE guidelines. This means that the ARRIVE guidelines have not made a significant impact and many published papers still do not provide enough information to be replicated, validated or utilised in retrospective analyses. This begs the question: why have reporting standards not improved five years after the ARRIVE guidelines were published?

A limitation of this study published on this topic in this thesis was the small number of journals in the non-supporting group (n = 2). However, the ARRIVE guidelines have been widely adopted by hundreds of journals and this made it difficult to find journals that do not support the guidelines. Additionally, the risk of a systematic bias from the inclusion of only two journals is most likely low because similarly poor reporting standards have also been observed in other studies (Schwarz *et al*., 2012; Bara and Joffe, 2014; Delgado-Ruiz *et al*., 2014; Ting *et al*., 2014; Avey *et al*., 2016; Gulin *et al*., 2016; Lin *et al*., 2016 and Nam *et al*., 2018).

While the comparison between journals that support or do not support the ARRIVE guidelines may be limited by the large number of journals that support the ARRIVE guidelines, other comparisons can still be made. Future work could include assessment of the adherence to the ARRIVE guidelines once more time has passed (*e.g.* 10 years later). Assessments of adherence should also account for the various enforcement methods that have been proposed, such as the use of a mandated checklist (Han *et al.,* 2017 and Macleod *et al*., 2017). Furthermore, when the revised ARRIVE guidelines are published (du Sert *et al*., 2018), reassessments should be performed to evaluate if the proposed tiered system is useful in the promoting improved reporting standards.

3.3 Pain research: looking forward

3.3.1 Use of appropriate assessment methods

Going forward, the fact that pain is a multidimensional phenomenon should be appreciated during pain research. This implies two changes in the way pain studies are primarily conducted: 1) a shift away from utilising only nociceptive evoked testing methods to assess the entire pain experience in animals and 2) the use of multiple types of pain assessment methods to evaluate the different dimensions of pain.

Firstly, the shift away from nociceptive evoked testing does not mean that these methods are irrelevant in pain research. Instead, pain researchers need to be more discerning of the type or quality of pain they are interested in and not use one type of assessment method as a proxy for another (Mogil and Crager, 2004). Afterall, it has been reported in both rodents and humans that the different qualities of pain are distinct and distinguishable (Gould, 2000 and De Rantere *et al*., 2015). These different aspects of pain may be evaluated with the combined application of nociceptive threshold testing, spontaneous behaviours and performance of ADL activities. Researchers may also want to consider the co-morbidities of pain, such as anxiety and depression, which have been observed in chronic pain patients and can be modelled similarly in rodent models (Mogil *et al*., 2010). The matching of the appropriate assessment method to the interested quality of pain will result in increased testing specificity and accuracy of the interested type of pain.

Secondly, the use of multiple types of assessment methods concurrently (termed the triangulation method) should be utilised to assess the entire pain experience (Bateson, 1991 and Roughan *et al*., 2014). This allows researchers to observe pain from different angles, and therefore build up a more complete picture of pain. For example, the use of multiple assessment methods demonstrated that the duration of different pain types may differ (*i.e.* mechanical hypersensitivity outlasts ongoing pain, De Rantere *et al*., 2015) and analgesic dose efficacy may differ depending on the assessment method used (Matsumiya *et al*., 2012; Waite *et al*., 2015 and Oliver *et al*., 2018). The use of multiple assessment methods may also help to identify more sensitive humane endpoints (Roughan *et al*., 2014 and Oliver *et al*., 2018). This is also applicable in the use of multiple nociceptive evoked tests to perform a QST protocol. The use of a QST protocol allows researchers to identify three distinct clusters of sensory profiles that are present across all types of patients with neuropathic diseases (Baron *et al* 2017). This has resulted in the reconsideration of neuropathic pain treatments not just due to disease types but rather the underlying pain mechanisms at play. Hence, the use of multiple assessment methods concomitantly builds up a more complete picture of the pain phenomenon and allows researchers to test hypothesis they may not have considered had they used only one type of assessment method.

Furthermore, the appropriate assessment methods for different types of pain can be inferred from human studies. Animal models have been utilised as models for human pain because humans and animals are similar physiologically (*i.e.* both possess similar nociceptors and brain structures to process pain which is abolished with analgesics) and behaviourally (*i.e.* both will react to pain and learn to avoid it; Bateson, 1991). Therefore, human studies can be used to inform animal studies on how different types of pain can manifest and what types of pain behaviours can be expected. For example, human neuropathic pain is known to include spontaneous pain (assessed with self-report), reduction of general well-being (with decreased motivation to perform normal activities) and abnormal sensory profiles (assessed with QST; Backonja *et al.,* 2013). It can be expected that animals with neuropathic pain conditions also experience the same symptoms, however, pain assessments are primarily limited to nociceptive evoked testing. Therefore, assessments of spontaneous pain and reduction of general wellbeing are relevant to characterising neuropathic pain and should also be assessed. While direct self-report by animals is impossible, animals can still ‘self-report’ their pain by the presence of spontaneous behaviours associated with pain (*i.e.* with the CPP test or grimacing; King *et al*., 2009 and Sotocinal *et al*., 2011). Assessment of general well-being of rats with neuropathic pain can be quantified with ADL activities such as burrowing (Andrews *et al*., 2012).

There is the continued need for the development, and validation (including identifying limitations) of novel pain assessment methods (Mogil, 2010). The development of new pain assessment methods includes the understanding of specific innate behaviour repertoires, such as grimacing, burrowing and grooming in rodents, and their specificity to pain and possible confounding factors. For example, grimacing in mice is known to be affected by the presence of male olfactory cues and therefore, the use of female observers or habituation to male observers may be required (Sorge *et al*., 2014). The appropriateness of the MGS and other grimace scales for use in chronic neuropathic pain models is contested (Langford *et al*., 2010 and Akintola *et al*., 2017). Follow-up studies are required to ascertain if the MGS is applicable for all neuropathic pain models or if it is only applicable in a few. Moreover, validation of new pain assessment methods is important, however, it should be noted that the validation of any assessment method is specific to the situation in which it was used (*e.g*. the pain assessment method may vary depending on the populations, environment and pain models it was tested with; Norman and Streiner, 2008). Validation usually includes assessment of face validity (*i.e*. do the items on the scale make sense?) and criterion validity (*i.e.* does the scale increase with pain and decrease with analgesia?) as well as reliability (*i.e.* is the scale consistent between and within raters?). Lastly, proper training and assessment of proficiency is important for the reliability of a scale (Zhang *et al*., 2019). Raters with higher proficiency are required to produce reliable and consistent data and to reduce inherent variability.

Overall, there is a need to ensure that the pain assessment methods used in animal models are relevant to the hypothesis tested and to the pain disease it is supposed to mimic. This includes the use of relevant and different types of assessment methods to characterise pain in animal models. Additionally, the continued development of new assessment methods that considers the animals' natural behavioural repertoire and the training of raters to use them are needed.

3.3.2. Use of technology to automate

The use of machine learning to automate pain assessment has been utilised in methods such as the HomeCageScan, a software that detects various behaviours of mice and records changes in their frequency and duration (Roughan *et al.,* 2009). The application of similar software to grimacing has been developed for image capture and scoring (Sotocinal *et al*., 2011 and Tuttle *et al.*, 2018). The use of such software reduces the work required to collect and score images for researchers. The combination of both software could also be applied as a clinical tool for round the clock monitoring. However, there are limitations that need to be understood.

Facial recognition software needs to be able to recognise a rat's face and some obstacles include: 1) head orientation, 2) lighting effects, 3) ability to detect facial landmarks (fur colour), 4) view obstruction (*i.e.* head movements or view blocked by objects) and 5) correct identification of expressions from facial landmarks (Sariyanidi *et al.*, 2015). The current version of the RFF software identifies the faces by using eyes and ears as identifying landmarks (Sotocinal *et al.*, 2011). It excludes unclear images by selecting images that are similar to the previous video frame. While these features make the software functional, it is still limited as a research tool in its current state. Rats should not be scored when they are performing certain actions as it will distort their facial expressions (*i.e.* during grooming, eating, sniffing and sleeping). During manual image extraction or real-time scoring, observers can make the distinction and decide when it is appropriate to score or extract an image. The RFF does not currently account for these behavioural influences. The RFF was developed while the rats were in a clear Plexiglas box under optimum lighting conditions. This is not the case in a home cage setting; light quality will differ depending on the rack level the cage is placed at and the rats would be blocked by enrichment items or cage mates. The use of infrared may be useful to deal with poor lighting or assessments during the dark cycle (Li and Deng, 2018) while the use of multiple cameras may be able to capture images around the obstructions. The RFF does not differentiate between individuals, preventing its use in group housing. The tracking of individual mice with the HomeCageScan has been demonstrated as possible and similar applications for the RFF may be possible as well (Bains *et al.*, 2016).

Subsequently, the scoring of multiple images requires a software that can recognise the different facial expressions (Li and Deng, 2018). This has been performed with a deep neural network to create the *a*MGS which binarily classifies images as "pain" or "no pain" (Tuttle *et al.*, 2018). These types of software typically require large training data sets and the *a*MGS was trained with nearly 6000 images which needed to be annotated manually (*i.e.* labelling of identifying facial landmarks; Li and Deng, 2018 and Tuttle *et al.*, 2018). While the *a*MGS was able to identify images into the two categories with similar accuracy to human raters (83-93%), the *a*MGS functioned best when images had scores at the extreme ends of the scale (*e.g.* 0 or 2) and poorly when scores were intermediate (*e.g.* 1). This shows that the current version of the *a*MGS lacks sensitivity and needs to be improved to capture subtle facial changes. This may be possible with 3D facial expression recognition (*i.e.* modeling the rat's face from 2D to 3D, thereby reducing the effect of head orientation and lighting and improving the recognition of subtler facial expression changes) or utilising visualization techniques to focus on landmarks that may improve discrimination (*i.e.* the identification of a classifier – orbital tightening and ear changes may allow better discrimination between "pain" and "no pain" mice, Dalla Costa *et al*., 2018 and Li and Deng, 2018). Lastly, the *a*MGS needs to be validated with image sets collected from other laboratories as the image sets used to train, test and validate the software were limited to two laboratories. It has been reported that the effectiveness of a program may be affected by different collection environments, construction indicators and annotators (Li and Deng, 2018).

Overall, the introduction of automation into pain assessments of animals will relieve the workload required to collect and score the images. However, current software is limited in their utility and can only serve as research tools. Improvements or development of new software to accurately identify animal faces in their home cage and to score subtle changes in facial expressions are required for clinical applications. The use of machine learning programs can be trained to overcome the many obstacles during the identification and scoring of rat faces. However, this requires more time for people to manually annotate additional images or videos. The eventual development of such programs will improve the sensitivity of pain assessments with round the clock monitoring. This will also improve animal welfare as painful animals can be identified immediately. However, there will always be the need for human observers who are competent because they are required to decide if an animal has reached a humane endpoint and intervention is required. Furthermore, while machine learning can only identify specifically what they have been programmed to identify, experienced animal care personnel are able to make judgements based on observing the entire animal. Lastly, though time-consuming, human observers can simply observe the animals in real-time without need for complex or expensive equipment which will be required for automated scoring.

3.4. How can we ARRIVE?

Despite the overwhelming support from journals for the ARRIVE guidelines, the general observation of low adherence to the guidelines five years after its publication suggests that more time may be required for the guidelines to take effect or that current reliance on voluntary adherence to reporting standards will never be successful. The CONSORT (CONsolidated Standards Of Reporting Trials) statement, the oldest and most popular clinical reporting guideline to date can be used to explore the potential effect of time on the ARRIVE guidelines.

3.3.2. A "case-study" of the CONSORT statement

The CONSORT statement was first developed in 1996 (Begg *et al.*) to promote better reporting standards of randomized controlled trials (RCTS). It has since been revised twice, in 2001 (Moher *et al.*, 2001) and in 2010 (Schluz *et al.*, 2010). The current version of the CONSORT checklist includes 25 items and a flowchart diagram. Similar to the ARRIVE guidelines, there are many journals that support the CONSORT statement, with 585 journals (around 50% of all medical journals) that currently endorse the guidelines (www.consort-statement.org). "Extensions" of the CONSORT statement have also been developed to accommodate different RCT studies (*e.g.* cluster RCTs, Campbell *et al.*, 2012)

Several studies assessing the impact of the CONSORT statement have been performed since its conception (Moher *et al.*, 2001; Plint *et al.*, 2006 and Turner *et al.*, 2012). These have shown that the introduction of the CONSORT statement has resulted in an improvement in reporting or adherence to the guidelines 14 years later. Additionally, journals that supported the guidelines were more likely to publish papers with higher reporting standards. Papers from CONSORT supporting journals do seem to report certain items better, for example, items 'allocation concealment' and 'sample size' were 81% and 61%, respectively, and were more likely to be reported in papers from journals that support the guidelines (Turner *et al.*, 2012). However, when comparisons between supporting and non-supporting journals were performed for the entire guidelines, the differences between the journal types was only 3%. When overall adherence to the CONSORT statement was assessed, a median of only 60-73% of the CONSORT items were reported (Setvanovic *et al.*, 2010; Munter *et al.*, 2015; Sarkis-Onofre *et al.*, 2010 and Tam *et al.*, 2017). Overall, 23 years after the introduction of the first CONSORT statement, journals that support the guidelines may publish more papers that report certain items, however, overall adherence to the guidelines are still suboptimal.

Poor reporting standards regarding the ARRIVE guidelines and CONSORT statement is not isolated. It was reported that out of 124 systematic reviews that assessed adherence to various guidelines, 87.9% of these reviews report suboptimal reporting (Jin *et al.*, 2018). These included the ARRIVE (4/4 [100%]), CONSORT (71/81 [88%]) as well as two other guidelines, PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses; 16/19 [84%]) and STROBE (Strengthening the reporting of observational studies in epidemiology; 7/8 [88%]).

Therefore, poor reporting is rampant across the scientific literature and while support for the guidelines may improve the reporting of certain items, additional measures are required for improvement.

3.3.3. Improving adherence to the ARRIVE guidelines

One criticism of the ARRIVE guidelines is its general nature. However, the guidelines are meant to be general, allowing application in a wide variety of research areas and used as the basis for more specific guidelines (du Sert *et al.*, 2018). Adaptation of the ARRIVE guidelines has already been applied at least once in the development of a guideline for experimental autoimmune encephalomyelitis, experts in the field built upon the ARRIVE guidelines and added additional recommendations important to the disease model (Amor and Baker, 2012). This has also been performed with the CONSORT statement where extensions of the guidelines are available when the CONSORT guidelines may fit well with other types of studies. It has also been criticized that the ARRIVE guidelines should be more specific for certain items to improve the replicability of the reported studies. This includes the increased emphasis on the complete reporting of analgesics and anesthetics used (Muhlhausler *et al.*, 2013 and Carbone and Austin, 2016) as well as the specificity of husbandry procedures (Hoojimans *et al.*, 2011).

A second criticism of the ARRIVE guidelines is its length, with the suggestion that full incorporation is too detailed and difficult to read (Carbone and Austin, 2016). When it is broken down into constituent parts, the ARRIVE guidelines can consist of multiple items and subitems, which limits the capacity of reviewers and editorial staff to easily evaluate adherence (Avey *et* *al*., 2010). In addition to the creation of shorter and more specific guidelines, as described above, it has been suggested that reporting guidelines should focus on core items that directly affect the validity of the study (Landis *et al*., 2012; Hooijmans and Ritskes-Hoitinga, 2013 and Reichlin *et al*., 2016). One such proposal was the emphasis on items that relate to bias and study design: randomization, blinding, sample size estimation and data handling (Landis *et al*., 2012). The focus on a few core items will allow for adherence assessments to be performed more easily and rapidly. A more general and shorter guideline which includes items common to many guidelines has also been proposed: harmonized animal research reporting principles (HARRP; Osborne *et al*., 2018). This guideline consists of eight items which were formed by the combination of some ARRIVE items (*e.g*. sample size and animal allocation are placed under the 'study design' item) and the exclusion of some items (*i.e.* title, abstract and generalisability/translation). This guideline may be easier to follow with a slightly shorter list of items, use of simpler language and a less rigid format. The concept of prioritizing and focusing on some items has been planned for the revision of the ARRIVE guidelines where items will be ranked into different levels of priority (du Sert *et al*., 2018). However, while this may ease the incorporation of the ARRIVE guidelines into the review process, it should still be emphasized that non- 'core' items remain important.

Lastly, it has been noted that authors do not fully comprehend the issues of poor reporting and thus do not appreciate the use of the ARRIVE guidelines to promote scientific rigor. This is evident when 56% and 90% of surveyed Swiss and Chinese researchers, respectively, report unawareness of the ARRIVE guideline (Reichlin *et al*., 2013 and Ma *et al*., 2017). Furthermore, most of the Swiss researchers surveyed could not correctly identify the correct methods to reduce different types of bias, suggesting that the studies they published are prone to biases (Reichlin *et al*., 2013). Therefore, the development of an explanation and elaboration document, as has been published to accompany the CONSORT statement should be considered (Moher *et al*., 2010). This has been proposed for the next revision of the ARRIVE guidelines (du Sert *et al*., 2018).

The observed continued suboptimal reporting standards of published RCTs even though the CONSORT statement has been published for 23 years and has been revised twice, it is evident that only revising the ARRIVE guidelines is insufficient. Therefore, to ensure improvement of guideline adherence additional or alternative solutions must be considered.

As noted above, the ARRIVE guidelines and its usefulness to ensure scientific rigor is still unknown to many researchers. Therefore, increasing awareness of the ARRIVE guidelines is paramount. The ARRIVE guidelines can be introduced to new and veteran researchers by its introduction into the university curriculum or promotion at scientific conferences. Multiple resources on the NC3Rs website (www.nc3rs.org) are available, including presentations with accompanying speaker notes and a webinar. Currently, many journals show support for the guidelines by mentioning the ARRIVE guidelines in their 'instructions to authors' page and ask authors to refer to it when writing up their manuscripts. The need to look at a second page of instructions may deter authors from doing so. Therefore, incorporation of the ARRIVE guidelines directly into the 'instructions to authors' page to present one coherent set of instructions may improve reporting adherence. This will also allow journals to customize the ARRIVE guidelines to emphasize certain items that are important in the field they represent.

In addition to an increased promotion of the ARRIVE guidelines, changes in the review process should be considered (McGrath and Lilley, 2015). The focus of this would be to minimise the time and increase the ease for editors and reviewers to assess adherence. One proposed method has been the submission of a mandated checklist alongside the manuscript (Loder *et al*., 2009). The use of such a checklist would ensure that the manuscripts submitted adheres to the ARRIVE guidelines. Authors are likely to be very motivated to complete the checklist if they know the review process will not proceed without it. This has been enforced by the Nature Publishing Group with reported success in improved reporting of items related to bias (Anon, 2013; Macleod *et al*., 2017 and Han *et al*., 2017). Another suggestion was to provide a template of the methods section, where the bulk of the items from the ARRIVE guidelines is located that can be included in the publication itself (Baker *et al*., 2014 and McGrath and Lilley, 2015). This would standardise the methods section thus simplifying the evaluation process and improving transparency. The use of technology such as text mining or machine learning can also improve evaluation with automation (Florez-Vargas *et al*., 2016 and Bahor *et al*., 2017). Lastly, a team of specialists can be employed to review specific aspects of the manuscript to ensure sufficient information is provided and that the statistical methods were correctly used (*e.g.* a team of statisticians; Baker *et al*., 2014).

Overall, the support of the ARRIVE guidelines by researchers, journals, fund granting agencies and other stakeholders demonstrate that current reporting standards need to improve. However, the publication of the ARRIVE guidelines and its support has not caused a meaningful improvement. There have been many suggestions of how reporting standards of scientific literature can improve, from the emphasis of a few key items to changes in the peer-review process. It remains to be seen if such measures can be implemented practically and if they will be effective in improving the reporting standards.

3.5. Additional factors to improve pain research

In addition to the improvements of pain assessment methods and better reporting standards, there are many other factors that require consideration. One such factor is the re-evaluation of the animal models used (Mogil *et al*., 2010). Animal pain models have historically been young, male and of a single strain whereas most chronic pain patients are middle-aged women of varying ethnicities (Greenspan *et al*., 2007). This discrepancy demonstrates that the effects of age, sex and genetic heterogeneity are overlooked during the use of animals as translational pain models. Additionally, it was suggested that naturally occurring models should be adopted as they will better reflect the human pain condition in terms of age and genetic diversity (Mogil *et al*., 2010 and Klinck *et al*., 2017).

Additionally, there is the need for the re-evaluation of what is defined as success in the research world (Rice *et al*., 2013). Currently, there is an overemphasis on the quantity of publications, citations, grant funding and the perceived impact of journals in which studies are published. In contrast, recognition for work that directly replicates significant pre-clinical findings are non-existent. There is a need to include longer term metrics where significant pre-clinical findings are translated to RCTs (*i.e.* results from pre-clinical research are replicated in RCTs; Rice *et al*., 2013). It should also include the methodical evaluation of studies and their raw data with the ARRIVE guidelines to evaluate potential study design flaws and deciding if the results are truly significant. Additionally, there is the need for the reduction of publication bias where studies with positive results are overwhelmingly more likely to be published (Rice *et al.,* 2013).

Lastly, the use and promotion of study design guidelines should be utilised to complement reporting guidelines. The use of a study design guideline will ensure that researchers have considered all required steps and will perform a well-designed study. When it is time to write up their study, they should have all information required to fulfill any reporting guidelines and turn in a well reported manuscript. One such guideline has already been proposed: Planning Research and Experimental Procedures on Animals: Recommendations for Excellence (PREPARE) guidelines (Smith *et al*., 2018). This guideline was proposed to direct researchers to the various factors that needs to be considered as well as steps that should be undertaken before a study is conducted (*e.g.* performance of a literature review). The use of both study design and reporting guidelines should improve the quality and reporting of future research.

3.6. Conclusion

Pain assessments in animals has traditionally depended on nociceptive evoked testing methods that do not assess the affective ongoing component of pain. This has been suggested as one of the reasons why novel analgesics that demonstrated efficacy in animal trials do not translate during human trials. It has been proposed that non-evoked spontaneous behaviours can be utilised to assess the affective ongoing component of pain. One such non-evoked spontaneous behavioural tool is the RGS.

This thesis has demonstrated that the RGS is a viable and practical tool that can be applied in a research and clinical setting. Real-time application of the RGS allows pain assessments to be performed easily and quickly, which in turn allows data to be generated quickly and analgesic intervention can be provided. Furthermore, the ability to use the RGS to assess acute and chronic visceral pain highlights the RGS as a useful pain assessment method which can be applied in a wider array of pain types than previously reported. Consequently, this means that the RGS may be a robust tool for a variety of different pain models to study the different pain mechanisms, identify pain biomarkers and assess the efficacy of novel analgesics. More importantly, the real-time application of the RGS for pain assessment will greatly improve the welfare of rats in experimental studies. Real-time application of the RGS allow researchers to rapidly identify the presence and intensity of pain and decide when analgesic intervention is required. The RGS is also a potential method for the pain assessment of all laboratory rats to ensure good animal welfare. Therefore, researchers no longer rely solely on nociceptive evoked testing methods as the only way to assess pain in animals. The use of such behavioural tools will improve the specificity of pain type assessed which will also improve translation to human pain as similar qualities of pain are assessed. This will also improve the welfare of animals involved in research as the affective and perhaps more relevant component of pain can now be assessed.

It has also been demonstrated that training and assessment of proficiency is required before using the RGS to ensure scoring reliability. It was demonstrated that there is a learning curve involved in the use of the RGS and training is required for new raters to improve.

Furthermore, similar proficiency or reliability with an experienced rater is possible after three training sessions and can be maintained for a few years. This demonstrates that training is important in the use of the RGS, a finding that should be considered when developing and adopting other pain scales.

Finally, this thesis has also demonstrated that five years after the publication of the ARRIVE guidelines, reporting standards have not improved meaningfully. This highlights that current measures to improve reporting standards are insufficient and changes to the guidelines or enforcement may be required for a meaningful improvement. Good reporting standards are imperative for scientific progress because sufficient information is required to validate and replicate study findings and for a study to be included in retrospective analysis that will maximise the information gathered. Within pain research, there are many factors that may confound outcome measures and they need to be accounted for. Reporting standards also have welfare and financial implications, this includes the animals and money spent on research that cannot be replicated to the use of additional animal and money resources to replicate findings.

Overall, future pain research should shift away from nociceptive evoked testing as the sole pain assessment method in pre-clinical animal research. Instead, researchers should embrace the use of behavioural tools, such as the RGS, or ADL activities, such as burrowing, to effectively and fully characterise the pain experience. The appropriate pain assessment methods can be inferred from the methods used in human pain studies due to similarities between animals and humans physiologically and behaviourally. There is also a need for the continued development of new pain assessment methods that consider the natural behavioural repertoire of animals. Pain assessment methods need to be validated for use in different situations and populations to identify potential limitations and ensure the assessment tool is appropriate for the situation. Additionally, the future of pain assessment will include automated pain assessments by machines which will reduce the work required by researchers and animal care personnel and allow for round the clock monitoring. However, there are limitations of the currently developed software that restrict their use as research tools only. Lastly, additional measures are required for the ARRIVE guidelines to meaningfully improve reporting standards of pre-clinical animal research. It remains to be seen if these measures will be effective and continuous assessments for the adherence to the ARRIVE guidelines is required.

It is the hope that the work in this thesis will encourage more researchers and animal care personnel to utilise the RGS, or other non-evoked spontaneous behaviour tools, as a pain assessment tool to research pain and to ensure the well-being of animals. It is also the hope that this thesis encourages a conscientious effort to improve the reporting standards of published papers.

4. Bibliography

Abbott, F. V., Franklin, K. B. J., & Connell, B. (1986). The stress of a novel environment reduces formalin pain - possible role of serotonin. *European Journal of Pharmacology, 126*(1-2), 141-144. doi:10.1016/0014-2999(86)90750-8

Adam, B., Liebregts, T., Gschossmann, J. M., Krippner, C., Scholl, F., Ruwe, M., & Holtmann, G. (2006). Severity of mucosal inflammation as a predictor for alterations of visceral sensory function in a rat model. *Pain, 123*(1-2), 179-186. doi:10.1016/j.pain.2006.02.029

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. doi:10.1016/j.ynpai.2017.10.001

Al–Chaer, E. D., Kawasaki, M., & Pasricha, P. J. (2000). A new model of chronic visceral hypersensitivity in adult rats induced by colon irritation during postnatal development. *Gastroenterology, 119*(5), 1276-1285. doi:10.1053/gast.2000.19576

Alex, P., Zachos, N. C., Nguyen, T., Gonzales, L., Chen, T. E., Conklin, L. S., . . . Li, X. H. (2009). Distinct Cytokine Patterns Identified from Multiplex Profiles of Murine DSS and TNBS-induced Colitis. *Inflammatory Bowel Diseases, 15*(3), 341-352. doi:10.1002/ibd.20753

Alleva, E., Caprioli, A., & Laviola, G. (1986). Postnatal social-environment affects morphine analgesia in male-mice. *Physiology & Behavior, 36*(4), 779-781. doi:10.1016/0031-9384(86)90368-9

Als-Nielsen, B., Chen, W. D., Gluud, C., & Kjaergard, L. L. (2003). Association of funding and conclusions in randomized drug trials - A reflection of treatment effect or adverse events? *Jama-Journal of the American Medical Association, 290*(7), 921-928. doi:10.1001/jama.290.7.921

Altholtz, L. Y., Fowler, K. A., Badura, L. L., & Kovacs, M. S. (2006). Comparison of the stress response in rats to repeated isoflurane or CO2 : O-2 anesthesia used for restraint during serial blood collection via the jugular vein. *Journal of the American Association for Laboratory Animal Science, 45*(3), 17-22.

Altman, D. G., Schulz, K. F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., . . . Grp, C. (2001). The revised CONSORT statement for reporting randomized trials: Explanation and elaboration. *Ann Intern Med, 134*(8), 663-694. doi:10.7326/0003-4819-134-8-200104170-00012

Amour, F. E., & Smith, D. L. (1941). A method for determining loss of pain sensation. *Journal of Pharmacology and Experimental Therapeutics, 72*(1), 74.

Andrews, N. A., Latremoliere, A., Basbaum, A. I., Mogil, J. S., Porreca, F., Rice, A. S., . . . Whiteside, G. (2016). Ensuring transparency and minimization of methodologic bias in preclinical pain research: PPRECISE considerations. *Pain, 157*(4), 901-909. doi:10.1097/j.pain.0000000000000458

Andrews, N., Legg, E., Lisak, D., Issop, Y., Richardson, D., Harper, S., . . . Rice, A. S. (2012). Spontaneous burrowing behaviour in the rat is reduced by peripheral nerve injury or inflammation associated pain. *Eur J Pain, 16*(4), 485-495. doi:10.1016/j.ejpain.2011.07.012

Anon. (2013). Reducing our irreproducibility. *Nature, 496*(7446), 398-398.

Arendt-Nielsen, L., & Yarnitsky, D. (2009). Experimental and clinical applications of quantitative sensory testing applied to skin, muscles and viscera. *J Pain, 10*(6), 556-572. doi:10.1016/j.jpain.2009.02.002

Arras, M., Rettich, A., Cinelli, P., Kasermann, H. P., & Burki, K. (2007). Assessment of post-laparotomy pain in laboratory mice by telemetric recording of heart rate and heart rate variability. *BMC Vet Res, 3*, 16. doi:10.1186/1746-6148-3-16

Asgar, J., Zhang, Y., Saloman, J. L., Wang, S., Chung, M. K., & Ro, J. Y. (2015). The role of TRPA1 in muscle pain and mechanical hypersensitivity under inflammatory conditions in rats. *Neuroscience, 310*, 206-215. doi:10.1016/j.neuroscience.2015.09.042

Attal, N., Bouhassira, D., Baron, R., Dostrovsky, J., Dworkin, R. H., Finnerup, N., . . . Treede, R. D. (2011). Assessing symptom profiles in neuropathic pain clinical trials: can it improve outcome? *Eur J Pain, 15*(5), 441-443. doi:10.1016/j.ejpain.2011.03.005

Avey, M. T., Moher, D., Sullivan, K. J., Fergusson, D., Griffin, G., Grimshaw, J. M., . . . McIntyre, L. (2016). The Devil Is in the Details: Incomplete Reporting in Preclinical Animal Research. *PLoS One, 11*(11). doi:10.1371/journal.pone.0166733

Backonja, M. M., & Stacey, B. (2004). Neuropathic pain symptoms relative to overall pain rating. *J Pain, 5*(9), 491-497. doi:10.1016/j.jpain.2004.09.001

Backonja, M., Attal, N., Baron, R., Bouhassira, D., Drangholt, M., Dyck, P. J., . . . Ziegler, D. (2013). Value of quantitative sensory testing in neurological and pain disorders: NeuPSIG consensus. *Pain, 154*(9), 1807-1819. doi:10.1016/j.pain.2013.05.047

Bahor, Z., Liao, J., Macleod, M. R., Bannach-Brown, A., McCann, S. K., Wever, K. E., . . . Sena, E. (2017). Risk of bias reporting in the recent animal focal cerebral ischaemia literature. *Clin Sci (Lond), 131*(20), 2525-2532. doi:10.1042/CS20160722

Bains, R. S., Cater, H. L., Sillito, R. R., Chartsias, A., Sneddon, D., Concas, D., . . . Armstrong, J. D. (2016). Analysis of Individual Mouse Activity in Group Housed Animals of Different Inbred Strains using a Novel Automated Home Cage Analysis System. *Front Behav Neurosci, 10*, 106. doi:10.3389/fnbeh.2016.00106

Baker, D., & Amor, S. (2012). Publication guidelines for refereeing and reporting on animal use in experimental autoimmune encephalomyelitis. *Journal of Neuroimmunology, 242*(1-2), 78-83. doi:10.1016/j.jneuroim.2011.11.003

Baker, D., Lidster, K., Sottomayor, A., & Amor, S. (2014). Two Years Later: Journals Are Not Yet Enforcing the ARRIVE Guidelines on Reporting Standards for Pre-Clinical Animal Studies. *Plos Biology, 12*(1). doi:10.1371/journal.pbio.1001756

Ballantyne, M., Stevens, B., McAllister, M., Dionne, K., & Jack, A. (1999). Validation of the premature infant pain profile in the clinical setting. *Clinical Journal of Pain, 15*(4), 297-303. doi:10.1097/00002508-199912000-00006

Banik, R. K., & Kabadi, R. A. (2013). A modified Hargreaves' method for assessing threshold temperatures for heat nociception. *J Neurosci Methods, 219*(1), 41-51. doi:10.1016/j.jneumeth.2013.06.005

Bara, M., & Joffe, A. R. (2014). The methodological quality of animal research in critical care: the public face of science. *Annals of Intensive Care, 4*, 9. doi:10.1186/s13613-014-0026-8

Baron, R., Maier, C., Attal, N., Binder, A., Bouhassira, D., Cruccu, G., . . . Treede, R. D. (2017). Peripheral neuropathic pain: a mechanism-related organizing principle based on sensory profiles. *Pain, 158*(2), 261-272. doi:10.1097/j.pain.0000000000000753

Barrot, M. (2012). Tests and models of nociception and pain in rodents. *Neuroscience, 211*, 39-50. doi:10.1016/j.neuroscience.2011.12.041

Bateson, P. (1991). Assessment of pain in animals. *Animal Behaviour, 42*, 827-839. doi:10.1016/s0003-3472(05)80127-7

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olki, I., . . . Amer, J. (1997). Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *Medizinische Klinik, 92*(11), 675-679.

Ben-Ami Bartal, I., Decety, J., & Mason, P. (2011). Empathy and pro-social behavior in rats. *Science, 334*(6061), 1427-1430. doi:10.1126/science.1210789

Bennett, G. J. (2012). What Is Spontaneous Pain and Who Has It? *Journal of Pain, 13*(10), 921-929. doi:10.1016/j.jpain.2012.05.008

Berge, O. G., Garciacabrera, I., & Hole, K. (1988). Response latencies in the tail-flick test depend on tail skin temperature. *Neuroscience Letters, 86*(3), 284-288. doi:10.1016/0304-3940(88)90497-1

Bianchi, M., & Panerai, A. E. (2002). Effects of lornoxicam, piroxicam, and meloxicam in a model of thermal hindpaw hyperalgesia induced by formalin injection in rat tail. *Pharmacological Research, 45*(2), 101-105. doi:10.1006/phrs.2001.0921

Bielefeldt, K., Davis, B., & Binion, D. G. (2009). Pain and Inflammatory Bowel Disease. *Inflammatory Bowel Diseases, 15*(5), 778-788. doi:10.1002/ibd.20848

Bjorn, A., Pudas-Tahka, S. M., Salantera, S., & Axelin, A. (2017). Video education for critical care nurses to assess pain with a behavioural pain assessment tool: A descriptive comparative study. *Intensive and Critical Care Nursing, 42*, 68-74. doi:10.1016/j.iccn.2017.02.010

Blanchard, R. J., Blanchard, D. C., Agullana, R., & Weiss, S. M. (1991). 22 khz alarm cries to presentation of a predator, by laboratory rats living in visible burrow systems. *Physiology & Behavior, 50*(5), 967-972. doi:10.1016/0031-9384(91)90423-l

Bland, J. M., & Altman, D. G. (2007). Agreement between methods of measurement with multiple observations per individual. *Journal of Biopharmaceutical Statistics, 17*(4), 571-582. doi:10.1080/10543400701329422

Boissy, A., Aubert, A., Desire, L., Greiveldinger, L., Delval, E., & Veissier, I. (2011). Cognitive sciences to relate ear postures to emotions in sheep. *Animal Welfare, 20*(1), 47-56.

Bolles, R. C. (1960). Grooming behavior in the rat. *Journal of Comparative and Physiological Psychology, 53*(3), 306-310. doi:10.1037/h0045421

Bove, G. (2006). Mechanical sensory threshold testing using nylon monofilaments: the pain field's "tin standard". *Pain, 124*(1-2), 13-17. doi:10.1016/j.pain.2006.06.020

Bramhall, M., Florez-Vargas, O., Stevens, R., Brass, A., & Cruickshank, S. (2015). Quality of methods reporting in animal models of colitis. *Inflamm Bowel Dis, 21*(6), 1248-1259. doi:10.1097/MIB.0000000000000369

Brenna, O., Furnes, M. W., Drozdov, I., van Beelen Granlund, A., Flatberg, A., Sandvik, A. K., . . . Gustafsson, B. I. (2013). Relevance of TNBS-colitis in rats: a methodological study with endoscopic, histologic and Transcriptomic [corrected] characterization and correlation to IBD. *PLoS One, 8*(1), e54543. doi:10.1371/journal.pone.0054543

Brondani, J. T., Mama, K. R., Luna, S. P. L., Wright, B. D., Niyom, S., Ambrosio, J., . . . Padovani, C. R. (2013). Validation of the English version of the UNESP-Botucatu multidimensional composite pain scale for assessing postoperative pain in cats. *Bmc Veterinary Research, 9*, 15. doi:10.1186/1746-6148-9-143

Bryden, L. A., Nicholson, J. R., Doods, H., & Pekcec, A. (2015). Deficits in spontaneous burrowing behavior in the rat bilateral monosodium iodoacetate model of osteoarthritis: an objective measure of pain-related behavior and analgesic efficacy. *Osteoarthritis Cartilage, 23*(9), 1605-1612. doi:10.1016/j.joca.2015.05.001

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365-376. doi:10.1038/nrn3475

Calvino, B., Besson, J. M., Boehrer, A., & Depaulis, A. (1996). Ultrasonic vocalization (22-28 kHz) in a model of chronic pain, the arthritic rat: Effects of analgesic drugs. *Neuroreport: An International Journal for the Rapid Communication of Research in Neuroscience, 7*(2), 581-584. doi:10.1097/00001756-199601310-00049

Calvo, G., Holden, E., Reid, J., Scott, E. M., Firth, A., Bell, A., . . . Nolan, A. M. (2014). Development of a behaviour-based measurement tool with defined intervention level for assessing acute pain in cats. *Journal of Small Animal Practice, 55*(12), 622-629. doi:10.1111/jsap.12280

Campbell, M. K., Piaggio, G., Elbourne, D. R., Altman, D. G., & Group, C. (2012). Consort 2010 statement: extension to cluster randomised trials. *BMJ, 345*, e5661. doi:10.1136/bmj.e5661

Campbell, R. D., Hecker, K. G., Biau, D. J., & Pang, D. S. J. (2014). Student Attainment of Proficiency in a Clinical Skill: The Assessment of Individual Learning Curves. *PLoS*

One, 9(2), 5. doi:10.1371/journal.pone.0088526

Carbone, L. (2011). Pain in Laboratory Animals: The Ethical and Regulatory Imperatives. *PLoS One, 6*(9). doi:10.1371/journal.pone.0021578

Carbone, L., & Austin, J. (2016). Pain and Laboratory Animals: Publication Practices for Better Data Reproducibility and Better Animal Welfare. *PLoS One, 11*(5), 24. doi:10.1371/journal.pone.0155001

Chang, L., Munakata, J., Mayer, E. A., Schmulson, M. J., Johnson, T. D., Bernstein, C. N., . . . Matin, K. (2000). Perceptual responses in patients with inflammatory and functional bowel disease. *Gut, 47*(4), 497-505. doi:10.1136/gut.47.4.497

Chaves, R. H. F., Souza, C. C., Furlaneto, I. P., Teixeira, R. K. C., Oliveira, C. P., Rodrigues, E. M., . . . Lima, A. R. (2018). Influence of tramadol on functional recovery of acute spinal cord injury in rats. *Acta Cir Bras, 33*(12), 1087-1094. doi:10.1590/s0102-865020180120000006

Chesler, E. J., Wilson, S. G., Lariviere, W. R., Rodriguez-Zas, S. L., & Mogil, J. S. (2002). Identification and ranking of genetic and laboratory environment factors influencing a behavioral trait, thermal nociception, via computational analysis of a large data archive. *Neuroscience and Biobehavioral Reviews, 26*(8), 907-923. doi:10.1016/s0149-7634(02)00103-3

Chi, H., Kawano, T., Tamura, T., Iwata, H., Takahashi, Y., Eguchi, S., . . . Yokoyama, M. (2013). Postoperative pain impairs subsequent performance on a spatial memory task via effects on N-methyl-D-aspartate receptor in aged rats. *Life Sci, 93*(25-26), 986-993. doi:10.1016/j.lfs.2013.10.028

Chizh, B. A., Greenspan, J. D., Casey, K. L., Nemenov, M. I., & Treede, R. D. (2008). Identifying biological markers of activity in human nociceptive pathways to facilitate analgesic drug development. *Pain, 140*(2), 249-253. doi:10.1016/j.pain.2008.09.024

Christensen, S. L. T., Petersen, S., Sorensen, D. B., Olesena, J., & Jansen-Olesen, I. (2016). Infusion of low dose glyceryl trinitrate has no consistent effect on burrowing behavior, running wheel activity and light sensitivity in female rats. *Journal of Pharmacological and Toxicological Methods, 80*, 43-50. doi:10.1016/j.vascn.2016.04.004

Ciuffreda, M. C., Tolva, V., Casana, R., Gnecchi, M., Vanoli, E., Spazzolini, C., . . . Calvillo, L. (2014). Rat Experimental Model of Myocardial Ischemia/Reperfusion Injury: An Ethical Approach to Set up the Analgesic Management of Acute Post-Surgical Pain. *PLoS One, 9*(4). doi:10.1371/journal.pone.0095913

Clark, J. A., Myers, P. H., Goelz, M. F., Thigpen, J. E., & Forsythe, D. B. (1997). Pica behavior associated with buprenorphine administration in the rat. *Laboratory Animal Science, 47*(3), 300-303.

Clavelou, P., Pajot, J., Dallel, R., & Raboisson, P. (1989). Application of the formalin test to the study of orofacial pain in the rat. *Neuroscience Letters, 103*(3), 349-353. doi:https://doi.org/10.1016/0304-3940(89)90125-0

Cloutier, S., Panksepp, J., & Newberry, R. C. (2012). Playful handling by caretakers reduces fear of humans in the laboratory rat. *Applied Animal Behaviour Science, 140*(3-4), 161-171. doi:10.1016/j.applanim.2012.06.001

Cobo, E., Cortes, J., Ribera, J. M., Cardellach, F., Selva-O'Callaghan, A., Kostov, B., . . . Vilardell, M. (2011). Effect of using reporting guidelines during peer review on quality of final manuscripts submitted to a biomedical journal: masked randomised trial. *BMJ, 343*, d6783. doi:10.1136/bmj.d6783

Cohen, J. A., & Price, E. O. (1979). Grooming in the norway rat - displacement activity or boundary-shift. *Behavioral and Neural Biology, 26*(2), 177-188. doi:10.1016/s0163-1047(79)92563-9

Cohn, J.F., Ambadar, Z. and Ekman, P. (2007). Observer-based measurement of facial exression with the Faction Action Coding System. In Coan, JA & Allen, JJB (Eds.), Handbook of emotion elicitation and assessment (pp.203-221). New York, NY: Oxford University Press.

Cooper, H. S., Murthy, S. N. S., Shah, R. S., & Sedergran, D. J. (1993). Clinicopathological study of dextran sulfate sodium experimental murine colitis. *Laboratory Investigation, 69*(2), 238-249.

Coudereau, J. P., Monier, C., Bourre, J. M., & Frances, H. (1997). Effect of isolation on pain threshold and on different effects of morphine. *Progress in Neuro-Psychopharmacology & Biological Psychiatry, 21*(6), 997-1018. doi:10.1016/s0278-5846(97)00094-8

Coutaux, A., Adam, F., Willer, J. C., & Le Bars, D. (2005). Hyperalgesia and allodynia: peripheral mechanisms. *Joint Bone Spine, 72*(5), 359-371. doi:10.1016/j.jbspin.2004.01.010

Currie, G. L., Delaney, A., Bennett, M. I., Dickenson, A. H., Egan, K. J., Vesterinen, H. M., . . . Fallon, M. T. (2013). Animal models of bone cancer pain: systematic review and meta-analyses. *Pain, 154*(6), 917-926. doi:10.1016/j.pain.2013.02.033

Dalla Costa, E., Bracci, D., Dai, F., Lebelt, D., & Minero, M. (2017). Do Different Emotional States Affect the Horse Grimace Scale Score? A Pilot Study. *Journal of Equine Veterinary Science, 54*, 114-117. doi:10.1016/j.jevs.2017.03.221

Dalla Costa, E., Minero, M., Lebelt, D., Stucke, D., Canali, E., & Leach, M. C. (2014). Development of the Horse Grimace Scale (HGS) as a pain assessment tool in horses undergoing routine castration. *PLoS One, 9*(3), e92281. doi:10.1371/journal.pone.0092281

Dalla Costa, E., Pascuzzo, R., Leach, M. C., Dai, F., Lebelt, D., Vantini, S., & Minero, M. (2018). Can grimace scales estimate the pain status in horses and mice? A statistical approach to identify a classifier. *PLoS One, 13*(8), e0200339. doi:10.1371/journal.pone.0200339

Davoody, L., Quiton, R. L., Lucas, J. M., Ji, Y., Keller, A., & Masri, R. (2011). Conditioned place preference reveals tonic pain in an animal model of central pain. *J Pain, 12*(8), 868-874. doi:10.1016/j.jpain.2011.01.010

de los Santos-Arteaga, M., Sierra-Dominguez, S. A., Fontanella, G. H., Delgado-Garcia, J. M., & Carrion, A. M. (2003). Analgesia induced by dietary restriction is mediated by the kappa-opioid system. *Journal of Neuroscience, 23*(35), 11120-11126.

de Oliveira, G. R. (2002). The construction of learning curves for basic skills in anesthetic procedures: An application for the cumulative sum method. *Anesthesia and Analgesia, 95*(2), 411-416. doi:10.1213/01.Ane.0000021360.03491.B7

De Rantere, D., Schuster, C. J., Reimer, J. N., & Pang, D. S. J. (2016). The relationship between the Rat Grimace Scale and mechanical hypersensitivity testing in three experimental pain models. *European Journal of Pain, 20*(3), 417-426. doi:10.1002/ejp.742

Deacon, R. M. J. (2006). Burrowing in rodents: a sensitive method for detecting behavioral dysfunction. *Nature Protocols, 1*(1), 118-121. doi:10.1038/nprot.2006.19

Deacon, R. M. J. (2009). Burrowing: A sensitive behavioural assay, tested in five species of laboratory rodents. *Behavioural Brain Research, 200*(1), 128-133. doi:10.1016/j.bbr.2009.01.007

Deacon, R. M. J., Croucher, A., & Rawlins, J. N. P. (2002). Hippocampal cytotoxic lesion effects on species-typical behaviours in mice. *Behavioural Brain Research, 132*(2), 203-213. doi:10.1016/s0166-4328(01)00401-6

Defensor, E. B., Corley, M. J., Blanchard, R. J., & Blanchard, D. C. (2012). Facial expressions of mice in aggressive and fearful contexts. *Physiol Behav, 107*(5), 680-685. doi:10.1016/j.physbeh.2012.03.024

Deiteren, A., Vermeulen, W., Moreels, T. G., Pelckmans, P. A., De Man, J. G., & De Winter, B. Y. (2014). The effect of chemically induced colitis, psychological stress and their combination on visceral pain in female Wistar rats. *Stress, 17*(5), 431-444. doi:10.3109/10253890.2014.951034

Delgado-Ruiz, R. A., Luis Calvo-Guirado, J., & Romanos, G. E. (2015). Critical size defects for bone regeneration experiments in rabbit calvariae: systematic review and quality evaluation using ARRIVE guidelines. *Clinical Oral Implants Research, 26*(8), 915-930. doi:10.1111/clr.12406

Denenberg, V. H., Rhoda, E. T., & Zarrow, M. X. (1969). Maternal Behavior in the Rat: An Investigation and Quantification of Nest Building. *Behaviour, 34*(1/2), 1-16.

Denmark, A., Tien, D., Wong, K., Chung, A., Cachat, J., Goodspeed, J., . . . Kalueff, A. V. (2010). The effects of chronic social defeat stress on mouse self-grooming behavior and its patterning. *Behav Brain Res, 208*(2), 553-559. doi:10.1016/j.bbr.2009.12.041

Deuis, J. R., Dvorakova, L. S., & Vetter, I. (2017). Methods Used to Evaluate Pain Behaviors in Rodents. *Front Mol Neurosci, 10*, 284. doi:10.3389/fnmol.2017.00284

Di Giminiani, P., Brierley, V. L., Scollo, A., Gottardo, F., Malcolm, E. M., Edwards, S. A., & Leach, M. C. (2016). 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. doi:10.3389/fvets.2016.00100

Di Girolamo, N., Giuffrida, M. A., Winter, A. L., & Reynders, R. M. (2017). Reporting and communication of randomisation procedures is suboptimal in veterinary trials. *Veterinary Record, 181*(8). doi:10.1136/vr.104035

Dicksved, J., Schreiber, O., Willing, B., Petersson, J., Rang, S., Phillipson, M., . . . Roos, S. (2012). Lactobacillus reuteri Maintains a Functional Mucosal Barrier during DSS Treatment Despite Mucus Layer Dysfunction. *PLoS One, 7*(9), 8. doi:10.1371/journal.pone.0046399

Dieleman, L. A., Palmen, M., Akol, H., Bloemena, E., Pena, A. S., Meuwissen, S. G. M., & van Rees, E. P. (1998). Chronic experimental colitis induced by dextran sulphate sodium (DSS) is characterized by Th1 and Th2 cytokines. *Clinical and Experimental Immunology, 114*(3), 385-391.

D'Mello, R., & Dickenson, A. H. (2008). Spinal cord mechanisms of pain. *Br J Anaesth, 101*(1), 8-16. doi:10.1093/bja/aen088

Dohoo, S. E., & Dohoo, I. R. (1996). Factors influencing the postoperative use of analgesics in dogs and cats by Canadian veterinarians. *Canadian Veterinary Journal-Revue Veterinaire Canadienne, 37*(9), 552-556.

Doodnaught, G. M., Benito, J., Monteiro, B. P., Beauchamp, G., Grasso, S. C., & Steagall, P. V. (2017). Agreement among undergraduate and graduate veterinary students and veterinary anesthesiologists on pain assessment in cats and dogs: A preliminary study. *Canadian Veterinary Journal-Revue Veterinaire Canadienne, 58*(8), 805-808.

Dore-Savard, L., Otis, V., Belleville, K., Lemire, M., Archambault, M., Tremblay, L., . . . Sarret, P. (2010). Behavioral, medical imaging and histopathological features of a new rat model of bone cancer pain. *PLoS One, 5*(10), e13774. doi:10.1371/journal.pone.0013774

Dowling, P., Klinker, F., Amaya, F., Paulus, W., & Liebetanz, D. (2009). Iron-deficiency sensitizes mice to acute pain stimuli and formalin-induced nociception. *J Nutr, 139*(11), 2087-2092. doi:10.3945/jn.109.112557

du Sert, N. P. (2011). Improving the reporting of animal research: when will we ARRIVE? *Disease Models & Mechanisms, 4*(3), 281-282. doi:10.1242/dmm.007971

du Sert, N. P. (2012). Maximising the output of osteoarthritis research: the ARRIVE guidelines. *Osteoarthritis Cartilage, 20*(4), 253-255. doi:10.1016/j.joca.2011.12.017

du Sert, N. P., Hurst, V., Ahluwalia, A., Alam, S., Altman, D. G., Avey, M. T., . . . Holgate, S. T. (2018). Revision of the ARRIVE guidelines: rationale and scope. *BMJ Open Science, 2*(1), e000002. doi:10.1136/bmjos-2018-000002

Eisenberg, E., Vos, B. P., & Strassman, A. M. (1993). The NMDA antagonist Memantine blocks pain behavior in a rat model of formalin-induced facial pain. *Pain, 54*(3), 301-307.

Engelhardt, G., Homma, D., Schlegel, K., Utzmann, R., & Schnitzler, C. (1995). Antiinflammatory, analgesic, antipyretic and related properties of meloxicam, a new nonsteroidal antiinflammatory agent with favorable gastrointestinal tolerance. *Inflammation Research, 44*(10), 423-433. doi:10.1007/bf01757699

Faller, K. M. E., McAndrew, D. J., Schneider, J. E., & Lygate, C. A. (2015). Refinement of analgesia following thoracotomy and experimental myocardial infarction using the Mouse Grimace Scale. *Experimental Physiology, 100*(2), 164-172. doi:10.1113/expphysiol.2014.083139

Farrell, K. E., Callister, R. J., & Keely, S. (2014). Understanding and targeting centrally mediated visceral pain in inflammatory bowel disease. *Front Pharmacol, 5*, 27. doi:10.3389/fphar.2014.00027

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). STATISTICAL-INFERENCE FOR COEFFICIENT-ALPHA. *Applied Psychological Measurement, 11*(1), 93-103. doi:10.1177/014662168701100107

Field, M. J., Bramwell, S., Hughes, J., & Singh, L. (1999). Detection of static and dynamic components of mechanical allodynia in rat models of neuropathic pain: are they signalled by distinct primary sensory neurones? *Pain, 83*(2), 303-311. doi:10.1016/s0304-3959(99)00111-6

Fisher, M., Feuerstein, G., Howells, D. W., Hurn, P. D., Kent, T. A., Savitz, S. I., . . . Grp, S. (2009). Update of the Stroke Therapy Academic Industry Roundtable Preclinical Recommendations. *Stroke, 40*(6), 2244-2250. doi:10.1161/strokeaha.108.541128

Florez-Vargas, O., Brass, A., Karystianis, G., Bramhall, M., Stevens, R., Cruickshank, S., & Nenadic, G. (2016). Bias in the reporting of sex and age in biomedical research on mouse models. *Elife, 5*, 14. doi:10.7554/eLife.13615

Frederickson, R. C. A., Burgis, V., & Edwards, J. D. (1977). Hyperalgesia induced by naloxone follows diurnal rhythm in responsivity to painful stimuli. *Science, 198*(4318), 756-758. doi:10.1126/science.561998

Freedman, L. P., Cockburn, I. M., & Simcoe, T. S. (2015). The Economics of Reproducibility in Preclinical Research. *Plos Biology, 13*(6). doi:10.1371/journal.pbio.1002165

Freedman, L. P., Venugopalan, G., & Wisman, R. (2017). Reproducibility2020: Progress and priorities. *F1000Res, 6*, 604. doi:10.12688/f1000research.11334.1

Frye, C. A., Cuevas, C. A., & Kanarek, R. B. (1993). Diet and estrous-cycle influence pain sensitivity in rats. *Pharmacology Biochemistry and Behavior, 45*(1), 255-260. doi:10.1016/0091-3057(93)90116-b

Fujita, M., Fukuda, T., Sato, Y., Takasusuki, T., & Tanaka, M. (2018). Allopregnanolone suppresses mechanical allodynia and internalization of neurokinin-1 receptors at the spinal dorsal horn in a rat postoperative pain model. *Korean J Pain, 31*(1), 10-15. doi:10.3344/kjp.2018.31.1.10

Gabriel, A. F., Paoletti, G., Della Seta, D., Panelli, R., Marcus, M. A., Farabollini, F., . . . Joosten, E. A. (2010). Enriched environment and the recovery from inflammatory pain: Social versus physical aspects and their interaction. *Behav Brain Res, 208*(1), 90-95. doi:10.1016/j.bbr.2009.11.015

Gambero, A., Marostica, M., Abdalla Saad, M. J., & Pedrazzoli, J., Jr. (2007). Mesenteric adipose tissue alterations resulting from experimental reactivated colitis. *Inflamm Bowel Dis, 13*(11), 1357-1364. doi:10.1002/ibd.20222

Gao, M., Long, H., Ma, W., Liao, L., Yang, X., Zhou, Y., . . . Lai, W. (2016). The role of periodontal ASIC3 in orofacial pain induced by experimental tooth movement in rats. *Eur J Orthod, 38*(6), 577-583. doi:10.1093/ejo/cjv082

Gao, Z., Cui, F., Cao, X., Wang, D., Li, X., & Li, T. (2017). Local infiltration of the surgical wounds with levobupivacaine, dexibuprofen, and norepinephrine to reduce postoperative pain: A randomized, vehicle-controlled, and preclinical study. *Biomed Pharmacother, 92*, 459-467. doi:10.1016/j.biopha.2017.05.038

Gasparetto, M., & Guariso, G. (2013). Highlights in IBD Epidemiology and Its Natural History in the Paediatric Age. *Gastroenterol Res Pract, 2013*, 829040. doi:10.1155/2013/829040

Gaudio, E., Taddei, G., Vetuschi, A., Sferra, R., Frieri, G., Ricciardi, G., & Caprilli, R. (1999). Dextran sulfate sodium (DSS) colitis in rats - Clinical, structural, and ultrastructural aspects. *Digestive Diseases and Sciences, 44*(7), 1458-1475. doi:10.1023/a:1026620322859

Geier, M. S., Butler, R. N., Giffard, P. M., & Howarth, G. S. (2007). Lactobacillus fermentum BR11, a potential new probiotic, alleviates symptoms of colitis induced by dextran sulfate sodium (DSS) in rats. *Int J Food Microbiol, 114*(3), 267-274. doi:10.1016/j.ijfoodmicro.2006.09.018

Gentsch, C., Lichtsteiner, M., Frischknecht, H. R., Feer, H., & Siegfried, B. (1988). Isolation-induced locomotor hyperactivity and hypoalgesia in rats are prevented by handling and reversed by resocialization. *Physiology & Behavior, 43*(1), 13-16. doi:10.1016/0031-9384(88)90091-1

Giamberardino, M. A., Valente, R., Debigontina, P., & Vecchiet, L. (1995). Artificial ureteral calculosis in rats - behavioral characterization of visceral pain episodes and their relationship with referred lumbar muscle hyperalgesia. *Pain, 61*(3), 459-469. doi:10.1016/0304-3959(94)00208-v

Giglio, C. A., Defino, H. L. A., Da-Silva, C. A., De-Souza, A. S., & Del Bel, E. A. (2006). Behavioral and physiological methods for early quantitative assessment of spinal cord injury and prognosis in rats. *Brazilian Journal of Medical and Biological Research, 39*(12), 1613-1623. doi:10.1590/s0100-879x2006001200013

Goldstein, D. J., Offen, W. W., Klein, E. G., Phebus, L. A., Hipskind, P., Johnson, K. W., & Ryan, R. E. (2001). Lanepitant, an NK-1 antagonist, in migraine prevention. *Cephalalgia, 21*(2), 102-106. doi:10.1046/j.1468-2982.2001.00161.x

Goncalves, F. D., Schneider, N., Mello, H. F., Passos, E. P., Meurer, L., Cirne-Lima, E., & Paz, A. H. D. (2013). Characterization of Acute Murine Dextran Sodium Sulfate (DSS) Colitis: Severity of Inflammation is Dependent on the DSS Molecular Weight and Concentration. *Acta Scientiae Veterinariae, 41*, 9.

Gorman, A. L., Yu, C. G., Ruenes, G. R., Daniels, L., & Yezierski, R. P. (2001). Conditions affecting the onset, severity, and progression of a spontaneous pain-like behavior after excitotoxic spinal cord injury. *J Pain, 2*(4), 229-240. doi:10.1054/jpai.2001.22788

Gould, H. J. (2000). Complete Freund's adjuvant-induced hyperalgesia: a human perception. *Pain, 85*(1-2), 301-303. doi:10.1016/s0304-3959(99)00289-4

Gould, S. A., Doods, H., Lamla, T., & Pekcec, A. (2016). Pharmacological characterization of intraplantar Complete Freund's Adjuvant-induced burrowing deficits. *Behav Brain Res, 301*, 142-151. doi:10.1016/j.bbr.2015.12.019

Grant, S., & Group, C.-S. (2019). The CONSORT-SPI 2018 extension: a new guideline for reporting social and psychological intervention trials. *Addiction, 114*(1), 4-8. doi:10.1111/add.14411

Greenspan, J. D., Craft, R. M., LeResche, L., Arendt-Nielsen, L., Berkley, K. J., Fillingim, R. B., . . . Pain, S. I. G. o. t. I. (2007). Studying sex and gender differences in pain and analgesia: a consensus report. *Pain, 132 Suppl 1*, S26-45. doi:10.1016/j.pain.2007.10.014

Griffin, G., Clark, J. M., Zurlo, J., & Ritskes-Hoitinga, M. (2014). Scientific uses of animals: harm-benefit analysis and complementary approaches to implementing the Three Rs. *Revue Scientifique Et Technique-Office International Des Epizooties, 33*(1), 265-272. doi:10.20506/rst.33.1.2283

Guesgen, M. J., Beausoleil, N. J., Leach, M., Minot, E. O., Stewart, M., & Stafford, K. J. (2016). Coding and quantification of a facial expression for pain in lambs. *Behav Processes, 132*, 49-56. doi:10.1016/j.beproc.2016.09.010

Guo, P., & Hu, S. P. (2017). Thalidomide alleviates postoperative pain and spatial memory deficit in aged rats. *Biomed Pharmacother, 95*, 583-588. doi:10.1016/j.biopha.2017.08.114

Häger, C., Biernot, S., Buettner, M., Glage, S., Keubler, L. M., Held, N., . . . Bleich, A. (2017). The Sheep Grimace Scale as an indicator of post-operative distress and pain in laboratory sheep. *PLoS One, 12*(4), e0175839. doi:10.1371/journal.pone.0175839

Hager, C., Keubler, L. M., Biernot, S., Dietrich, J., Buchheister, S., Buettner, M., & Bleich, A. (2015). Time to Integrate to Nest Test Evaluation in a Mouse DSS-Colitis Model. *PLoS One, 10*(12), e0143824. doi:10.1371/journal.pone.0143824

Haidet, K. K., Tate, J., Divirgilio-Thomas, D., Kolanowski, A., & Happ, M. B. (2009). Methods to Improve Reliability of Video-Recorded Behavioral Data. *Research in Nursing & Health, 32*(4), 465-474. doi:10.1002/nur.20334

Hall, L. J., Faivre, E., Quinlan, A., Shanahan, F., Nally, K., & Melgar, S. (2011). Induction and Activation of Adaptive Immune Populations During Acute and Chronic Phases of a Murine Model of Experimental Colitis. *Digestive Diseases and Sciences, 56*(1), 79-89. doi:10.1007/s10620-010-1240-3

Han, J. S., Bird, G. C., Li, W., Jones, J., & Neugebauer, V. (2005). Computerized analysis of audible and ultrasonic vocalizations of rats as a standardized measure of pain-related behavior. *J Neurosci Methods, 141*(2), 261-269. doi:10.1016/j.jneumeth.2004.07.005

Han, S., Olonisakin, T. F., Pribis, J. P., Zupetic, J., Yoon, J. H., Holleran, K. M., . . . Lee, J. S. (2017). A checklist is associated with increased quality of reporting preclinical biomedical research: A systematic review. *PLoS One, 12*(9). doi:10.1371/journal.pone.0183591

Haney, M., & Miczek, K. A. (1993). Ultrasounds during agonistic interactions between female rats (rattus-norvegicus). *Journal of Comparative Psychology, 107*(4), 373-379. doi:10.1037//0735-7036.107.4.373

Hargraves, W. A., & Hentall, I. D. (2005). Analgesic effects of dietary caloric restriction in adult mice. *Pain, 114*(3), 455-461. doi:10.1016/j.pain.2005.01.010

Hargreaves, K., Dubner, R., Brown, F., Flores, C., & Joris, J. (1988). A NEW AND SENSITIVE METHOD FOR MEASURING THERMAL NOCICEPTION IN CUTANEOUS HYPERALGESIA. *Pain, 32*(1), 77-88. doi:10.1016/0304-3959(88)90026-7

Harris, H. M., Carpenter, J. M., Black, J. R., Smitherman, T. A., & Sufka, K. J. (2017). The effects of repeated nitroglycerin administrations in rats; modeling migraine-related endpoints and chronification. *J Neurosci Methods, 284*, 63-70. doi:10.1016/j.jneumeth.2017.04.010

Hassan, E. A., Ramadan, H. K., Ismael, A. A., Mohamed, K. F., El-Attar, M. M., & Alhelali, I. (2017). Noninvasive biomarkers as surrogate predictors of clinical and endoscopic remission after infliximab induction in patients with refractory ulcerative colitis. *Saudi Journal of Gastroenterology, 23*(4), 238-245. doi:10.4103/sjg.SJG_599_16

He, Y., Tian, X., Hu, X., Porreca, F., & Wang, Z. J. (2012). Negative reinforcement reveals non-evoked ongoing pain in mice with tissue or nerve injury. *J Pain, 13*(6), 598-607. doi:10.1016/j.jpain.2012.03.011

Heinrich, M., Mechea, A., & Hoffmann, F. (2016). Improving postoperative pain management in children by providing regular training and an updated pain therapy concept. *Eur J Pain, 20*(4), 586-593. doi:10.1002/ejp.770

Hess, S. E., Rohr, S., Dufour, B. D., Gaskill, B. N., Pajor, E. A., & Garner, J. P. (2008). Home Improvement: C57BL/6J Mice Given More Naturalistic Nesting Materials Build Better Nests. *Journal of the American Association for Laboratory Animal Science, 47*(6), 25-31.

Hewson, C. J., Dohoo, I. R., & Lemke, K. A. (2006). Factors affecting the use of postincisional analgesics in dogs and cats by Canadian veterinarians in 2001. *Canadian Veterinary Journal-Revue Veterinaire Canadienne, 47*(5), 453-459.

Hewson, C. J., Dohoo, I. R., Lemke, K. A., & Barkema, H. W. (2007). Factors affecting Canadian veterinarians' use of analgesics when dehorning beef and dairy calves. *Canadian Veterinary Journal-Revue Veterinaire Canadienne, 48*(11), 1129-1136.

Hill, R. (2000). NK1 (substance P) receptor antagonists – why are they not analgesic in humans? *Trends in Pharmacological Sciences, 21*(7), 244-246. doi:https://doi.org/10.1016/S0165-6147(00)01502-9

Holden, E., Calvo, G., Collins, M., Bell, A., Reid, J., Scott, E. M., & Nolan, A. M. (2014).

Evaluation of facial expression in acute pain in cats. *Journal of Small Animal Practice, 55*(12), 615-621. doi:10.1111/jsap.12283

Holman, L., Head, M. L., Lanfear, R., & Jennions, M. D. (2015). Evidence of Experimental Bias in the Life Sciences: Why We Need Blind Data Recording. *Plos Biology, 13*(7), 12. doi:10.1371/journal.pbio.1002190

Hooijmans, C. R., & Ritskes-Hoitinga, M. (2013). Progress in Using Systematic Reviews of Animal Studies to Improve Translational Research. *Plos Medicine, 10*(7), 4. doi:10.1371/journal.pmed.1001482

Huang, W. L., Calvo, M., Karu, K., Olausen, H. R., Bathgate, G., Okuse, K., . . . Rice, A. S. C. (2013). A clinically relevant rodent model of the HIV antiretroviral drug stavudine induced painful peripheral neuropathy. *Pain, 154*(4), 560-575. doi:10.1016/j.pain.2012.12.023

Hugonnard, M., Leblond, A., Keroack, S., Cadore, J. L., & Troncy, E. (2004). Attitudes and concerns of French veterinarians towards pain and analgesia in dogs and cats. *Vet Anaesth Analg, 31*(3), 154-163. doi:10.1111/j.1467-2987.2004.00175.x

Ibrahiem, E. H. I., Nigam, V. N., Brailovsky, C. A., Madarnas, P., & Elhilali, M. (1983). Orthotopic implantation of primary n- 4-(5-nitro-2-furyl)-2-thiazolyl formamide-induced bladder-cancer in bladder submucosa - an animal-model for bladder-cancer study. *Cancer Research, 43*(2), 617-622.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *Plos Medicine, 2*(8), 696-701. doi:10.1371/journal.pmed.0020124

Jain, P., Hassan, A. M., Koyani, C. N., Mayerhofer, R., Reichmann, F., Farzi, A., . . . Holzer, P. (2015). Behavioral and molecular processing of visceral pain in the brain of mice: impact of colitis and psychological stress. *Front Behav Neurosci, 9*, 177. doi:10.3389/fnbeh.2015.00177

Jeger, V., Arrigo, M., Hildenbrand, F. F., Muller, D., Jirkof, P., Hauffe, T., . . . Rudiger, A. (2017). Improving animal welfare using continuous nalbuphine infusion in a long-term rat model of sepsis. *Intensive Care Med Exp, 5*(1), 23. doi:10.1186/s40635-017-0137-2

Jegstrup, I. M., Vestergaard, R., Vach, W., & Ritskes-Hoitinga, M. (2005). Nest-building behaviour in male rats from three inbred strains: BN/HsdCpb, BDIX/OrIIco and LEW/Mol. *Animal Welfare, 14*(2), 149-156.

Jin, Y., Sanger, N., Shams, I., Luo, C., Shahid, H., Li, G., . . . Samaan, Z. (2018). Does the medical literature remain inadequately described despite having reporting guidelines for 21 years? - A systematic review of reviews: an update. *J Multidiscip Healthc, 11*, 495-510. doi:10.2147/JMDH.S155103

Jirkof, P., Cesarovic, N., Rettich, A., Nicholls, F., Seifert, B., & Arras, M. (2010). Burrowing behavior as an indicator of post-laparotomy pain in mice. *Front Behav Neurosci, 4*, 165. doi:10.3389/fnbeh.2010.00165

Jirkof, P., Fleischmann, T., Cesarovic, N., Rettich, A., Vogel, J., & Arras, M. (2013). Assessment of postsurgical distress and pain in laboratory mice by nest complexity scoring. *Lab Anim, 47*(3), 153-161. doi:10.1177/0023677213475603

Johansen, J. P., Fields, H. L., & Manning, B. H. (2001). The affective component of pain in rodents: Direct evidence for a contribution of the anterior cingulate cortex. *Proceedings of the National Academy of Sciences of the United States of America, 98*(14), 8077-8082. doi:10.1073/pnas.141218998

Jolles, J., Rompa-Barendregt, J., & Gispen, W. H. (1979). Novelty and grooming behavior in the rat. *Behavioral & Neural Biology, 25*(4), 563-572. doi:10.1016/S0163-1047(79)90362-5

Jourdan, D., Ardid, D., & Eschalier, A. (2002). Analysis of ultrasonic vocalisation does not allow chronic pain to be evaluated in rats. *Pain, 95*(1-2), 165-173. doi:10.1016/s0304-3959(01)00394-3

Jourdan, D., Ardid, D., Chapuy, E., Eschalier, A., & Lebars, D. (1995). Audible and ultrasonic vocalization elicited by single electrical nociceptive stimuli to the tail in the rat. *Pain, 63*(2), 237-249. doi:10.1016/0304-3959(95)00049-x

Jurjus, A. R., Khoury, N. N., & Reimund, J. M. (2004). Animal models of inflammatory bowel disease. *J Pharmacol Toxicol Methods, 50*(2), 81-92. doi:10.1016/j.vascn.2003.12.002

Kalueff, A. V., Stewart, A. M., Song, C., Berridge, K. C., Graybiel, A. M., & Fentress, J. C. (2016). Neurobiology of rodent self-grooming and its value for translational neuroscience. *Nat Rev Neurosci, 17*(1), 45-59. doi:10.1038/nrn.2015.8

Kawano, T., Eguchi, S., Iwata, H., Yamanaka, D., Tateiwa, H., Locatelli, F. M., & Yokoyama, M. (2016). Effects and underlying mechanisms of endotoxemia on post-incisional pain in rats. *Life Sci, 148*, 145-153. doi:10.1016/j.lfs.2016.01.046

Kawano, T., Takahashi, T., Iwata, H., Morikawa, A., Imori, S., Waki, S., . . . Yokoyama, M. (2014). Effects of ketoprofen for prevention of postoperative cognitive dysfunction in aged rats. *J Anesth, 28*(6), 932-936. doi:10.1007/s00540-014-1821-y

Keating, S. C., Thomas, A. A., Flecknell, P. A., & Leach, M. C. (2012). Evaluation of EMLA cream for preventing pain during tattooing of rabbits: changes in physiological, behavioural and facial expression responses. *PLoS One, 7*(9), e44437. doi:10.1371/journal.pone.0044437

Khoo, S. Y., Lay, B. P. P., Joya, J., & McNally, G. P. (2018). Local anaesthetic refinement of pentobarbital euthanasia reduces abdominal writhing without affecting immunohistochemical endpoints in rats. *Lab Anim, 52*(2), 152-162. doi:10.1177/0023677217721260

Kihara, N., de la Fuente, S. G., Fujino, K., Takahashi, T., Pappas, T. N., & Mantyh, C. R. (2003). Vanilloid receptor-1 containing primary sensory neurones mediate dextran sulphate sodium induced colitis in rats. *Gut, 52*(5), 713-719. doi:10.1136/gut.52.5.713

Kilkenny, C., Browne, W. J., Cuthill, I. C., Emerson, M., & Altman, D. G. (2010). Improving Bioscience Research Reporting: The ARRIVE Guidelines for Reporting Animal Research. *Plos Biology, 8*(6). doi:10.1371/journal.pbio.1000412

Kilkenny, C., Parsons, N., Kadyszewski, E., Festing, M. F. W., Cuthill, I. C., Fry, D., . . . Altman, D. G. (2009). Survey of the Quality of Experimental Design, Statistical Analysis and Reporting of Research Using Animals. *PLoS One, 4*(11). doi:10.1371/journal.pone.0007824

King, T., Vera-Portocarrero, L., Gutierrez, T., Vanderah, T. W., Dussor, G., Lai, J., . . . Porreca, F. (2009). Unmasking the tonic-aversive state in neuropathic pain. *Nat Neurosci, 12*(11), 1364-1366. doi:10.1038/nn.2407

Kitajima, S., Takuma, S., & Morimoto, M. (2000). Histological analysis of murine colitis induced by dextran sulfate sodium of different molecular weights. *Experimental Animals, 49*(1), 9-15. doi:10.1538/expanim.49.9

Klinck, M. P., Mogil, J. S., Moreau, M., Lascelles, B. D. X., Flecknell, P. A., Poitte, T., & Troncy, E. (2017). Translational pain assessment: could natural animal models be the missing link? *Pain, 158*(9), 1633-1646. doi:10.1097/j.pain.0000000000000978

Klopfenstein, C. E., Herrmann, F. R., Mamie, C., Van Gessel, E., & Forster, A. (2000). Pain intensity and pain relief after surgery - A comparison between patients' reported assessments and nurses' and physicians' observations. *Acta Anaesthesiologica Scandinavica, 44*(1), 58-62. doi:10.1034/j.1399-6576.2000.440111.x

Knutson, B., Burgdorf, J., & Panksepp, J. (1998). Anticipation of play elicits high-frequency ultrasonic vocalizations in young rats. *Journal of Comparative Psychology, 112*(1), 65-73. doi:10.1037/0735-7036.112.1.65

Korat, P. S., & Kapupara, P. P. (2017). Local infiltration of the surgical wound with levobupivacaine, ibuprofen, and epinephrine in postoperative pain: An experimental study. *Biomed Pharmacother, 96*, 104-111. doi:10.1016/j.biopha.2017.09.131

Kullmann, F., Messmann, H., Alt, M., Gross, V., Bocker, T., Schölmerich, J., & Rüschoff, J. (2001). Clinical and histopathological features of dextran sulfate sodium induced acute and chronic colitis associated with dysplasia in rats. *International Journal of Colorectal Disease, 16*(4), 238-246. doi:10.1007/s003840100311

Kunz, M., Scharmann, S., Hemmeter, U., Schepelmann, K., & Lautenbacher, S. (2007). The facial expression of pain in patients with dementia. *Pain, 133*(1-3), 221-228. doi:10.1016/j.pain.2007.09.007

Kurejova, M., Nattenmuller, U., Hildebrandt, U., Selvaraj, D., Stosser, S., & Kuner, R. (2010). An improved behavioural assay demonstrates that ultrasound vocalizations constitute a reliable indicator of chronic cancer pain and neuropathic pain. *Molecular Pain, 6*, 7. doi:10.1186/1744-8069-6-18

Kurina, L. M., Goldacre, M. J., Yeates, D., & Gill, L. E. (2001). Depression and anxiety in people with inflammatory bowel disease. *Journal of Epidemiology and Community Health, 55*(10), 716-720. doi:10.1136/jech.55.10.716

Kuzmic P. 2015. Critical values of F-statistics. Available at http://www.biokin.com/tools/f-critical.html (accessed 26 February 2018).

Landis, S. C., Amara, S. G., Asadullah, K., Austin, C. P., Blumenstein, R., Bradley, E. W., . . . Silberberg, S. D. (2012). A call for transparent reporting to optimize the predictive value of preclinical research. *Nature, 490*(7419), 187-191. doi:10.1038/nature11556

Langford, D. J., Bailey, A. L., Chanda, M. L., Clarke, S. E., Drummond, T. E., Echols, S., . . . Mogil, J. S. (2010). Coding of facial expressions of pain in the laboratory mouse. *Nature Methods, 7*(6), 447-U452. doi:10.1038/nmeth.1455

Langford, D. J., Crager, S. E., Shehzad, Z., Smith, S. B., Sotocinal, S. G., Levenstadt, J. S., . . . Mogil, J. S. (2006). Social modulation of pain as evidence for empathy in mice. *Science, 312*(5782), 1967-1970. doi:10.1126/science.1128322

Lapointe, T. K., Basso, L., Iftinca, M. C., Flynn, R., Chapman, K., Dietrich, G., . . . Altier, C. (2015). TRPV1 sensitization mediates postinflammatory visceral pain following acute colitis. *Am J Physiol Gastrointest Liver Physiol, 309*(2), G87-99. doi:10.1152/ajpgi.00421.2014

Larauche, M., Mulak, A., & Tache, Y. (2012). Stress and visceral pain: from animal models to clinical therapies. *Exp Neurol, 233*(1), 49-67. doi:10.1016/j.expneurol.2011.04.020

Lariviere, W. R., Wilson, S. G., Laughlin, T. M., Kokayeff, A., West, E. E., Adhikari, S. M., . . . Mogil, J. S. (2002). Heritability of nociception. III. Genetic relationships among commonly used assays of nociception and hypersensitivity. *Pain, 97*(1-2), 75-86. doi:10.1016/s0304-3959(01)00492-4

Larsson, M. H., Rapp, L., & Lindstrom, E. (2006). Effect of DSS-induced colitis on visceral sensitivity to colorectal distension in mice. *Neurogastroenterol Motil, 18*(2), 144-152. doi:10.1111/j.1365-2982.2005.00736.x

Latremoliere, A., & Woolf, C. J. (2009). Central sensitization: a generator of pain hypersensitivity by central neural plasticity. *J Pain, 10*(9), 895-926. doi:10.1016/j.jpain.2009.06.012

Lau, W., Dykstra, C., Thevarkunnel, S., Silenieks, L. B., de Lannoy, I. A. M., Lee, D. K. H., & Higgins, G. A. (2013). A back translation of pregabalin and carbamazepine against evoked and non-evoked endpoints in the rat spared nerve injury model of neuropathic pain. *Neuropharmacology, 73*, 204-215. doi:10.1016/j.neuropharm.2013.05.023

Le Bars, D., Gozariu, M., & Cadden, S. W. (2001). Animal models of nociception. *Pharmacological Reviews, 53*(4), 597-652.

Leach, M. C., Bowell, V. A., Allan, T. F., & Morton, D. B. (2002). Degrees of aversion shown by rats and mice to different concentrations of inhalational anaesthetics. *Veterinary Record, 150*(26), 808-815. doi:10.1136/vr.150.26.808

Leach, M. C., Coulter, C. A., Richardson, C. A., & Flecknell, P. A. (2011). Are We Looking in the Wrong Place? Implications for Behavioural-Based Pain Assessment in Rabbits (Oryctolagus cuniculi) and Beyond? *PLoS One, 6*(3), 9.

doi:10.1371/journal.pone.0013347

Legg, E. D., Novejarque, A., & Rice, A. S. (2009). The Three Ages of Rat: the influence of rodent age on affective and cognitive outcome measures in peripheral neuropathic pain. *Pain, 144*(1-2), 12-13. doi:10.1016/j.pain.2009.04.023

Leung, V., Benoit-Biancamano, M. O. & Pang, D. S. J. (In press). Performance of behavioral essays: the Rat Grimace Scale, burrowing and a composite behavior score to identify visceral pain in an acute and chronic colitis model. *Pain Rep*. ID: PAINREPORTS-D-18-0096

Leung, V., Rousseau-Blass, F., Beauchamp, G., & Pang, D. S. J. (2018). ARRIVE has not ARRIVEd: Support for the ARRIVE (Animal Research: Reporting of in vivo Experiments) guidelines does not improve the reporting quality of papers in animal welfare, analgesia or anesthesia. *PLoS One, 13*(5), e0197882. doi:10.1371/journal.pone.0197882

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. doi:10.1038/srep31667

Li, J. X. (2013). The application of conditioning paradigms in the measurement of pain. *Eur J Pharmacol, 716*(1-3), 158-168. doi:10.1016/j.ejphar.2013.03.002

Li, S., & Deng, W. (2018). Deep Facial Expression Recognition: A Survey. *arXiv* :1804.08348

Liao, L., Long, H., Zhang, L., Chen, H., Zhou, Y., Ye, N., & Lai, W. (2014). Evaluation of pain in rats through facial expression following experimental tooth movement. *Eur J Oral Sci, 122*(2), 121-124. doi:10.1111/eos.12110

Lim, M. Y., Chen, H. C., & Omar, M. A. (2014). Assessment of post-operative pain in cats: a case study on veterinary students of Universiti Putra Malaysia. *J Vet Med Educ, 41*(2), 197-203. doi:10.3138/jvme.0713-099R1

Liu, P., Okun, A., Ren, J., Guo, R. C., Ossipov, M. H., Xie, J., . . . Porreca, F. (2011). Ongoing pain in the MIA model of osteoarthritis. *Neurosci Lett, 493*(3), 72-75. doi:10.1016/j.neulet.2011.01.027

Liu, Y., Zhao, X., Mai, Y., Li, X., Wang, J., Chen, L., . . . Feng, Y. (2016). Adherence to ARRIVE Guidelines in Chinese Journal Reports on Neoplasms in Animals. *PLoS One, 11*(5). doi:10.1371/journal.pone.0154657

Loder, E. W., & Penzien, D. B. (2009). Improving the Quality of Research Reporting:HeadacheSteps Up to the Plate. *Headache: The Journal of Head and Face Pain, 49*(3), 335-340. doi:10.1111/j.1526-4610.2009.01356.x

Long, H., Liao, L., Gao, M., Ma, W., Zhou, Y., Jian, F., . . . Lai, W. (2015). Periodontal CGRP contributes to orofacial pain following experimental tooth movement in rats. *Neuropeptides, 52*, 31-37. doi:10.1016/j.npep.2015.06.006

Long, H., Liao, L., Zhou, Y., Shan, D., Gao, M., Huang, R., . . . Lai, W. (2017). A novel technique of delivering viral vectors to trigeminal ganglia in rats. *Eur J Oral Sci, 125*(1), 1-7. doi:10.1111/eos.12326

Lorenz, J., Beck, H., & Bromm, B. (1997). Cognitive performance, mood and experimental pain before and during morphine-induced analgesia in patients with chronic non-malignant pain. *Pain, 73*(3), 369-375. doi:10.1016/s0304-3959(97)00123-1

Ludolph, A. C., Bendotti, C., Blaugrund, E., Chio, A., Greensmith, L., Loeffler, J.-P., . . . von Horsten, S. (2010). Guidelines for preclinical animal research in ALS/MND: A consensus meeting. *Amyotrophic Lateral Sclerosis, 11*(1-2), 38-45. doi:10.3109/17482960903545334

Ma, B., Xu, J. K., Wu, W. J., Liu, H. Y., Kou, C. K., Liu, N., & Zhao, L. L. (2017). Survey of basic medical researchers on the awareness of animal experimental designs and reporting standards in China. *PLoS One, 12*(4), 12. doi:10.1371/journal.pone.0174530

MacCallum, C. J. (2010). Reporting Animal Studies: Good Science and a Duty of Care. *Plos Biology, 8*(6), 2. doi:10.1371/journal.pbio.1000413

Macleod, M. R. (2017). Findings of a retrospective, controlled cohort study of the impact of a change in Nature journals&#039; editorial policy for life sciences research on the completeness of reporting study design and execution. *bioRxiv*, 187245. doi:10.1101/187245

Macleod, M. R., McLean, A. L., Kyriakopoulou, A., Serghiou, S., de Wilde, A., Sherratt, N., . . . Sena, E. S. (2015). Risk of Bias in Reports of In Vivo Research: A Focus for Improvement. *Plos Biology, 13*(10). doi:10.1371/journal.pbio.1002273

Macleod, M. R., van der Worp, H. B., Sena, E. S., Howells, D. W., Dirnagl, U., & Donnan, G. A. (2008). Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. *Stroke, 39*(10), 2824-2829. doi:10.1161/strokeaha.108.515957

MacRae, A. M., Joanna Makowska, I., & Fraser, D. (2018). Initial evaluation of facial expressions and behaviours of harbour seal pups (Phoca vitulina) in response to tagging and microchipping. *Applied Animal Behaviour Science, 205*, 167-174. doi:10.1016/j.applanim.2018.05.001

Makowska, I. J., & Weary, D. M. (2016). The importance of burrowing, climbing and standing upright for laboratory rats. *R Soc Open Sci, 3*(6), 160136. doi:10.1098/rsos.160136

Mao, J. (2009). Translational pain research: achievements and challenges. *J Pain, 10*(10), 1001-1011. doi:10.1016/j.jpain.2009.06.002

Martin, Y. B., & Avendano, C. (2009). Effects of removal of dietary polyunsaturated fatty acids on plasma extravasation and mechanical allodynia in a trigeminal neuropathic pain model. *Mol Pain, 5*, 8. doi:10.1186/1744-8069-5-8

Matsumiya, L. C., Sorge, R. E., Sotocinal, S. G., Tabaka, J. M., Wieskopf, J. S., Zaloum, A., . . . Mogil, J. S. (2012). Using the Mouse Grimace Scale to Reevaluate the Efficacy of Postoperative Analgesics in Laboratory Mice. *Journal of the American Association for Laboratory Animal Science, 51*(1), 42-49.

McGrath, J. C., & Lilley, E. (2015). Implementing guidelines on reporting research using animals (ARRIVE *etc*.): new requirements for publication in BJP. *British Journal of Pharmacology, 172*(13), 3189-3193. doi:10.1111/bph.12955

McLennan, K. M., Rebelo, C. J. B., Corke, M. J., Holmes, M. A., Leach, M. C., & Constantino-Casas, F. (2016). Development of a facial expression scale using footrot and mastitis as models of pain in sheep. *Applied Animal Behaviour Science, 176*, 19-26. doi:10.1016/j.applanim.2016.01.007

Medhurst, S. J., Walker, K., Bowes, M., Kidd, B. L., Glatt, M., Muller, M., . . . Urban, L. (2002). A rat model of bone cancer pain. *Pain, 96*(1-2), 129-140. doi:10.1016/s0304-3959(01)00437-7

Melgar, S., Karlsson, A., & Michaelsson, E. M. (2005). Acute colitis induced by dextran sulfate sodium progresses to chronicity in C57BL/6 but not in BALB/c mice: correlation between symptoms and inflammation. *American Journal of Physiology-Gastrointestinal and Liver Physiology, 288*(6), G1328-G1338. doi:10.1152/ajpgi.00467.2004

Messaoudi, M., Desor, D., Grasmuck, V., Joyeux, M., Langlois, A., & Roman, F. J. (1999). Behavioral evaluation of visceral pain in a rat model of colonic inflammation. *Neuroreport, 10*(5), 1137-1141. doi:10.1097/00001756-199904060-00043

Mich, P. M., Hellyer, P. W., Kogan, L., & Schoenfeld-Tacher, R. (2010). Effects of a pilot training program on veterinary students' pain knowledge, attitude, and assessment skills. *J Vet Med Educ, 37*(4), 358-368. doi:10.3138/jvme.37.4.358

Millecamps, M., Etienne, M., Jourdan, D., Eschalier, A., & Ardid, D. (2004). Decrease in non-selective, non-sustained attention induced by a chronic visceral inflammatory state as a new pain evaluation in rats. *Pain, 109*(3), 214-224. doi:10.1016/j.pain.2003.12.028

Miller, A. L., & Leach, M. C. (2015). The Mouse Grimace Scale: A Clinically Useful Tool? *PLoS One, 10*(9). doi:10.1371/journal.pone.0136000

Miller, A. L., Golledge, H. D., & Leach, M. C. (2016). The Influence of Isoflurane Anaesthesia on the Rat Grimace Scale. *PLoS One, 11*(11), e0166652. doi:10.1371/journal.pone.0166652

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 Molecules and Diseases, 57*, 58-66. doi:10.1016/j.bcmd.2015.12.006

Moehring, F., O'Hara, C. L., & Stucky, C. L. (2016). Bedding Material Affects Mechanical Thresholds, Heat Thresholds, and Texture Preference. *J Pain, 17*(1), 50-64. doi:10.1016/j.jpain.2015.08.014

Mogil, J. S. (2009). Animal models of pain: progress and challenges. *Nat Rev Neurosci, 10*(4), 283-294. doi:10.1038/nrn2606

Mogil, J. S., & Crager, S. E. (2004). What should we be measuring in behavioral studies of chronic pain in animals? *Pain, 112*(1-2), 12-15. doi:10.1016/j.pain.2004.09.028

Mogil, J. S., Davis, K. D., & Derbyshire, S. W. (2010). The necessity of animal models in pain research. *Pain, 151*(1), 12-17. doi:10.1016/j.pain.2010.07.015

Mogil, J. S., Miermeister, F., Seifert, F., Strasburg, K., Zimmermann, K., Reinold, H., . . . Reeh, P. W. (2005). Variable sensitivity to noxious heat is mediated by differential expression of the CGRP gene. *Proceedings of the National Academy of Sciences of the United States of America, 102*(36), 12938-12943. doi:10.1073/pnas.0503264102

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gotzsche, P. C., Devereaux, P. J., . . . Altman, D. G. (2010). CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ, 340*, c869. doi:10.1136/bmj.c869

Moher, D., Jones, A., Lepage, L., & Grp, C. (2001). Use of the CONSORT statement and quality of reports of randomized trials - A comparative before-and-after evaluation. *Jama-Journal of the American Medical Association, 285*(15), 1992-1995. doi:10.1001/jama.285.15.1992

Moher, D., Schulz, K. F., & Altman, D. G. (2005). The CONSORT Statement. Revised recommendations on improving the quality of reports on parallel-design randomized studies. *Schmerz, 19*(2), 156-+. doi:10.1007/s004832-004-0380-9

Mullard, J., Berger, J. M., Ellis, A. D., & Dyson, S. (2017). Development of an ethogram to describe facial expressions in ridden horses (FEReq). *Journal of Veterinary Behavior-Clinical Applications and Research, 18*, 7-12. doi:10.1016/j.jveb.2016.11.005

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., . . . Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*(1). doi:10.1038/s41562-016-0021

Munter, N. H., Stevanovic, A., Rossaint, R., Stoppe, C., Sanders, R. D., & Coburn, M. (2015). CONSORT item adherence in top ranked anaesthesiology journals in 2011: a retrospective analysis. *Eur J Anaesthesiol, 32*(2), 117-125. doi:10.1097/EJA.0000000000000176

Muralidharan, A., Kuo, A., Jacob, M., Lourdesamy, J. S., De Carvalho, L. M. P., Nicholson, J. R., . . . Smith, M. T. (2016). Comparison of Burrowing and Stimuli-Evoked Pain Behaviors as End-Points in Rat Models of Inflammatory Pain and Peripheral Neuropathic Pain. *Frontiers in Behavioral Neuroscience, 10*. doi:10.3389/fnbeh.2016.00088

Naito, H., Okumura, T., Inoue, M., & Suzuki, Y. (2006). Ultrasonic vocalization response elicited in adjuvant-induced arthritic rats as a useful method for evaluating analgesic drugs. *Experimental Animals, 55*(2), 125-129. doi:10.1538/expanim.55.125

Nakashima, S. F., Ukezono, M., Nishida, H., Sudo, R., & Takano, Y. (2015). Receiving of emotional signal of pain from conspecifics in laboratory rats. *R Soc Open Sci, 2*(4), 140381. doi:10.1098/rsos.140381

Nam, M. H., Chun, M. S., Seong, J. K., & Kim, H. G. (2018). Ensuring reproducibility and ethics in animal experiments reporting in Korea using the ARRIVE guideline. *Lab Anim Res, 34*(1), 11-19. doi:10.5625/lar.2018.34.1.11

Negus, S. S., Neddenriep, B., Altarifi, A. A., Carroll, F. I., Leitl, M. D., & Miller, L. L. (2015). Effects of ketoprofen, morphine, and kappa opioids on pain-related depression of nesting in mice. *Pain, 156*(6), 1153-1160. doi:10.1097/j.pain.0000000000000171

Ness, T. J., & Gebhart, G. F. (1987). CHARACTERIZATION OF NEURONAL RESPONSES TO NOXIOUS VISCERAL AND SOMATIC STIMULI IN THE MEDIAL LUMBOSACRAL SPINAL-CORD OF THE RAT. *Journal of Neurophysiology, 57*(6), 1867-1892.

Nishikawa, T., & Tanaka, M. (1978). ALTERED BEHAVIORAL-RESPONSES TO INTENSE FOOT SHOCK IN SOCIALLY-ISOLATED RATS. *Pharmacology Biochemistry and Behavior, 8*(1), 61-67. doi:10.1016/0091-3057(78)90124-7

Nunamaker, E. A., Goldman, J. L., Adams, C. R., & Fortman, J. D. (2018). Evaluation of Analgesic Efficacy of Meloxicam and 2 Formulations of Buprenorphine after Laparotomy in Female Sprague-Dawley Rats. *J Am Assoc Lab Anim Sci, 57*(5), 498-507. doi:10.30802/AALAS-JAALAS-17-000129

Nyuyki, K. D., & Pittman, Q. J. (2015). Toward a better understanding of the central consequences of intestinal inflammation. *Ann N Y Acad Sci, 1351*, 149-154. doi:10.1111/nyas.12935

Ogren, S. O., & Berge, O. G. (1984). TEST-DEPENDENT VARIATIONS IN THE ANTINOCICEPTIVE EFFECT OF PARA-CHLOROAMPHETAMINE-INDUCED RELEASE OF 5-HYDROXYTRYPTAMINE. *Neuropharmacology, 23*(8), 915-924. doi:10.1016/0028-3908(84)90005-4

Okayasu, I., Hatakeyama, S., Yamada, M., Ohkusa, T., Inagaki, Y., & Nakaya, R. (1990). A NOVEL METHOD IN THE INDUCTION OF RELIABLE EXPERIMENTAL ACUTE AND CHRONIC ULCERATIVE-COLITIS IN MICE. *Gastroenterology, 98*(3), 694-702. doi:10.1016/0016-5085(90)90290-h

Okun, A., DeFelice, M., Eyde, N., Ren, J. Y., Mercado, R., King, T., & Porreca, F. (2011). Transient inflammation-induced ongoing pain is driven by TRPV1 sensitive afferents. *Molecular Pain, 7*, 11. doi:10.1186/1744-8069-7-4

Okun, A., Liu, P., Davis, P., Ren, J., Remeniuk, B., Brion, T., . . . Porreca, F. (2012). Afferent drive elicits ongoing pain in a model of advanced osteoarthritis. *Pain, 153*(4), 924-933. doi:10.1016/j.pain.2012.01.022

Oliver, V. L., Athavale, S., Simon, K. E., Kendall, L. V., Nemzek, J. A., & Lofgren, J. L. (2017). Evaluation of Pain Assessment Techniques and Analgesia Efficacy in a Female Guinea Pig (Cavia porcellus) Model of Surgical Pain. *Journal of the American Association for Laboratory Animal Science, 56*(4), 425-435.

Oliver, V. L., Thurston, S. E., & Lofgren, J. L. (2018). Using Cageside Measures to Evaluate Analgesic Efficacy in Mice (Mus musculus) after Surgery. *Journal of the American Association for Laboratory Animal Science, 57*(2), 186-201.

Oliver, V., De Rantere, D., Ritchie, R., Chisholm, J., Hecker, K. G., & Pang, D. S. (2014). Psychometric assessment of the Rat Grimace Scale and development of an analgesic intervention score. *PLoS One, 9*(5), e97882. doi:10.1371/journal.pone.0097882

Ong, C. K. S., Lirk, P., Seymour, R. A., & Jenkins, B. J. (2005). The efficacy of preemptive analgesia for acute postoperative pain management: A meta-analysis. *Anesthesia and Analgesia, 100*(3), 757-773. doi:10.1213/01.Ane.0000144428.98767.0e

Osborne, N., Avey, M. T., Anestidou, L., Ritskes-Hoitinga, M., & Griffin, G. (2018). Improving animal research reporting standards: HARRP, the first step of a unified approach by ICLAS to improve animal research reporting standards worldwide. *EMBO Rep, 19*(5). doi:10.15252/embr.201846069

Osman, N., Adawi, D., Ahrne, S., Jeppsson, B., & Molin, G. (2004). Modulation of the effect of dextran sulfate sodium-induced acute colitis by the administration of different probiotic strains of Lactobacillus and Bifidobacterium. *Digestive Diseases and Sciences, 49*(2), 320-327. doi:10.1023/b:Ddas.0000017459.59088.43

Otis, C., Guillot, M., Moreau, M., Martel-Pelletier, J., Pelletier, J. P., Beaudry, F., & Troncy, E. (2017). Spinal neuropeptide modulation, functional assessment and cartilage lesions in a monosodium iodoacetate rat model of osteoarthritis. *Neuropeptides, 65*, 56-62. doi:10.1016/j.npep.2017.04.009

Pacharinsak, C., & Beitz, A. (2008). Animal models of cancer pain. *Comparative Medicine, 58*(3), 220-233.

Palmen, M., Dieleman, L. A., Vanderende, M. B., Uyterlinde, A., Pena, A. S., Meuwissen, S. G. M., & Vanrees, E. P. (1995). NONLYMPHOID AND LYMPHOID-CELLS IN ACUTE, CHRONIC AND RELAPSING EXPERIMENTAL COLITIS. *Clinical and Experimental Immunology, 99*(2), 226-232.

Perez, J., Ware, M. A., Chevalier, S., Gougeon, R., & Shir, Y. (2005). Dietary omega-3 fatty acids may be associated with increased neuropathic pain in nerve-injured rats. *Anesth Analg, 101*(2), 444-448, table of contents. doi:10.1213/01.ANE.0000158469.11775.52

Pham, T. M., Hagman, B., Codita, A., Van Loo, P. L., Strommer, L., & Baumans, V. (2010). Housing environment influences the need for pain relief during post-operative recovery in mice. *Physiol Behav, 99*(5), 663-668. doi:10.1016/j.physbeh.2010.01.038

Philips, B. H., Weisshaar, C. L., & Winkelstein, B. A. (2017). Use of the Rat Grimace Scale to Evaluate Neuropathic Pain in a Model of Cervical Radiculopathy. *Comparative Medicine, 67*(1), 34-42.

Pilcher, C. W. T., & Browne, J. (1982). ENVIRONMENTAL CROWDING MODIFIES RESPONDING TO NOXIOUS STIMULI AND THE EFFECTS OF MU-AGONISTS AND KAPPA-AGONISTS. *Life Sciences, 31*(12-1), 1213-1216. doi:10.1016/0024-3205(82)90345-9

Pitcher, G. M., & Henry, J. L. (2004). Nociceptive response to innocuous mechanical stimulation is mediated via myelinated afferents and NK-1 receptor activation in a rat model of neuropathic pain. *Exp Neurol, 186*(2), 173-197. doi:10.1016/j.expneurol.2003.10.019

Plint, A. C., Moher, D., Morrison, A., Schulz, K., Altman, D. G., Hill, C., & Gaboury, I. (2006). Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. *Medical Journal of Australia, 185*(5), 263-267.

Plone, M. A., Emerich, D. F., & Lindner, M. D. (1996). Individual differences in the hotplate test and effects of habituation on sensitivity to morphine. *Pain, 66*(2-3), 265-270. doi:10.1016/0304-3959(96)03048-5

Plunkett, J. A., Yu, C. G., Easton, J. M., Bethea, J. R., & Yezierski, R. P. (2001). Effects of interleukin-10 (IL-10) on pain behavior and gene expression following excitotoxic spinal cord injury in the rat. *Exp Neurol, 168*(1), 144-154. doi:10.1006/exnr.2000.7604

Prefontaine, L., Helie, P., & Vachon, P. (2015). Postoperative pain in Sprague Dawley rats after liver biopsy by laparotomy versus laparoscopy. *Lab Animal, 44*(5), 174-178. doi:10.1038/laban.731

Prkachin, K. M. (1992). THE CONSISTENCY OF FACIAL EXPRESSIONS OF PAIN - A COMPARISON ACROSS MODALITIES. *Pain, 51*(3), 297-306. doi:10.1016/0304-3959(92)90213-u

Puglisiallegra, S., & Oliverio, A. (1983). SOCIAL-ISOLATION - EFFECTS ON PAIN THRESHOLD AND STRESS-INDUCED ANALGESIA. *Pharmacology Biochemistry and Behavior, 19*(4), 679-681. doi:10.1016/0091-3057(83)90344-1

Qin, C., Malykhina, A. P., Akbarali, H. I., Greenwood-Van Meerveld, B., & Foreman, R. D. (2008). Acute colitis enhances responsiveness of lumbosacral spinal neurons to colorectal distension in rats. *Dig Dis Sci, 53*(1), 141-148. doi:10.1007/s10620-007-9835-z

Qu, C., King, T., Okun, A., Lai, J., Fields, H. L., & Porreca, F. (2011). Lesion of the rostral anterior cingulate cortex eliminates the aversiveness of spontaneous neuropathic pain following partial or complete axotomy. *Pain, 152*(7), 1641-1648. doi:10.1016/j.pain.2011.03.002

Radhakrishnan, R., Moore, S. A., & Sluka, K. A. (2003). Unilateral carrageenan injection into muscle or joint induces chronic bilateral hyperalgesia in rats. *Pain, 104*(3), 567-577. doi:10.1016/s0304-3959(03)00114-3

Raekallio, M., Heinonen, K. M., Kuussaari, J., & Vainio, O. (2003). Pain Alleviation in Animals: Attitudes and Practices of Finnish Veterinarians. *The Veterinary Journal, 165*(2), 131-135. doi:10.1016/s1090-0233(02)00186-7

Randhawa, P. K., Singh, K., Singh, N., & Jaggi, A. S. (2014). A review on chemical-induced inflammatory bowel disease models in rodents. *Korean J Physiol Pharmacol, 18*(4), 279-288. doi:10.4196/kjpp.2014.18.4.279

Reichlin, T. S., Vogt, L., & Wurbel, H. (2016). The Researchers' View of Scientific Rigor-Survey on the Conduct and Reporting of In Vivo Research. *PLoS One, 11*(12), 20. doi:10.1371/journal.pone.0165999

Reijgwart, M. L., Schoemaker, N. J., Pascuzzo, R., Leach, M. C., Stodel, M., de Nies, L., . . . van Zeeland, Y. R. A. (2017). The composition and initial evaluation of a grimace scale in ferrets after surgical implantation of a telemetry probe. *PLoS One, 12*(11), e0187986. doi:10.1371/journal.pone.0187986

Remeniuk, B., Sukhtankar, D., Okun, A., Navratilova, E., Xie, J. Y., King, T., & Porreca, F. (2015). Behavioral and neurochemical analysis of ongoing bone cancer pain in rats. *Pain, 156*(10), 1864-1873. doi:10.1097/j.pain.0000000000000218

Rialland, P., Authier, S., Guillot, M., del Castillo, J. R. E., Veilleux-Lemieux, D., Frank, D., . . . Troncy, E. (2012). Validation of Orthopedic Postoperative Pain Assessment Methods for Dogs: A Prospective, Blinded, Randomized, Placebo-Controlled Study. *PLoS One, 7*(11). doi:10.1371/journal.pone.0049480

Rice, A. S. C., Cimino-Brown, D., Eisenach, J. C., Kontinen, V. K., Lacroix-Fralish, M. L., Machin, I., . . . Preclinical Pain, C. (2008). Animal models and the prediction of efficacy in clinical trials of analgesic drugs: A critical appraisal and call for uniform reporting standards. *Pain, 139*(2), 243-247. doi:10.1016/j.pain.2008.08.017

Rice, A. S. C., Finnerup, N. B., Kemp, H. I., Currie, G. L., & Baron, R. (2018). Sensory profiling in animal models of neuropathic pain: a call for back-translation. *Pain, 159*(5), 819-824. doi:10.1097/j.pain.0000000000001138

Rice, A. S. C., Morland, R., Huang, W., Currie, G. L., Sena, E. S., & Macleod, M. R. (2013). Transparency in the reporting of in vivo pre-clinical pain research: The relevance and implications of the ARRIVE (Animal Research: Reporting In Vivo Experiments) guidelines. *Scand J Pain, 4*(2), 58-62. doi:10.1016/j.sjpain.2013.02.002

Rock, M. L., Karas, A. Z., Rodriguez, K. B. G., Gallo, M. S., Pritchett-Corning, K., Karas, R. H., . . . Gaskill, B. N. (2014). The Time-to-Integrate-to-Nest Test as an Indicator of Wellbeing in Laboratory Mice. *Journal of the American Association for Laboratory Animal Science, 53*(1), 24-28.

Roper, T. J. (1973). NESTING MATERIAL AS A REINFORCER FOR FEMALE MICE. *Animal Behaviour, 21*(NOV), 733-740. doi:10.1016/s0003-3472(73)80099-5

Ross-Huot, M. C., Laferriere, A., Gi, C. M., Khorashadi, M., Schricker, T., & Coderre, T. J. (2011). Effects of glycemic regulation on chronic postischemia pain. *Anesthesiology, 115*(3), 614-625. doi:10.1097/ALN.0b013e31822a63c9

Roughan, J. V., & Flecknell, P. A. (2000). Effects of surgery and analgesic administration on spontaneous behaviour in singly housed rats. *Res Vet Sci, 69*(3), 283-288. doi:10.1053/rvsc.2000.0430

Roughan, J. V., & Flecknell, P. A. (2001). Behavioural effects of laparotomy and analgesic effects of ketoprofen and carprofen in rats. *Pain, 90*(1-2), 65-74. doi:10.1016/s0304-3959(00)00387-0

Roughan, J. V., & Flecknell, P. A. (2002). Buprenorphine: a reappraisal of its antinociceptive effects and therapeutic use in alleviating post-operative pain in animals. *Laboratory Animals, 36*(3), 322-343. doi:10.1258/002367702320162423

Roughan, J. V., & Flecknell, P. A. (2003). Evaluation of a short duration behaviour-based post-operative pain scoring system in rats. *European Journal of Pain, 7*(5), 397-406. doi:10.1016/s1090-3801(02)00140-4

Roughan, J. V., & Flecknell, P. A. (2006). Training in behaviour-based post-operative pain scoring in rats—An evaluation based on improved recognition of analgesic requirements. *Applied Animal Behaviour Science, 96*(3-4), 327-342. doi:10.1016/j.applanim.2005.06.012

Roughan, J. V., & Flecknell, R. A. (2004). Behaviour-based assessment of the duration of laparotomy-induced abdominal pain and the analgesic effects of carprofen and buprenorphine in rats. *Behavioural Pharmacology, 15*(7), 461-472. doi:10.1097/00008877-200411000-00002

Roughan, J. V., Bertrand, H. G. M. J., & Isles, H. M. (2016). Meloxicam prevents COX-2-mediated post-surgical inflammation but not pain following laparotomy in mice. *European Journal of Pain, 20*(2), 231-240. doi:10.1002/ejp.712

Roughan, J. V., Coulter, C. A., Flecknell, P. A., Thomas, H. D., & Sufka, K. J. (2014). The Conditioned Place Preference Test for Assessing Welfare Consequences and Potential Refinements in a Mouse Bladder Cancer Model. *PLoS One, 9*(8). doi:10.1371/journal.pone.0103362

Roughan, J. V., Flecknell, P. A., & Davies, B. R. (2004). Behavioural assessment of the effects of tumour growth in rats and the influence of the analgesics carprofen and meloxicam. *Laboratory Animals, 38*(3), 286-296. doi:10.1258/002367704323133673

Rutten, K., Robens, A., Read, S. J., & Christoph, T. (2014). Pharmacological validation of a refined burrowing paradigm for prediction of analgesic efficacy in a rat model of sub-chronic knee joint inflammation. *European Journal of Pain, 18*(2), 213-222. doi:10.1002/j.1532-2149.2013.00359.x

Rutten, K., Schiene, K., Robens, A., Leipelt, A., Pasqualon, T., Read, S. J., & Christoph, T. (2014). Burrowing as a non-reflex behavioural readout for analgesic action in a rat model of sub-chronic knee joint inflammation. *European Journal of Pain, 18*(2), 204-212. doi:10.1002/j.1532-2149.2013.00358.x

Saine, L., Helie, P., & Vachon, P. (2016). Effects of fentanyl on pain and motor behaviors following a collagenase-induced intracerebral hemorrhage in rats. *J Pain Res, 9*, 1039-1048. doi:10.2147/JPR.S121415

Sariyanidi, E., Gunes, H., & Cavallaro, A. (2015). Automatic Analysis of Facial Affect: A Survey of Registration, Representation, and Recognition. *IEEE Trans Pattern Anal Mach Intell, 37*(6), 1113-1133. doi:10.1109/TPAMI.2014.2366127

Sarkis-Onofre, R., Poletto-Neto, V., Cenci, M. S., Pereira-Cenci, T., & Moher, D. (2017). Impact of the CONSORT Statement endorsement in the completeness of reporting of randomized clinical trials in restorative dentistry. *J Dent, 58*, 54-59. doi:10.1016/j.jdent.2017.01.009

Satoh, H., Morimoto, Y., Arai, T., Asanuma, H., Kawauchi, S., Seguchi, K., . . . Murai, M. (2007). Intravesical ultrasonography for tumor staging in an orthotopically implanted rat model of bladder cancer. *J Urol, 177*(3), 1169-1173. doi:10.1016/j.juro.2006.10.038

Sayette, M. A., Cohn, J. F., Wertz, J. M., Perrott, M. A., & Parrott, D. J. (2001). A psychometric evaluation of the facial action coding system for assessing spontaneous expression. *Journal of Nonverbal Behavior, 25*(3), 167-185. doi:10.1023/a:1010671109788

Schaap, M. W. H., Uilenreef, J. J., Mitsogiannis, M. D., van 't Klooster, J. G., Arndt, S. S., & Hellebrekers, L. J. (2012). Optimizing the dosing interval of buprenorphine in a multimodal postoperative analgesic strategy in the rat: minimizing side-effects without affecting weight gain and food intake. *Laboratory Animals, 46*(4), 287-292. doi:10.1258/la.2012.012058

Schneider, L. E., Henley, K. Y., Turner, O. A., Pat, B., Niedzielko, T. L., & Floyd, C. L. (2017). Application of the Rat Grimace Scale as a Marker of Supraspinal Pain Sensation after Cervical Spinal Cord Injury. *J Neurotrauma, 34*(21), 2982-2993. doi:10.1089/neu.2016.4665

Schulz, K. F., Altman, D. G., Moher, D., & Group, C. (2010). CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ, 340*, c332. doi:10.1136/bmj.c332

Schwarting, R. K. W. (2018). Ultrasonic vocalization in juvenile and adult male rats: A comparison among stocks. *Physiol Behav, 191*, 1-11. doi:10.1016/j.physbeh.2018.03.023

Schwarz, F., Iglhaut, G., & Becker, J. (2012). Quality assessment of reporting of animal studies on pathogenesis and treatment of peri-implant mucositis and peri-implantitis. A systematic review using the ARRIVE guidelines. *Journal of Clinical Periodontology, 39*, 63-72. doi:10.1111/j.1600-051X.2011.01838.x

Sena, E. S., van der Worp, H. B., Bath, P. M., Howells, D. W., & Macleod, M. R. (2010). Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. *PLoS Biol, 8*(3), e1000344. doi:10.1371/journal.pbio.1000344

Senko, T., Olexova, L., Mokosakova, M., & Krskova, L. (2017). Angiotensin II enhancement during pregnancy influences the emotionality of rat offspring (Rattus norvegicus) in adulthood. Potential use of the Rat Grimace Scale. *Neuroendocrinology Letters, 38*(2), 117-123.

Sharkey, K. A. (2006). Visceral sensation and colitis: inflammation and hypersensitivity do not always go hand in hand. *Neurogastroenterol Motil, 18*(2), 87-90. doi:10.1111/j.1365-2982.2005.00743.x

Sherwin, C. M., Haug, E., Terkelsen, N., & Vadgama, M. (2004). Studies on the motivation for burrowing by laboratory mice. *Applied Animal Behaviour Science, 88*(3-4), 343-358. doi:10.1016/j.applanim.2004.03.009

Shi, X. Z., Winston, J. H., & Sarna, S. K. (2011). Differential immune and genetic responses in rat models of Crohn's colitis and ulcerative colitis. *Am J Physiol Gastrointest Liver Physiol, 300*(1), G41-51. doi:10.1152/ajpgi.00358.2010

Shineman, D. W., Basi, G. S., Bizon, J. L., Colton, C. A., Greenberg, B. D., Hollister, B. A., . . . Fillit, H. M. (2011). Accelerating drug discovery for Alzheimer's disease: best practices for preclinical animal studies. *Alzheimers Research & Therapy, 3*(5). doi:10.1186/alzrt90

Silva, M. A. D., Pimenta, C. A. D., & da Cruz, D. A. M. (2013). Pain assessment and training: the impact on pain control after cardiac surgery. *Revista Da Escola De Enfermagem Da Usp, 47*(1), 83-91.

Silvoniemi, M., Vasankari, T., Vahlberg, T., Vuorinen, E., Clemens, K. E., & Salminen, E. (2012). Physicians' self-assessment of cancer pain treatment skills--more training required. *Support Care Cancer, 20*(11), 2747-2753. doi:10.1007/s00520-012-1396-9

Smith, A. J., Clutton, R. E., Lilley, E., Hansen, K. E. A., & Brattelid, T. (2018). PREPARE: guidelines for planning animal research and testing. *Lab Anim, 52*(2), 135-141. doi:10.1177/0023677217724823

Smith, J. A., Birke, L., & Sadler, D. (1997). Reporting animal use in scientific papers. *Laboratory Animals, 31*(4), 312-317. doi:10.1258/002367797780596176

Smith, S. B., Crager, S. E., & Mogil, J. S. (2004). Paclitaxel-induced neuropathic hypersensitivity in mice: responses in 10 inbred mouse strains. *Life Sci, 74*(21), 2593-2604. doi:10.1016/j.lfs.2004.01.002

Solomon, P. E., Prkachin, K. M., & Farewell, V. (1997). Enhancing sensitivity to facial expression of pain. *Pain, 71*(3), 279-284. doi:10.1016/s0304-3959(97)03377-0

Sorge, R. E., LaCroix-Fralish, M. L., Tuttle, A. H., Sotocinal, S. G., Austin, J. S., Ritchie, J., . . . Mogil, J. S. (2011). Spinal cord Toll-like receptor 4 mediates inflammatory and neuropathic hypersensitivity in male but not female mice. *J Neurosci, 31*(43), 15450-15454. doi:10.1523/JNEUROSCI.3859-11.2011

Sorge, R. E., Mapplebeck, J. C., Rosen, S., Beggs, S., Taves, S., Alexander, J. K., . . . Mogil, J. S. (2015). Different immune cells mediate mechanical pain hypersensitivity in male and female mice. *Nat Neurosci, 18*(8), 1081-1083. doi:10.1038/nn.4053

Sorge, R. E., Martin, L. J., Isbester, K. A., Sotocinal, S. G., Rosen, S., Tuttle, A. H., . . . Mogil, J. S. (2014). Olfactory exposure to males, including men, causes stress and related analgesia in rodents. *Nature Methods, 11*(6), 629-+. doi:10.1038/nmeth.2935

Sotocinal, S. G., Sorge, R. E., Zaloum, A., Tuttle, A. H., Martin, L. J., Wieskopf, J. S., . . . Mogil, J. S. (2011). The Rat Grimace Scale: A partially automated method for quantifying pain in the laboratory rat via facial expressions. *Molecular Pain, 7*. doi:10.1186/1744-8069-7-55

Sperry, M. M., Yu, Y. H., Welch, R. L., Granquist, E. J., & Winkelstein, B. A. (2018). Grading facial expression is a sensitive means to detect grimace differences in orofacial pain in a rat model. *Scientific Reports, 8*. doi:10.1038/s41598-018-32297-2

Stevanovic, A., Schmitz, S., Rossaint, R., Schurholz, T., & Coburn, M. (2015). CONSORT item reporting quality in the top ten ranked journals of critical care medicine in 2011: a retrospective analysis. *PLoS One, 10*(5), e0128061. doi:10.1371/journal.pone.0128061

Stevens, A., Shamseer, L., Weinstein, E., Yazdi, F., Turner, L., Thielman, J., . . . Moher, D. (2014). Relation of completeness of reporting of health research to journals' endorsement of reporting guidelines: systematic review. *Bmj-British Medical Journal, 348*. doi:10.1136/bmj.g3804

Stokes, E. L., Flecknell, P. A., & Richardson, C. A. (2009). Reported analgesic and anaesthetic administration to rodents undergoing experimental surgical procedures. *Laboratory Animals, 43*(2), 149-154. doi:10.1258/la.2008.008020

Streiner DL and Norman GR. 2008. Reliability. In: Streiner DL and Normal GR, eds. Health Measurement Scales. New York: Oxford University Press, 167–210.

Stryjek, R., Modlinska, K., & Pisula, W. (2012). Species specific behavioural patterns (digging and swimming) and reaction to novel objects in wild type, Wistar, Sprague-Dawley and Brown Norway rats. *PLoS One, 7*(7), e40642. doi:10.1371/journal.pone.0040642

Stucchi, A. F., Shofer, S., Leeman, S., Materne, O., Beer, E., McClung, J., . . . Becker, J. M. (2000). NK-1 antagonist reduces colonic inflammation and oxidative stress in dextran sulfate-induced colitis in rats. *American Journal of Physiology-Gastrointestinal and Liver Physiology, 279*(6), G1298-G1306.

Studlack, P. E., Keledjian, K., Farooq, T., Akintola, T., Gerzanich, V., Simard, J. M., & Keller, A. (2018). Blast-induced brain injury in rats leads to transient vestibulomotor deficits and persistent orofacial pain. *Brain Injury, 32*(13-14), 1866-1878. doi:10.1080/02699052.2018.1536282

Sufka, K. J. (1994). Conditioned place preference paradigm - a novel-approach for analgesic drug assessment against chronic pain. *Pain, 58*(3), 355-366. doi:10.1016/0304-3959(94)90130-9

Sun, P., Zhou, K., Wang, S., Li, P., Chen, S., Lin, G., . . . Wang, T. (2013). Involvement of MAPK/NF-kappaB signaling in the activation of the cholinergic anti-inflammatory pathway in experimental colitis by chronic vagus nerve stimulation. *PLoS One, 8*(8), e69424. doi:10.1371/journal.pone.0069424

Tall, J. M. (2009). Housing supplementation decreases the magnitude of inflammation-induced nociception in rats. *Behav Brain Res, 197*(1), 230-233. doi:10.1016/j.bbr.2008.08.010

Tam, W. W., Lo, K. K., & Khalechelvam, P. (2017). Endorsement of PRISMA statement and quality of systematic reviews and meta-analyses published in nursing journals: a cross-sectional study. *BMJ Open, 7*(2), e013905. doi:10.1136/bmjopen-2016-013905

Thomas, A., Miller, A., Roughan, J., Malik, A., Haylor, K., Sandersen, C., . . . Leach, M. (2016). Efficacy of Intrathecal Morphine in a Model of Surgical Pain in Rats. *PLoS One, 11*(10), e0163909. doi:10.1371/journal.pone.0163909

Thomas, D. A., & Barfield, R. J. (1985). ULTRASONIC VOCALIZATION OF THE FEMALE RAT (RATTUS-NORVEGICUS) DURING MATING. *Animal Behaviour, 33*(AUG), 720-725. doi:10.1016/s0003-3472(85)80002-6

Ting, K. H. J., Hill, C. L., & Whittle, S. L. (2015). Quality of reporting of interventional animal studies in rheumatology: a systematic review using the ARRIVE guidelines. *International Journal of Rheumatic Diseases, 18*(5), 488-494. doi:10.1111/1756-185x.12699

Tjolsen, A., Rosland, J. H., Berge, O. G., & Hole, K. (1991). THE INCREASING-TEMPERATURE HOT-PLATE TEST - AN IMPROVED TEST OF NOCICEPTION IN MICE AND RATS. *Journal of Pharmacological Methods, 25*(3), 241-250. doi:10.1016/0160-5402(91)90014-v

Tobin, J. M., Delbridge, L. M. D., Di Nicolantonio, R., & Bhathal, P. (2004). Development of colorectal sensitization is associated with increased eosinophils and mast cells in dextran sulfate sodium-treated rats. *Digestive Diseases and Sciences, 49*(7-8), 1302-1310. doi:10.1023/B:DDAS.0000037827.07367.2d

Togawa, J. I., Nagase, H., Tanaka, K., Inamori, M., Nakajima, A., Ueno, N., . . . Sekihara, H. (2002). Oral administration of lactoferrin reduces colitis in rats via modulation of the immune system and correction of cytokine imbalance. *Journal of Gastroenterology and Hepatology, 17*(12), 1291-1298. doi:10.1046/j.1440-1746.2002.02868.x

Totsch, S. K., Waite, M. E., Tomkovich, A., Quinn, T. L., Gower, B. A., & Sorge, R. E. (2016). Total Western Diet Alters Mechanical and Thermal Sensitivity and Prolongs Hypersensitivity Following Complete Freund's Adjuvant in Mice. *J Pain, 17*(1), 119-125. doi:10.1016/j.jpain.2015.10.006

Tuboly, G., Benedek, G., & Horvath, G. (2009). Selective disturbance of pain sensitivity after social isolation. *Physiol Behav, 96*(1), 18-22. doi:10.1016/j.physbeh.2008.07.030

Turner, L., Shamseer, L., Altman, D. G., Weeks, L., Peters, J., Kober, T., . . . Moher, D. (2012). Consolidated standards of reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals. *Cochrane Database of Systematic Reviews*(11). doi:10.1002/14651858.MR000030.pub2

Tuttle, A. H., Molinaro, M. J., Jethwa, J. F., Sotocinal, S. G., Prieto, J. C., Styner, M. A., . . . Zylka, M. J. (2018). A deep neural network to assess spontaneous pain from mouse facial expressions. *Mol Pain, 14*, 1744806918763658. doi:10.1177/1744806918763658

Van de Weerd, H. A., Van Loo, P. L. P., Van Zutphen, L. F. M., Koolhaas, J. M., & Baumans, V. (1998). Strength of preference for nesting material as environmental enrichment for laboratory mice. *Applied Animal Behaviour Science, 55*(3-4), 369-382. doi:10.1016/s0168-1591(97)00043-9

Van Loo, R. L. R., & Baumans, V. (2004). The importance of learning young: the use of nesting material in laboratory rats. *Laboratory Animals, 38*(1), 17-24. doi:10.1258/00236770460734353 van Luijk, J., Bakker, B., Rovers, M. M., Ritskes-Hoitinga, M., de Vries, R. B. M., & Leenaars, M. (2014). Systematic Reviews of Animal Studies; Missing Link in Translational Research? *PLoS One, 9*(3), 5. doi:10.1371/journal.pone.0089981

Veigas, J. M., Williams, P. J., Halade, G., Rahman, M. M., Yoneda, T., & Fernandes, G. (2011). Fish oil concentrate delays sensitivity to thermal nociception in mice. *Pharmacol Res, 63*(5), 377-382. doi:10.1016/j.phrs.2011.02.004

Verma-Gandhu, M., Verdu, E. F., Bercik, P., Blennerhassett, P. A., Al-Mutawaly, N., Ghia, J. E., & Collins, S. M. (2007). Visceral pain perception is determined by the duration of colitis and associated neuropeptide expression in the mouse. *Gut, 56*(3), 358-364. doi:10.1136/gut.2006.100016

Vesterinen, H. M., Sena, E. S., Ffrench-Constant, C., Williams, A., Chandran, S., & Macleod, M. R. (2010). Improving the translational hit of experimental treatments in multiple sclerosis. *Multiple Sclerosis Journal, 16*(9), 1044-1055. doi:10.1177/1352458510379612

Vetuschi, A., Latella, G., Sferra, R., Caprilli, R., & Gaudio, E. (2002). Increased proliferation and apoptosis of colonic epithelial cells in dextran sulfate sodium-induced colitis in rats. *Digestive Diseases and Sciences, 47*(7), 1447-1457. doi:10.1023/a:1015931128583

Viscardi, A. V., Hunniford, M., Lawlis, P., Leach, M., & Turner, P. V. (2017). 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. doi:10.3389/fvets.2017.00051

Vitkova, J., Loucka, M., Bocek, J., & Vaculin, S. (2015). The effect of acclimatization and ambient temperature on heat withdrawal threshold in rats. *Eur J Pain, 19*(1), 21-27. doi:10.1002/ejp.515

Vivian, J. A., & Miczek, K. A. (1993). Morphine attenuates ultrasonic vocalization during agonistic encounters in adult male-rats. *Psychopharmacology, 111*(3), 367-375. doi:10.1007/bf02244954

Vogt, L., Reichlin, T. S., Nathues, C., & Wuerbel, H. (2016). Authorization of Animal Experiments Is Based on Confidence Rather than Evidence of Scientific Rigor. *Plos Biology, 14*(12). doi:10.1371/journal.pbio.2000598

Vollert, J., Maier, C., Attal, N., Bennett, D. L. H., Bouhassira, D., Enax-Krumova, E. K., . . . Baron, R. (2017). Stratifying patients with peripheral neuropathic pain based on sensory profiles: algorithm and sample size recommendations. *Pain, 158*(8), 1446-1455. doi:10.1097/j.pain.0000000000000935

Vos, B. P., Hans, G., & Adriaensen, H. (1998). Behavioral assessment of facial pain in rats: face grooming patterns after painful and non-painful sensory disturbances in the territory of the rat's infraorbital nerve. *Pain, 76*(1-2), 173-178.

Vos, B., Strassman, A., & Maciewicz, R. (1994). Behavioral evidence of trigeminal neuropathic pain following chronic constriction injury to the rat's infraorbital nerve. *14*(5), 2708-2723. doi:10.1523/JNEUROSCI.14-05-02708.1994 %J The Journal of Neuroscience

Vowinkel, T., Kalogeris, K. J., Mori, M., Krieglstein, C. F., & Granger, D. N. (2004). Impact of dextran sulphate sodium load on the severity of inflammation in experimental colitis. *Faseb Journal, 18*(5), A1265-A1265.

Waite, M. E., Tomkovich, A., Quinn, T. L., Schumann, A. P., Dewberry, L. S., Totsch, S. K., & Sorge, R. E. (2015). Efficacy of Common Analgesics for Postsurgical Pain in Rats. *Journal of the American Association for Laboratory Animal Science, 54*(4), 420-425.

Wallace, V. C., Norbury, T. A., & Rice, A. S. (2005). Ultrasound vocalisation by rodents does not correlate with behavioural measures of persistent pain. *Eur J Pain, 9*(4), 445-452. doi:10.1016/j.ejpain.2004.10.006

Wallas, T. R., Winterson, B. J., Ransil, B. J., & Bove, G. M. (2003). Paw withdrawal thresholds and persistent hindlimb flexion in experimental mononeuropathies. *The Journal of Pain, 4*(4), 222-230. doi:10.1016/s1526-5900(03)00619-9

Whittaker, A. L., Leach, M. C., Preston, F. L., Lymn, K. A., & Howarth, G. S. (2016). Effects of acute chemotherapy-induced mucositis on spontaneous behaviour and the grimace scale in laboratory rats. *Lab Anim, 50*(2), 108-118. doi:10.1177/0023677215595554

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. *Laboratory Animals, 49*(1), 30-39. doi:10.1177/0023677214546913

Williams, A. C. D. (2002). Facial expression of pain: An evolutionary account. *Behavioral and Brain Sciences, 25*(4), 439-+. doi:10.1017/s0140525x02000080

Williams, V. M., Lascelles, B. D. X., & Robson, M. C. (2005). Current attitudes to, and use of, peri-operative analgesia in dogs and cats by veterinarians in New Zealand. *New Zealand Veterinary Journal, 53*(3), 193-202. doi:10.1080/00480169.2005.36504

Williams, W. O., Riskin, D. K., & Mott, K. M. (2008). Ultrasonic sound as an indicator of acute pain in laboratory mice. *Journal of the American Association for Laboratory Animal Science, 47*(1), 8-10.

Wodarski, R., Delaney, A., Ultenius, C., Morland, R., Andrews, N., Baastrup, C., . . . Rice, A. S. C. (2016). Cross-centre replication of suppressed burrowing behaviour as an ethologically relevant pain outcome measure in the rat: a prospective multicentre study. *Pain, 157*(10), 2350-2365. doi:10.1097/j.pain.0000000000000657

Woolf, C. J. (2010). What is this thing called pain? *J Clin Invest, 120*(11), 3742-3744. doi:10.1172/JCI45178

Woolfe, G., & Macdonald, A. D. (1944). The evaluation of the analgesic action of pethidine hydrochloride (demerol). *Journal of Pharmacology and Experimental Therapeutics, 80*(3), 300.

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*(2), 488-503. doi:10.1097/j.pain.0000000000000393

Xiao, Z., McCallum, T. J., Brown, K. M., Miller, G. G., Halls, S. B., Parney, I., & Moore, R. B. (1999). Characterization of a novel transplantable orthotopic rat bladder transitional cell tumour model. *British Journal of Cancer, 81*(4), 638-646. doi:10.1038/sj.bjc.6690741

Yamanaka, D., Kawano, T., Nishigaki, A., Aoyama, B., Tateiwa, H., Shigematsu-Locatelli, M., . . . Yokoyama, M. (2017). The preventive effects of dexmedetomidine on endotoxin-induced exacerbated post-incisional pain in rats. *J Anesth, 31*(5), 664-671. doi:10.1007/s00540-017-2374-7

Yang, J. P., Yao, M., Jiang, X. H., & Wang, L. N. (2006). Establishment of model of visceral pain due to colorectal distension and its behavioral assessment in rats. *World Journal of Gastroenterology, 12*(17), 2781-2784. doi:10.3748/wjg.v12.i17.2781

Yehuda, S., Leprohongreenwood, C. E., Dixon, L. M., & Coscina, D. V. (1986). Effects of dietary-fat on pain threshold, thermoregulation and motor-activity in rats. *Pharmacology Biochemistry and Behavior, 24*(6), 1775-1777.

Yeomans, D. C., Pirec, V., & Proudfit, H. K. (1996). Nociceptive responses to high and low rates of noxious cutaneous heating are mediated by different nociceptors in the rat: Behavioral evidence. *Pain, 68*(1), 133-140. doi:10.1016/s0304-3959(96)03176-4

Yousef, M. A. A., Dionigi, P., Marconi, S., Calligaro, A., Cornaglia, A. I., Alfonsi, E., & Auricchio, F. (2015). Successful Reconstruction of Nerve Defects Using Distraction Neurogenesis with a New Experimental Device. *Basic and Clinical Neuroscience, 6*(4), 253-264.

Zhang, E., Leung, V. & Pang, D. S. J. (2019). Influence of rater training on inter- and intrarater reliability when using the Rat Grimace Scale. *J Am Assoc Lab Anim Sci,* DOI**:** https://doi.org/10.30802/AALAS-JAALAS-18-000044

Zhang, N., Li, D., Shao, J., & Wang, X. (2015). Animal models for bladder cancer: The model establishment and evaluation (Review). *Oncol Lett, 9*(4), 1515-1519. doi:10.3892/ol.2015.2888

Zhou, Q., Price, D. D., Caudle, R. M., & Verne, G. N. (2008). Visceral and somatic hypersensitivity in TNBS-induced colitis in rats. *Dig Dis Sci, 53*(2), 429-435. doi:10.1007/s10620-007-9881-6

Zois, C. D., Katsanos, K. H., Kosmidou, M., & Tsianos, E. V. (2010). Neurologic manifestations in inflammatory bowel diseases: current knowledge and novel insights. *J Crohns Colitis, 4*(2), 115-124. doi:10.1016/j.crohns.2009.10.005

5. Appendices

5.1. Appendix A

5.1.1. Summary of all pain models used, and action units identified for the grimace scales of different species

	Action units							Pain model	Reference
Grimace Scale	Area above eyes	Eyes	Nose	Cheek	Ear	Whiskers	Mouth		
Mouse	-	Orbital tightening	Nose bulge	Cheek bulge	Ear position	Whisker change	-	Acetic acid abdominal constriction test	Langford *et al.* (2010)
Rat	-	Orbital tightening	Nose/cheek flattening		Ear changes	Whisker change	-	Intra plantar CFA Intra plantar Kaolin/Carrageenan Laparotomy	Sotocinal *et al*. (2011)
Rabbit	-	Orbital tightening	Pointed nose	Cheek flattening	~~Ear changes~~	~~Whisker change~~	-	Ear tattoo with or without EMLA cream	Keating *et al*. (2012)
Cat	-	-	Nose/muzzle shape	-	Ear position	-	-	Post-operative treatment for injuries	Holden *et al*. (2014)
Horse	Tension above the eye area	Orbital tightening	Strained nostrils and flattening of profile	Prominent strained chewing muscles	Stiffly backwards ears	-	Mouth strained and pronounced chin	Castration	Dalla Costa *et al*. (2014)
Sheep	-	Orbital tightening	Abnormal nostril and philtrum shape	Cheek tightening	Abnormal ear position	-	Abnormal lip and jaw profile	Foot rot Mastitis	McLennan *et al.* (2016)

Sheep		Orbital tightening	-	-	Ear and head position	-	Flehming	Osteotomy	Hager *et al.* (2017)
Lamb	-	Orbital tightening	Nose changes	Cheek flattening	Ear position	-	Mouth changes	Tail docking with rubber ring	Guesgen *et al.* (2016)
Piglet	Tension above the eyes Temporal tension Forehead profile	Orbital tightening	Snout plate changes Snout angle ~~Nostril dilation~~	Cheek tension	-	-	~~Upper lip contraction~~ ~~Lower jaw profile~~	Tail docking	Di Giminiani *et al.* (2016)
Piglet	-	Orbital tightening	Cheek tightening/nose bulge		Ear position	-	-	Tail docking and castration	Viscardi *et al.* (2017)
Ferret	-	Orbital tightening	Nose bulging	Cheek bulging	Ear changes	Whisker retraction	-	Telemetry probe	Reijgwart *et al* (2017)
Seal	-	Eye change	Nose change	-	-	Whisker change	Mouth change	Tagging and chipping of hind flipper	MacRae *et al.* (2018)

Actions units that are crossed out were identified but were not included in the study performed.

5.1.2. Summary of the different methods of validation for the grimace scales of different species

Grimace Scale	Reliability			Construct Validity		Criterion Validity	Accuracy of experienced rater	References
	Inter-rater	Intra-rater	Internal consistency	Increase with 'pain'	Decrease with analgesia			
Mouse	ICC = 0.90	-	α=0.89	Yes	Yes; Morphine	-	97%	Langford *et al*. (2010)
Rat	ICC = 0.90	ICC = 0.83	α= 0.84	Yes	Yes; Morphine	-	82%	Sotocinal *et al*. (2011) Oliver *et al*. (2014)
Rabbit	ICC = 0.91	-	-	Yes	Yes; EMLA	Yes; physiological measurements	84%	Keating *et al*. (2014)
Cat		-	-	Yes	-	Yes; cats classified based on numerical rating scale	98%	Holden *et al*. (2014)
Horse	ICC = 0.92	ICC = 0.85	-	Yes	No; flunixin-meglumine	Yes; composite pain scale	73%	Dalla Costa *et al*. (2014; 2016)
Sheep	ICC = 0.86	-	Each AU correlates with other AUs and total score	Yes	Yes; antibiotics and meloxicam	Yes; lameness and lesion score	67%	McLennan *et al.* (2016)
Sheep	ICC = 0.92	-		Yes	-	Yes; clinical severity score	68%	Hager *et al*. (2017)
Lamb	W = 0.60	-	-	Yes	-	-	-	Guesgen *et al*. (2016)
Piglet	ICC = 0.97	-	-	Only orbital tightening	-	No correlation	-	Di Giminiani *et al*. (2016)
Piglet	ICC = 0.57	ICC > 0.90	-	Yes	No; EMLA & meloxicam	Yes; active and inactive behaviours	-	Viscardi *et al*. (2017)

Ferret	ICC = 0.89	ICC = 0.67	-	Only orbital tightening	-	-	89% (only orbital tightening)	Reijgwart *et al* (2017)
Seal	-	-	-	Only orbital tightening	Yes; buprenorphine	Yes; increase orbital tightening coincides with decreased activity	95%	MacRae *et al*. (2018)

EMLA = Eutectic Mixture of Local Anesthetics

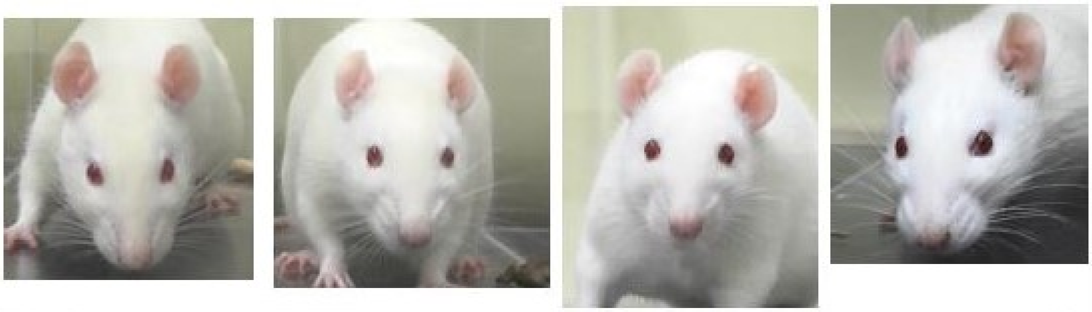
5.2. Appendix B – RGS training manual

To be used in conjunction with training manual from Sotocinal et al. 2011 Mol Pain 7:55

Eyes: 0

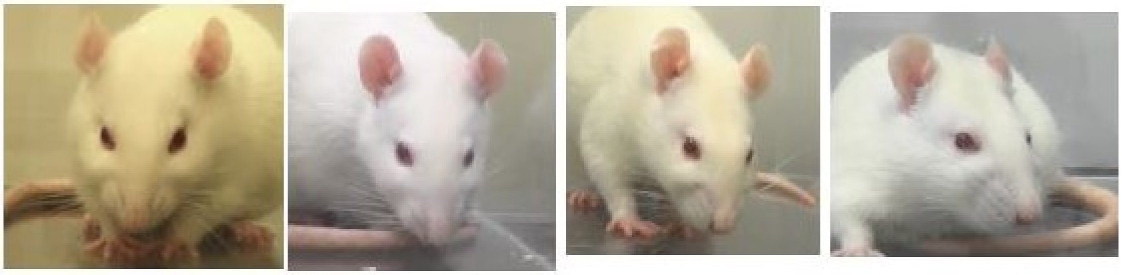
Sotocinal et al. 2011 – "Rats in pain display a narrowing of the orbital area, a tightly closed eyelid, or an eye squeeze. An eye squeeze is defined as the orbital muscles around the eyes being contracted. The nictitating membrane may be visible around the eye and becomes more pronounced as the pain intensifies. As a guideline, any eye closure that reduces the eye size by more than half should be coded as a "2" [...] Photographs of sleeping rats should not be taken and/or coded"

Pang lab – Eyes are completely bulgy, eye lids not showing



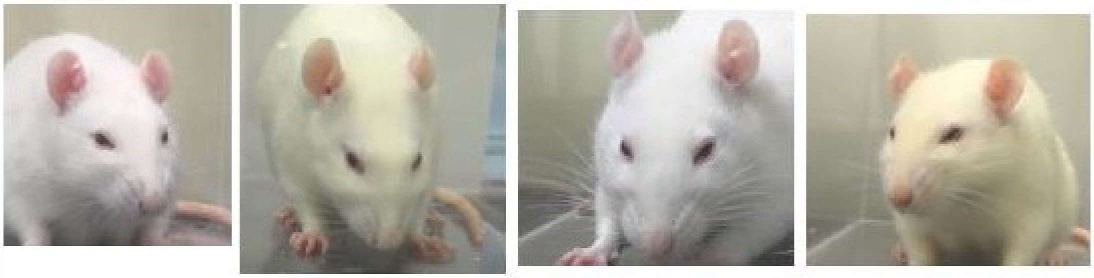
Eyes: 1

Pang lab – Eyes are slightly sunken into their head, tension in the eye lids, <50% closed



Eyes: 2

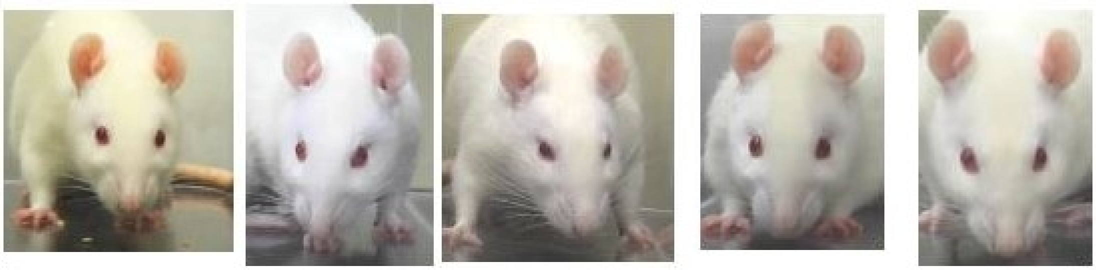
Pang lab – eyes are squeezed together, >50% closed



Ears: 0

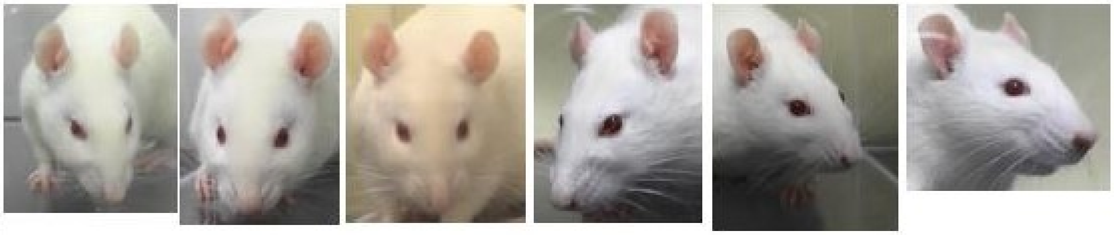
Sotocinal et al. 2011 – "The ears of rats in pain may be curled and pointed more than in the baseline position. In the baseline position ears are roughly perpendicular to the head, face forward, and are angled slightly backward. Importantly, the ears also have a rounded shape. In pain, the ears tend to fold, curl inwards and are angled forward. This curling of the ears tends to result in a 'pointed' shape of the ears. In pronounced pain states, the ears are angled outward and are held close to 45° away from both the perpendicular axis and the nose. As a result, the space between the ears may appear wider relative to baseline."

Pang lab – Ears are rounded, facing forwards and approximately perpendicular to the head.



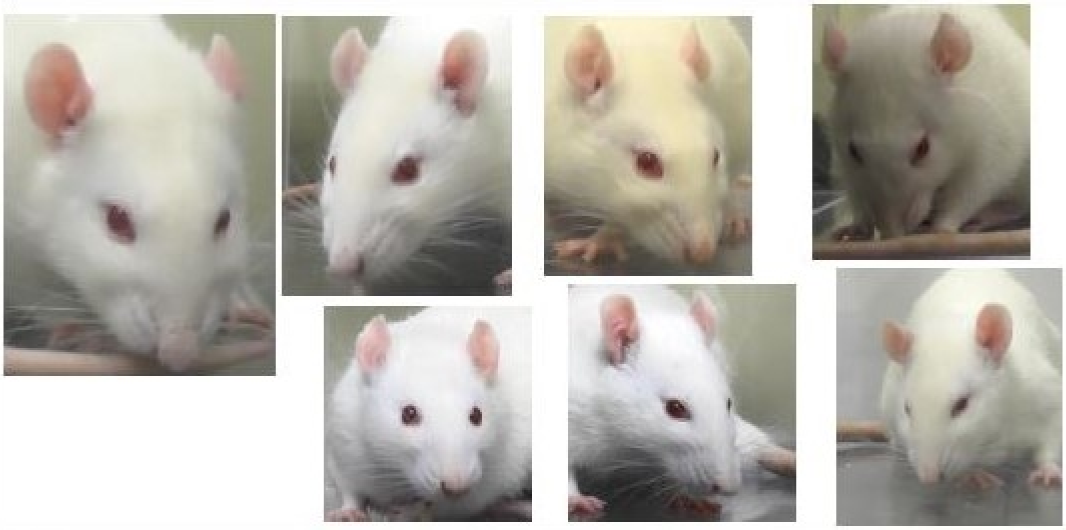
Ears: 1

Pang lab – Ears are slightly rotated outwards, a bit further apart, more slender (curled)



Ears: 2

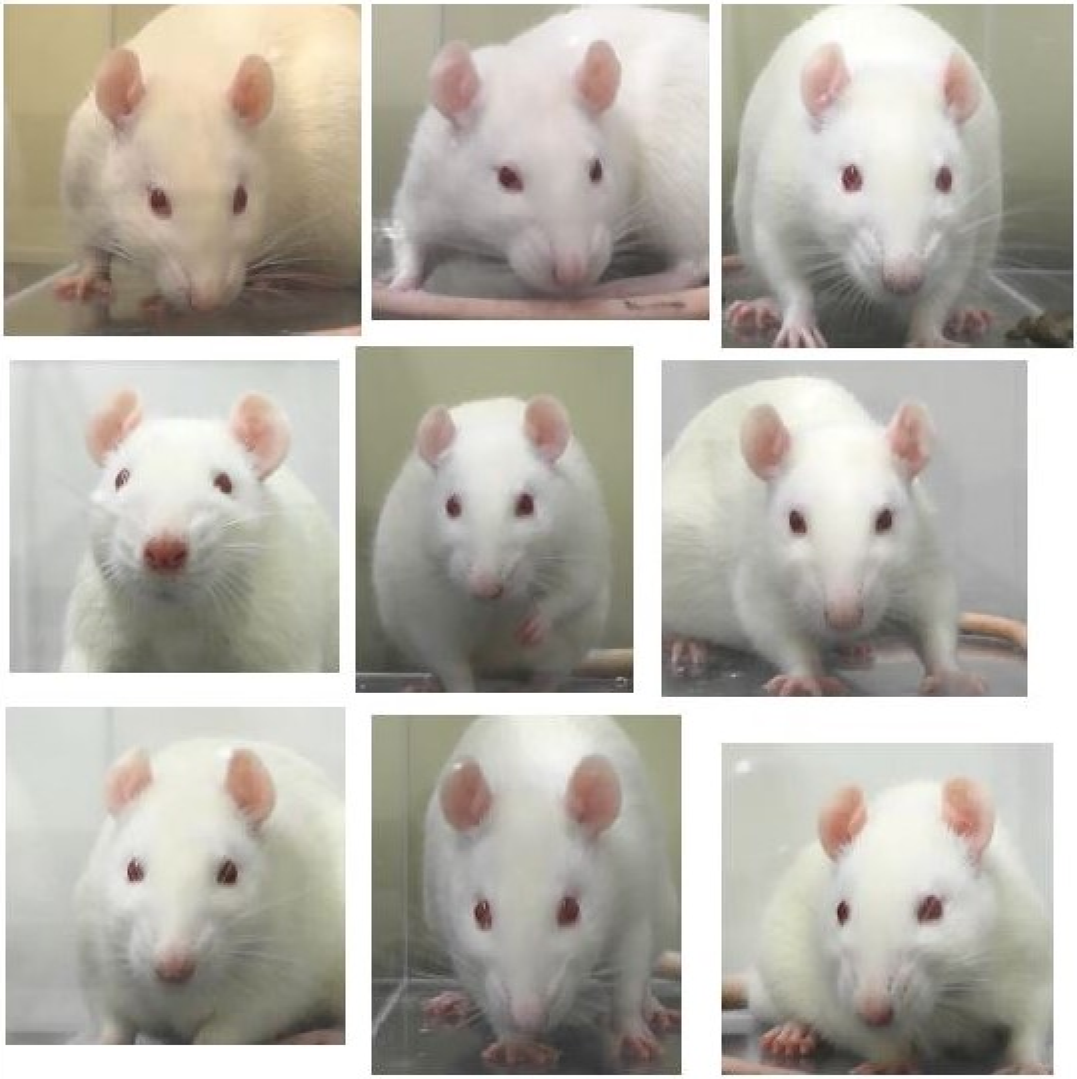
Pang lab – Ears are rotated outwards, base of ears tend to be further apart, and ear shape is usually curled or narrowed.



Nose/cheek: 0

Sotocinal et al. 2011 – "Rats in pain display a lack of bulge on top of the nose (i.e., a flattening of the nose). In the 'no pain' condition a clear bulge is present at the bridge of the nose. The whisker pads are also rounded and slightly puffed out, leaving a clear crease between the pads and the cheek. When in pain, the bridge of the nose flattens and elongates, causing the whisker pads to flatten. At this time the crease between the pads and the cheek is no longer present. In frontal headshots, the nose may appear narrower and longer"

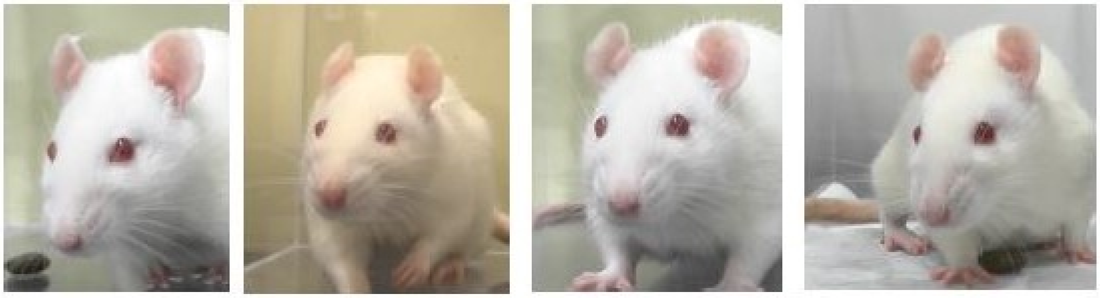
Pang lab – Whisker pad/muzzle is round, cheeks are round & relaxed, obvious crease between nose and cheek



Whiskers: 0

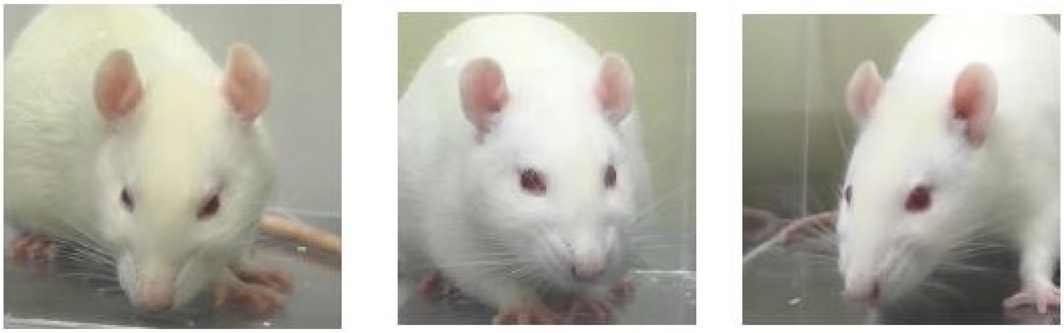
Sotocinal et al. 2011 – "Rats in pain have whiskers that have moved from the baseline position and orientation. Whiskers start relaxed and drooping slightly downward and, as pain progresses, tension in the pads increases and they become angled back along the head. In pain, the whisker pad is contracted causing the whiskers to bunch and be directed outwards away from the face. This gives the appearance of the whiskers as 'standing on end'. As follicles become tense, whiskers are closer together and are less distinct."

Pang lab – Whiskers are spread, relaxed, and floppy/droopy at the ends



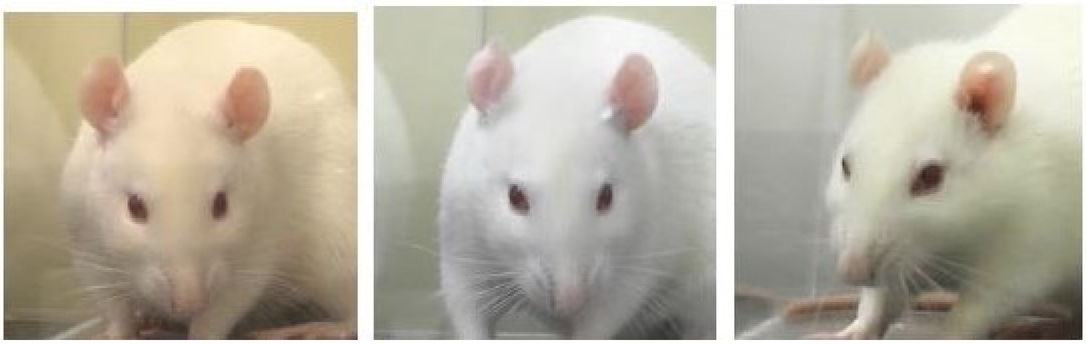
Whiskers: 1

Pang lab – Whiskers are straight at the ends and are pulled towards cheeks



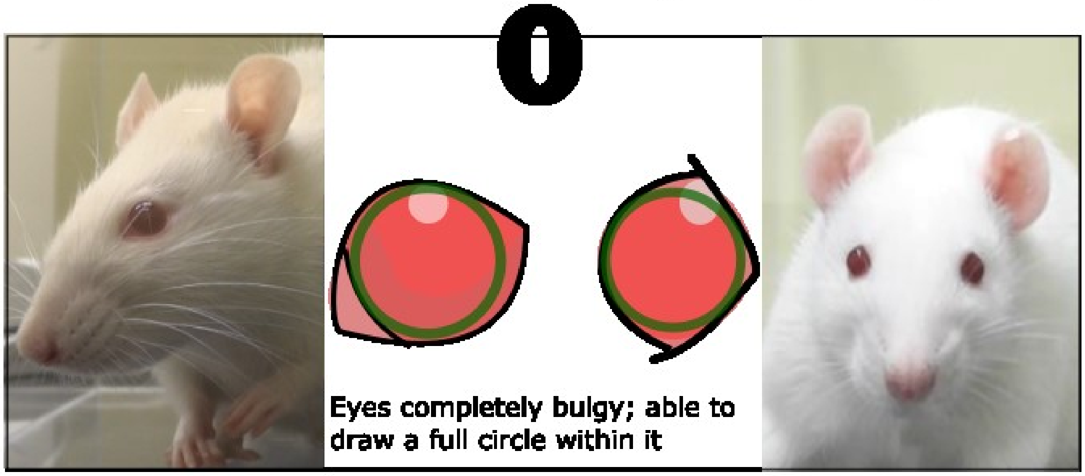
Whiskers: 2

Pang lab – Whiskers are straight at the ends and top whiskers are horizontal or pointed away from the cheeks



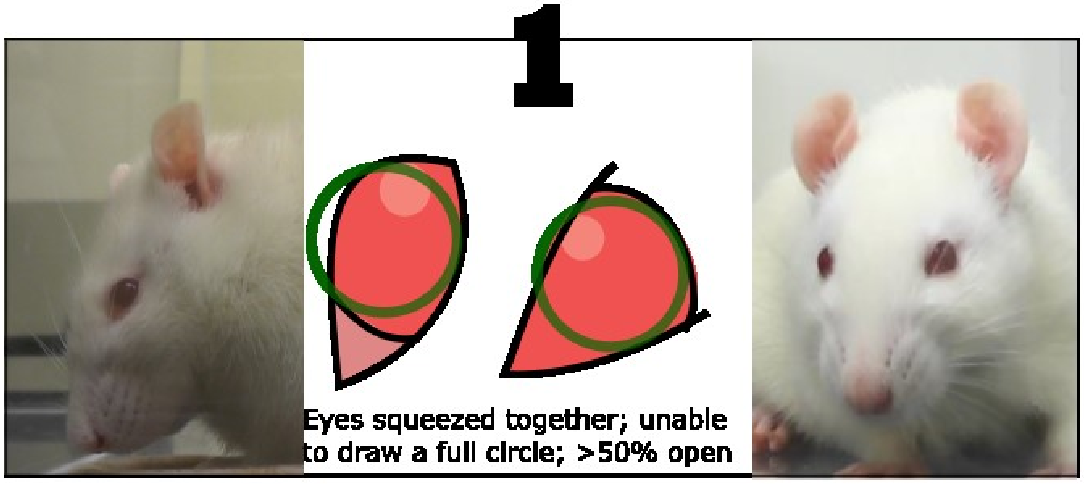
Rat Grimace Scale (RGS) - Eyes

0



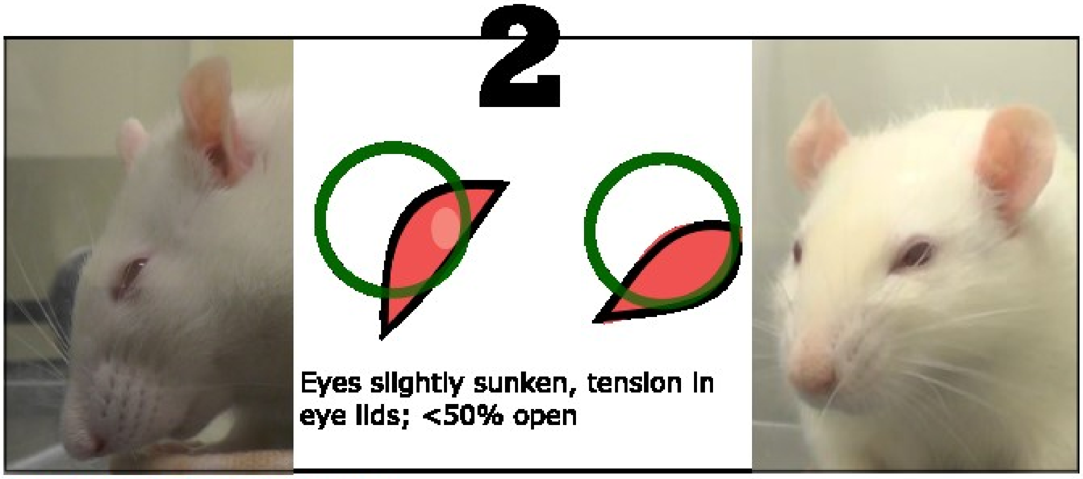
Eyes completely bulgy; able to draw a full circle within it

1



Eyes squeezed together; unable to draw a full circle; >50% open

2

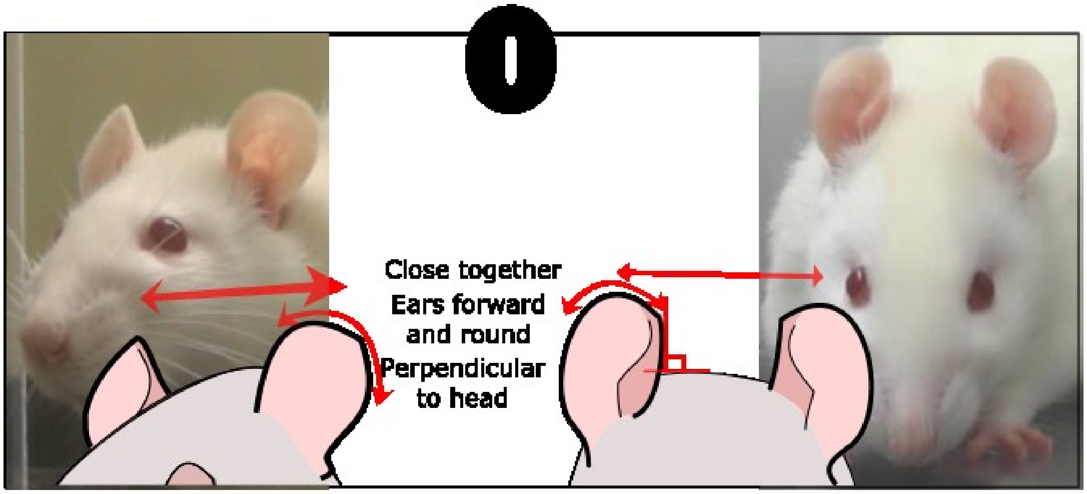


Eyes slightly sunken, tension in eye lids; <50% open

Do not use sleeping pictures

Rat Grimace Scale (RGS) - Ears

0

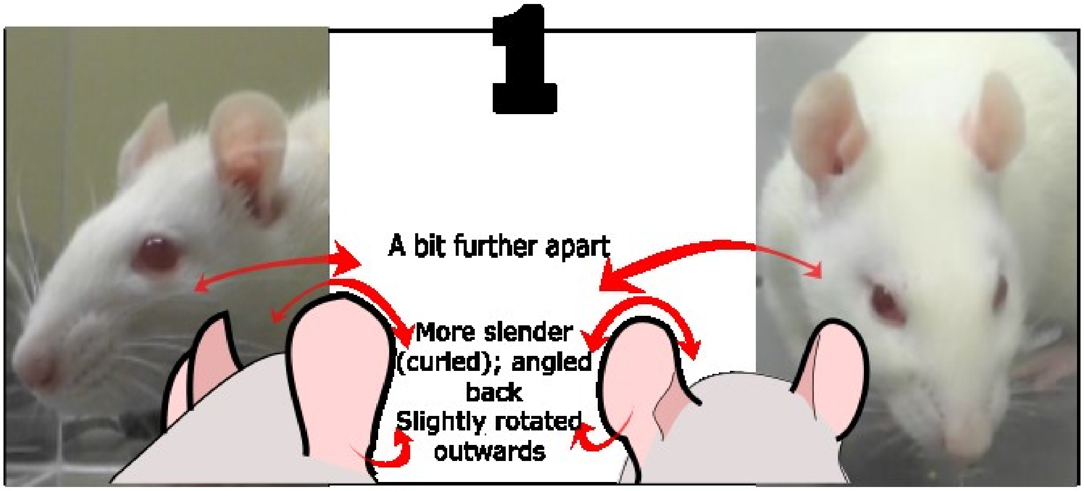


Close together

Ears forward and round

Perpendicular to head

1

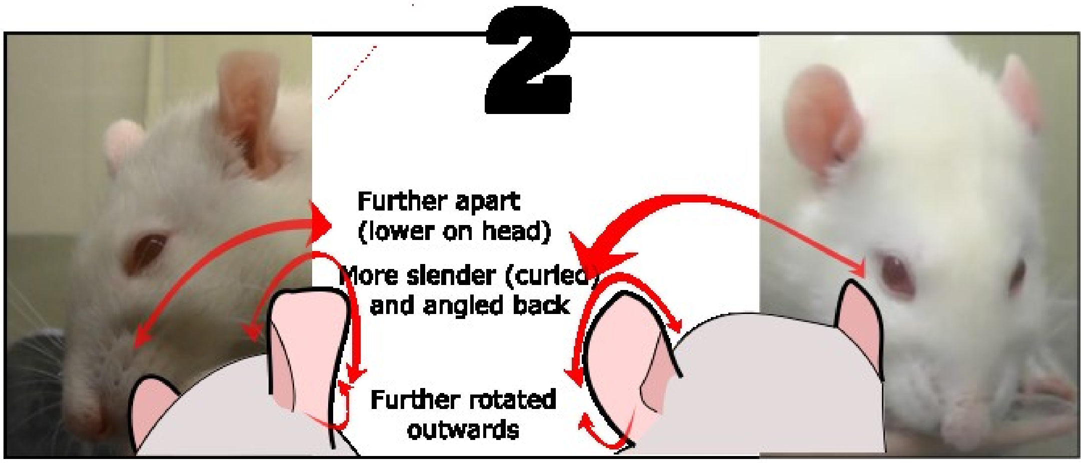


A bit further apart

More slender (curled); angled back

Slightly rotated outwards

2



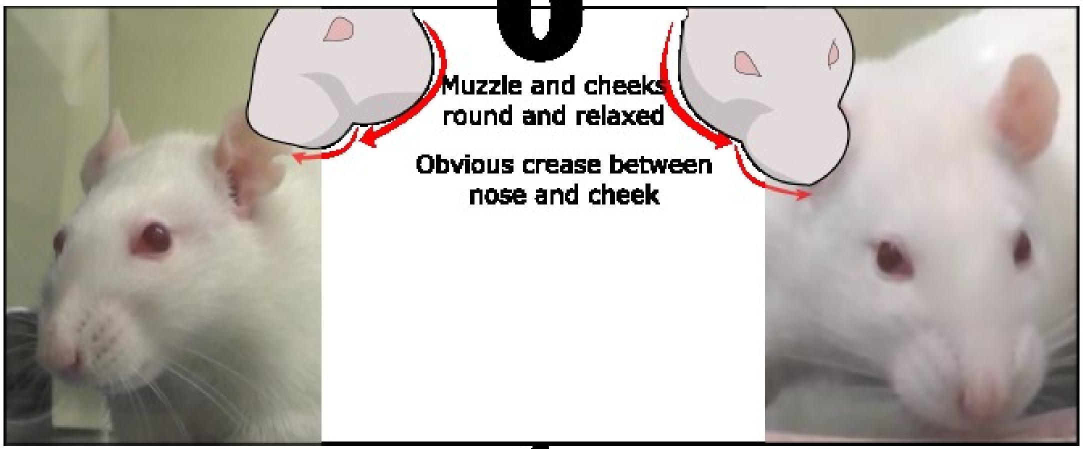
Further apart (lower on head)

More slender (curled) and angled back

Further rotated outwards

Rat Grimace Scale (RGS) - Nose/Cheek

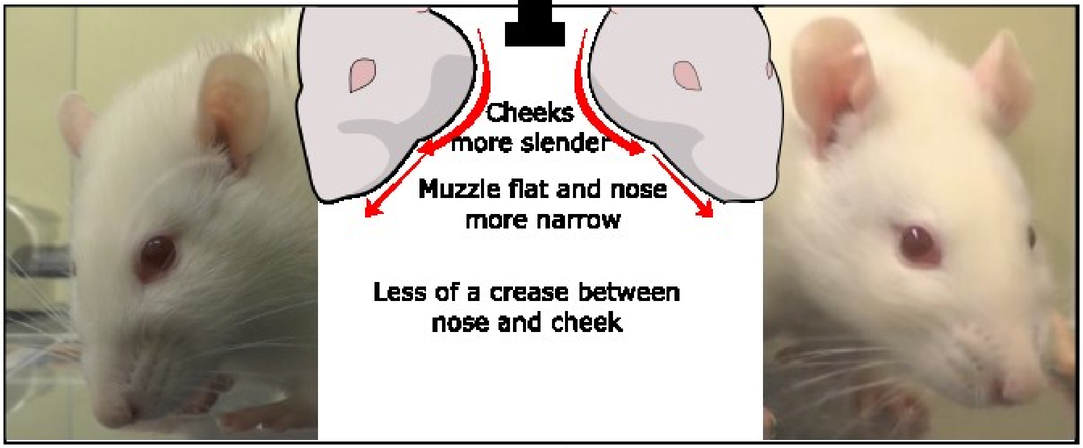
0



Muzzle and cheeks round and relaxed

Obvious crease between nose and cheek

1

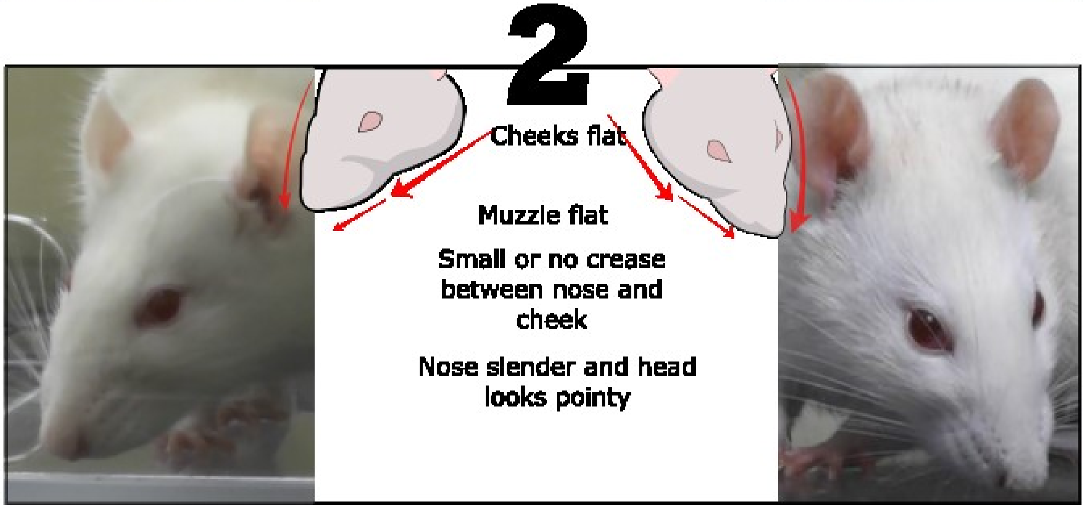


Cheeks more slender

Muzzle flat and nose more narrow

Less of a crease between nose and cheek

2



Cheeks flat

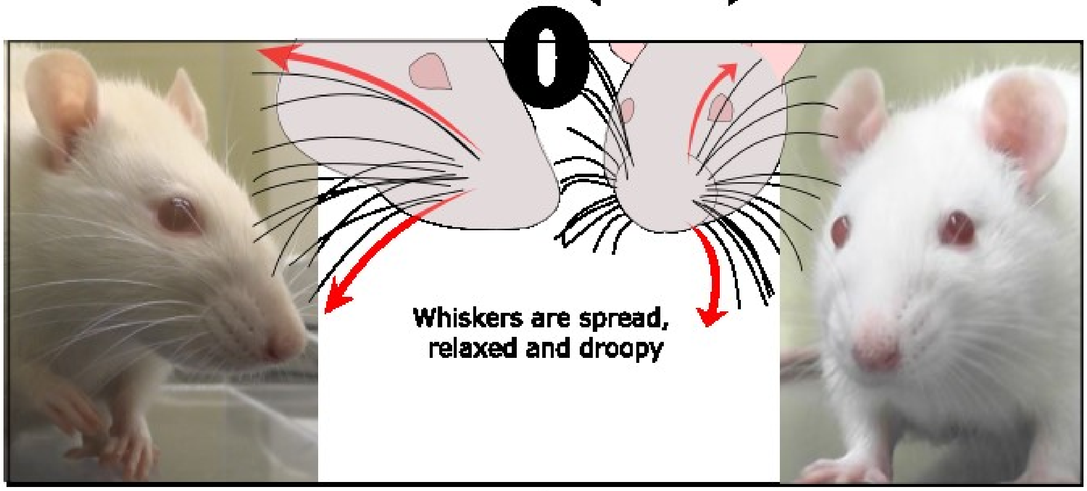
Muzzle flat

Small or no crease between nose and cheek

Nose slender and head looks pointy

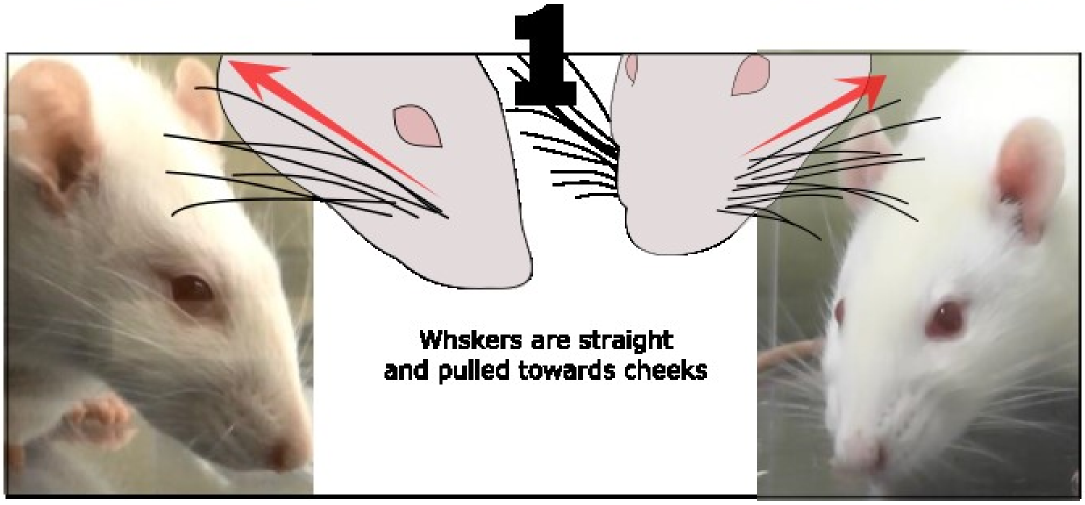
Rat Grimace Scale (RGS) - Whiskers

0



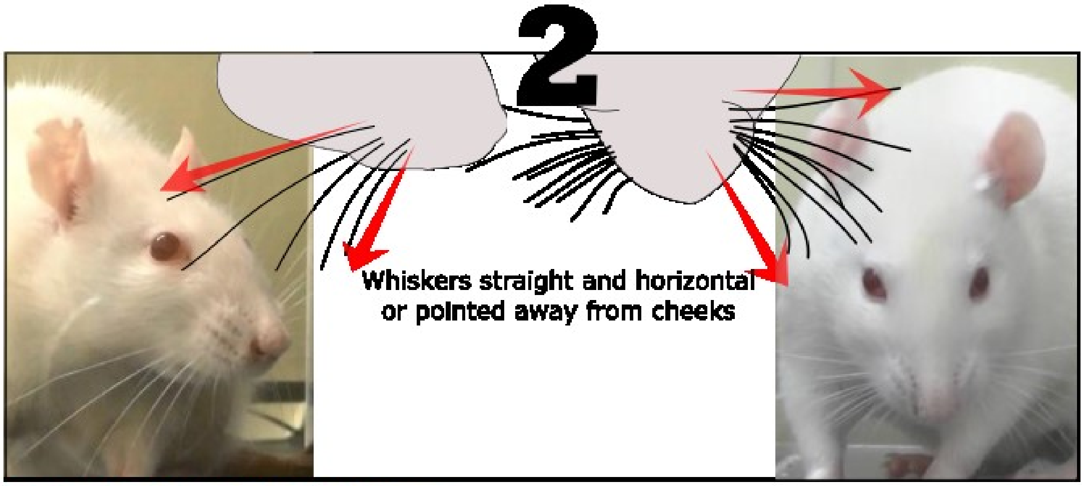
Whiskers are spread, relaxed and droopy

1



Whskers are straight and pulled towards cheeks

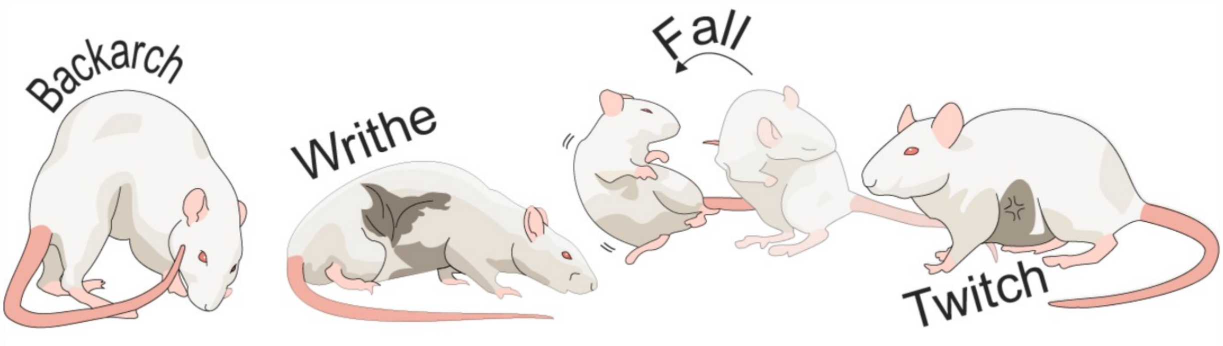
2



Whiskers straight and horizontal or pointed away from cheeks

Ignore whiskers right behind nose

5.3. Appendix C – composite behaviours in DSS colitis model



Legend: Cartoons of the four behaviours displayed by DSS-treated rats. Back arch = feline-like vertical stretch while inactive or walking; writhe = contraction of abdominal muscles; stagger/fall = loss of balance while rearing or grooming; twitch = transient involuntary muscle contraction.