The Short Form of the Glasgow Composite Measure Pain Scale in Post-Operative Analgesia Studies in Dogs: A Scoping Review

Barbara Testa

DVM (Hons), MRCVS

Submitted in fulfilment of the requirements for
the Degree of Master of Veterinary Medicine

School of Veterinary Medicine
College of Medical, Veterinary & Life Sciences
University of Glasgow

August 2022

Abstract

The measurement and treatment of acute pain in animals are essential from a welfare perspective. Valid pain-related outcome measures are also crucial for ensuring reliable and translatable findings in veterinary clinical trials. The short form of the Glasgow Composite Measure Pain Scale (GCMPS-SF) is a multi-item behavioural pain assessment tool, developed and validated using a psychometric approach, to measure acute pain in the dog. The psychometric approach refers to a scientific method used to develop tools intended to measure complex and multifaceted constructs like pain. Relevant words and expressions related to pain are collected, refined, and classified into domains and associated categories through a multi-step approach that involves the participation of a large sample of pain sufferers (or observers for non-verbal patients) and a pool of experts in the field. Ultimately, the instrument developed is tested by clinical studies to assess its validity, reliability, and responsiveness. While this approach has been widely adopted to reliably assess pain in humans, the GCMPS-SF is at present the only validated tool to measure acute pain in dogs developed using this methodology.

The GCMPS-SF comprises four sections (section A: observation of resting behaviours from a distance, section B and C: evaluation of interactive behaviours, section D: assessment of the overall attitude of the patient), with instructions for completion provided at the beginning of each section. The questionnaire encompasses two categories within each section, incorporating a total of six behavioural categories. These categories are associated with multiple descriptive expressions of pain, assigned an individual score each and ordered in an increased level of severity within the category.

We conducted a scoping review through systematic search of the literature to identify prospective research studies that have used the GCMPS-SF. We aimed to describe the contexts in which it has been used, verify the correct use of the scale, examine whether these studies are well-designed and adequately powered, and determine whether statistically significant differences in GCMPS-SF scores appear clinically relevant.

We identified 114 eligible studies, indicating widespread use of the scale.

We documented a limited number of modifications to the scale and intervention level, which would alter its validity, and a variety of methods to analyse the data derived from the scale.

We also documented many deficiencies in reporting of experimental design in terms of the observers used, the underlying hypothesis of the research, the statement of primary outcome, the use of *a priori* sample size calculations, blinding and randomisation strategies. These deficiencies in reporting and study design may predispose to both Type I and Type II statistical errors in the small animal pain literature. Results of our analyses also suggest that methodological factors affected study outcomes in our dataset. The probability of finding a statistically significant difference was 7 times higher in studies that used negative control groups, 3 times higher when the GCMPS-SF scores were used as a primary outcome, and 12 times higher if the pain scale was modified.

Finally, we documented a wide range (1.00 to 11.0) of actual effect sizes in GCMPS-SF scores, with approximately 30% of the values below 1.60, and a median largest actual effect size of 2.00 in trials that declared statistical significance. With the consideration that clinical relevance is perhaps more anchored to the intervention level with the GCMPS-SF, rather than to a minimum difference in pain scores, we question whether some of the differences detected, albeit statistically significant, are clinically relevant without accounting for their position on the scale.

Based on our findings, we encourage methodologically sound study design, high quality of reporting, and a more robust use of the scale and derived data to ensure attainment of reliable and translatable outcomes.

Table of contents

Abstract 2
List of tables 9
List of Figures 10
List of accompanying material 12
Preface 13
Acknowledgments 14
Author's declaration 15
Chapter 1 Introduction 16
1.1. Pain and nociception 16
1.2. Pain classification 17
1.2.1. Classification based on the nature of the stimulus 17
1.2.2. Classification based on the duration of the pain experience 19
1.3. Methodology for the recognition and quantification of acute pain in animals 20
1.3.1. The use of objective measures 21
1.3.2. Facial expressions 23
1.3.3. Behaviour-based pain scoring systems 26
1.3.3.1. Unidimensional pain scoring systems 26
1.3.3.2. Composite pain scoring systems 28
1.4. Concepts underlying the scientific development of pain scales using robust methodology 33
1.4.1. The psychometric theory 33
1.4.2. Validation process 34
1.4.2.1. Content validity 35
1.4.2.2. Criterion validity 36
1.4.2.3. Construct validity 36
1.4.3. Reliability 37
1.4.4. Responsiveness 37
1.4.5. Utility 38
1.4.6. Interval level measurement 39

1.5. Application of scientific methodology to the development of the Glasgow Composite Measure Pain Scale (GCMPS) for measurement of acute pain in dogs 41
1.5.1. Development of the first prototype scale 41
1.5.2. Development of the prototype into an interval level scale 42
1.5.3. Development of the short form of the GCMPS (GCMPS-SF) from the interval level prototype and derivation of an intervention score for provision of rescue analgesia..... 45
1.5.3.1. Development of the short form of the GCMPS 45
1.5.3.2. Derivation of an intervention level for provision of additional analgesia 49
1.5.3.3. Validation 50
1.6. Effects of confounding factors on the use of behavioural pain scoring systems 51
1.7. Effects of confounding factors and potential sources of bias on the GCMPS-SF scores 52
1.7.1. Observer-related factors 52
1.7.1.1. Experience of the individual 52
1.7.1.2. Number of observers 53
1.7.2. Patient-related factors 54
1.7.2.1. Anxiety 54
1.7.2.2. Temperament 55
1.7.3. The effect of sedative / analgesic drugs 56
1.8. The GCMPS-SF and acute pain study methodology 58
1.8.1. Hypothesis testing 59
1.8.2. Statistical power 59
1.8.3. Power analysis and sample size estimation 60
1.8.4. Significance level (alpha) and *p*-value 61
1.8.5. Confidence interval (CI) 62
1.8.6. Group sizes 64
1.8.7. Variability 65
1.8.8. Effect size 66
1.8.9. Control groups 67

1.8.10. Pain measurement instrument 68
1.8.11. Rescue analgesia provision 69
1.8.12. Data analysis 72
1.8.12.1. Analysis of data after administration of rescue analgesia 80
1.8.12.2. Survival analysis 82
1.8.13. Statistical errors 83
1.8.13.1. Type I statistical error 85
1.8.13.2. Type II statistical error 87
1.8.13.3. Interplay between Type I and Type II statistical errors 89
1.8.14. Study design 91
1.8.14.1. Hypotheses 91
1.8.14.1.1. Superiority design 91
1.8.14.1.2. Equivalence design 91
1.8.14.1.3. Non-inferiority design 91
1.8.14.2. Controlled *versus* observational 92
1.8.14.3. Blinding 93
1.8.14.4. Randomisation 94
1.8.14.5. Single *versus* multicentre 95
1.8.14.6. Clinical *versus* experimental 95
1.8.15. Minimum clinically important difference (MCID) 96
1.9. The use of evidence synthesis to evaluate research conduct 98
1.9.1. Systematic review 100
1.9.2. Scoping review 100
1.9.3. Indications for the conduct of scoping reviews 101
1.10. Aims and objectives of this project 102
1.10.1. Popularity 103
1.10.2. GCMPS-SF use 104
1.10.3. Study design and power 104
1.10.4. Actual effect size 105
1.10.5. Summary of the objectives and how they will be covered in this thesis 105
Preface to chapter 2 107

Chapter 2 The short form of the Glasgow Composite Measure Pain Scale in post-operative analgesia studies in dogs: a scoping review 109

2.1. Abstract 109

2.2. Introduction 110

2.3. Methods 111

2.3.1. Literature search 111

2.3.2. Inclusion criteria 112

2.3.3. Data extraction and appraisal 112

2.3.4. Variables describing the publications 112

2.3.5. Variables describing the use of the GCMPS-SF and measured data 113

2.3.6. Variables describing the study design 114

2.3.7. Statistical analysis 116

2.4. Results 116

2.4.1. Variables describing the publications 116

2.4.2. Variables describing the use of the GCMPS-SF and measured data 120

2.4.3. Variables describing the study design and power 123

2.5. Discussion 125

2.5.1. Appropriate use of the GCMPS-SF and derived data 125

2.5.2. The design of cute pain clinical trials 127

2.5.3. Limitations 128

2.5.4. Conclusions 129

2.6. Supplementary material 129

Chapter 3 Association of study design factors with the finding of a significant difference 130

3.1. Introduction 130

3.2. Methods 132

3.3. Results 133

3.4. Discussion 136

3.4.1. Limitations 141

3.4.2. Conclusions 141

Chapter 4 Maximum difference detected in GCMPS-SF scores between groups in studies that declared statistical significance 142

4.1. Introduction 142

4.2. Methods 144
4.3. Results 146
4.4. Discussion 150
4.4.1. Comparison of findings between surgical procedures 150
4.4.2. Comparison of findings between interventions 151
4.4.3. General considerations irrespective of the division in subgroups 152
4.4.3.1. Desired effect size 153
4.4.3.2. Actual effect size 154
4.4.4. Limitations 155
4.4.5. Conclusions 156
Chapter 5 Summary discussion 157
5.1. General discussion 157
5.1.1. Limitations 162
5.1.2. Conclusions 163
Appendices 165
List of references 196

List of Tables

Table 1.1 Most popular behavioural pain scales to measure acute pain in dogs and their points of difference…. 32

Table 1.2 Summary of a few common statistical parametric tests and their nonparametric equivalents …… 79

Table 1.3 Summary of statistical errors …. 84

Table 1.4 Characteristics of traditional literature reviews, scoping reviews, and systematic reviews. Reprinted from Munn *et al* 2018 …… 99

Table 2.1 Variables describing the publications included in the review …… 118

Table 2.2 Variables from publications in the review describing how the CMPS-SF was used …… 121

Table 2.3 A summary of handling data from the CMPS-SF and the statistical techniques used. 104 studies are included in this table and the 10 observational studies in the review omitted …… 122

Table 2.4 Variables describing features of study design in the publications …. 124

Table 3.1 Results of multivariable binomial logistic regression to estimate the odds of finding a statistical difference in GCMPS-SF scores or GCMPS-SF guided rescue analgesia versus not finding a difference. The analysis was conducted on the 83 controlled studies with a superiority hypothesis and formal statistical testing. The McFaddens R^2 and AIC for the model were 0.251 and 114, respectively ……. 135

Table 4.1 Summary of the type of surgeries grouped into broader categories 145

Table 4.2 Summary of type of intervention utilised by the 30 included studies ……. 149

List of Figures

Fig. 1.1 Mouse Grimace Scale. Reprinted from Langford *et al* 2010 25

Fig. 1.2 Examples of mono-dimensional pain scoring systems. Reprinted from Reid *et al* 2013 27

Fig. 1.3 CSU Scale for acute pain in dogs. Reprinted from Mich *et al* 2010 30

Fig. 1.4 GCMPS. Reprinted from Reid *et al* 2007 44

Fig. 1.5 GCMPS-SF. Reprinted from Reid *et al* 2007 48

Fig. 1.6 Example of different values of confidence level determining the width of the CI. Reprinted and adapted from www.365datascience.com 63

Fig. 1.7 Example of normal distribution. Reprinted and adapted from www.scribbr.com 74

Fig. 1.8 Examples of positive (a) and negative (b) skewed distribution. Reprinted and adapted from www.scribbr.com 74

Fig. 1.9 Quartiles and Interquartile range (IQR). Reprinted from www.scribbr.com 76

Fig. 1.10 Null hypothesis probability curve and probability of Type I statistical error. Reprinted and adapted from www.scribbr.com 86

Fig. 1.11 Alternative hypothesis probability curve and probability of Type II statistical error. Reprinted and adapted from www.scribbr.com 88

Fig. 1.12 Interplay between Type I and Type II error rate. Reprinted and adapted from www.scribbr.com 90

Fig. 2.1 Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flowchart showing the number of studies included in each stage of the review 117

Fig. 2.2 The number of publications using the CMPS-SF by year of publication between 2007 and 2019. The trendline represents a 2-year rolling average of the number of publications 117

Figure 4.1 Box plot of the largest difference found in GCMPS-SF scores in the 30 studies that declared a statistically significant difference 147

Figure 4.2 Box plot of the largest difference in GCMPS-sf scores detected in soft tissue and orthopaedic surgeries 147

Figure 4.3 Box plot of the largest difference in GCMPS-sf scores detected in TPLO (tibial plateau levelling osteotomy) and OVH (ovariohysterectomy) surgeries 149

Figure 4.4 Box plot of the largest difference in GCMPS-sf scores based on the intervention category (drug, regional anaesthesia, and surgery) 149

List of accompanying material

Appendix 1 Omission of the definitions of descriptors in the GCMPS-SF – authors' rationale..... 165

Appendix 2 Search strategy 166

Appendix 3 Studies Using the GCMPS-SF included in the review. Supplementary table and supplementary references .....….. 168

Appendix 4 Univariable tests investigating the association between study design factors and the finding of a statistically significant difference 188

Appendix 5 Summary of all the procedures used in the 30 clinical trials that reported the largest difference in GCMPS-sf scores detected 195

Preface

The author and the primary supervisor, Dr Andrew Bell, conceived and designed the study. Literature search was performed by the author. Data extraction and data analysis were performed by the author with input from the primary supervisor. The primary supervisor gave a considerable contribution to statistical analysis. The thesis and associated manuscripts were prepared by the author with mentorship from the supervisors, Dr Andrew Bell and Prof Pamela J. Murison.

The study presented in chapter 2 has been published in a peer-reviewed journal [Testa B, Reid J, Scott ME, Murison PJ and Bell AM (2021) The Short Form of the Glasgow Composite Measure Pain Scale in Post-operative Analgesia Studies in Dogs: A Scoping Review. Front. Vet. Sci. 8:751949. doi: 10.3389/fvets.2021.751949].

Acknowledgments

Above all, I would like to thank my supervisors, Dr Andrew Bell and Prof Pamela J Murison, for the invaluable mentorship provided throughout the study. This thesis would have not been possible without their input. I am extremely grateful to Dr Andrew Bell, whose guidance and expertise were irreplaceable for statistical analysis and for every step of this research project.

I would like to thank Newmetrica Ltd. for generously funding open access fees for the study published in chapter 2.

I would also like to thank my supervisors, and all the anaesthesia team at the University of Glasgow, for their support and guidance during my four years of residency. They all contributed so much to my professional and personal growth. A special thank goes to Prof Pamela J Murison, whose level of teaching, care, and knowledge is the model of the anaesthesiologist I wish I could become.

I would like to thank my Italian and international friends, who have always been there to support me in good and bad times. I owe them my mental integrity and the strength to complete this journey.

Finally, a thank from the bottom of my heart is for my dad, whose encouragement has supported me throughout my life and my veterinary carrier.

Author’s declaration

I declare that this dissertation is the result of my own work, and that it is an original project not submitted for any other degree or professional qualification at the University of Glasgow or any other Institution. Contribution of others and replication of figures are explicitly acknowledged and referenced.

Barbara Testa

CHAPTER 1
INTRODUCTION

Formal methods for the recognition and quantification of pain in animals have been the object of interest of numerous studies, due to the importance of reliable and reproducible measurement of pain in a number of different contexts.
Clinically, an accurate assessment of the individual patient's level of pain enables the clinician to provide adequate analgesia, and to titrate it according to the patient response. Measurement of pain is important also in veterinary and translational clinical research where the assessment of the efficacy of analgesic interventions relies on the valid measurement of this abstract construct. Recently, the translational value of naturally occurring companion animal models has been discussed in various settings. Kol and colleagues highlighted (Kol *et al* 2015) how canine cancer, which accounts for the cause of death in approximately 50% of dogs above 10 years of age, represented a statistically powerful model. While the appraisal of translational value started with comparative oncology, it has more recently moved to investigate the role of naturally occurring animal models to pain research. On this subject, Klinck *et al* (2017) stressed how various aspects, such as veterinary subject diversity, the pathophysiological similarities to humans, the fact that pets share the same environmental diversity as their owners, could yield better generalizability of findings and improved translational potential.

1.1. Pain and nociception

To understand the complexity of the subject, a preliminary distinction must be made between "pain" and "nociception". While "nociception" is the sensory mechanism that allows animals to sense and avoid potentially tissue-damaging insults, thus strictly representing the neural process of encoding noxious stimuli, the term "pain" refers to a much more complex and comprehensive sensory and emotional experience derived from the central elaboration of the noxious stimulus and associated with physiological and behavioural changes (McKune *et al* 2015, Bell 2018, Mischkowski *et al* 2018).

To describe the multi-faceted pain experience, the International Association for the Study of Pain (IASP) defined it as an "Unpleasant sensory and emotional experience associated with, or resembling that associated with, actual or potential tissue damage" (Raja *et al* 2020).

1.2. Pain classification

Conscious perception and elaboration of the noxious stimulus represents the result of a complex interaction between inhibitory and facilitatory peripheral and central nervous system pathways, which can result in different types of pain depending on the nature and duration of the primary trigger, and on the resulting transient or permanent changes in the above-mentioned pathways (McKune *et al* 2015). With respect to the nature of the primary trigger and the relative changes generated in the pain pathways, pain can be classified as nociceptive, neuropathic, and nociplastic (Fitzcharles *et al* 2021), and it can be divided into acute and chronic based on the duration of the pain experience, irrespective of its nature (Spacek 2006).

1.2.1. Classification based on the nature of the stimulus

Nociceptive pain derives from the activation of nociceptors by thermal, mechanical or chemical stimuli, which are processed by a normally functioning somatosensory system (Moore 2016), and is associated with injury or disease of somatic tissues such as skin, muscle, tendons, bone and joints (Goldman *et al* 2020). A type of nociceptive pain is inflammatory pain, which results from activation and sensitisation of nociceptor terminals by inflammatory mediators such as bradykinins, cytokines, prostaglandins, leukotrienes, serotonin, histamine, calcitonin gene-related peptide (CGRP), substance P, purines such as ATP, protons, free radicals, lipids, chemokines, and neurotrophines such as nerve growth factor (NGF) (Bell 2018). Nociceptive and inflammatory pain states are commonly associated with acute injury and trauma induced by surgery (McKune *et al* 2015), and represent an adaptive mechanism to prevent further damage

which terminates with the completion of the healing process. Clinical signs elicited by nociceptive pain include a variety of behavioural responses (examples of which are aggression, vocalisation, and restlessness), and attention to or aversive reactions to touch of the painful/injured area. Overall, despite the variety in responses, clinical signs associated with nociceptive pain are often predictably relatable to the presence and the degree of tissue damage (Hernandez-Avalos *et al* 2019).

Neuropathic pain is a maladaptive phenomenon caused by a disease or injury affecting the somatosensory system which persists beyond resolution of the initiating cause (Moore 2016). While the initiating stimulus might be associated with nociceptive pain, intense chronic nociceptive pain that outlasts the original insult determines structural and functional changes in the peripheral nervous system, spinal cord and brain that typically characterise neuropathic pain states. Key changes underlying the development of neuropathic pain involve injury-induced hyperexcitability of afferent neurons, which generates ectopic action potentials; peripheral sensitisation, characterised by intrinsic hyperexcitability and reduced threshold of peripheral nociceptors; central sensitisation, caused by repeated release of excitatory molecules in the dorsal horn of the spinal cord which leads to altered expression of multiple receptors (voltage-gated sodium channels, AMPA and NMDA receptors) and altered modulation of inhibitory pathways; and persistent pathologic activation of microglia within the CNS with release of inflammatory mediators that perpetuate changes in central modulation of painful stimuli and hyperexcitability of nociceptive neurons (Moore 2016). Common neuropathic pain states include diabetic neuropathy, spinal cord lesions and central post-stroke pain (Goldman *et al* 2020). Clinical signs that characterise neuropathic pain are frequently more subtle than those associated with nociceptive pain. They can manifest as decreased general activity, changes in posture, altered demeanour or appetite, phantom scratching, and vocalisation in the absence of a clear painful stimulus (Moore 2016). Human patients report sharp spontaneous pain and dysaesthesia (i.e., numbness and tingling), and the same can be assumed for companion animals.

Nociplastic pain does not show clear evidence of tissue damage or lesions or diseases to the somatosensory system, but is associated with altered peripheral and central processing and modulation of pain (Herzberg *et al* 2021). The exact

mechanisms of nociplastic pain have yet to be elucidated and fibromyalgia represents an example of this pain type in humans. The symptoms observed in nociplastic pain include peripheral widespread and/or intense multifocal pain without obvious identifiable tissue or nerve damage, and central-related signs such as fatigue, sleep, memory, and mood alterations (Fitzcharles *et al* 2021).

The term "mixed pain" refers to conditions characterised by the co-existence of different types of pain. Cancer pain represents a typical example both in human (Goldman *et al* 2020) and in veterinary medicine (McKune *et al* 2015), displaying neuropathic and inflammatory components.

1.2.2. Classification based on the duration of the pain experience

Acute pain is defined as 'pain of recent onset and probable limited duration. It usually has an identifiable temporal and causal relationship to injury or disease' (Ready *et al* 1992). Frequently associated with a surgical stimulus (McKune *et al* 2015), acute pain is largely nociceptive and inflammatory in aetiology, resolves within days or weeks, ceases with the healing of injured tissues (thus being self-limiting) and serves the purpose of conditioning the animal's response in order to avoid or minimise further exposure to the potentially damaging stimulus and set the circumstances for the healing process (Grichnik *et al* 1991).
In contrast, chronic pain is of longer duration (arbitrarily defined as pain lasting for three months or longer – Mathews *et al* 2014 – even though a precise temporal distinction between acute and chronic pain is not entirely clear), tends to outlast the original insult beyond the healing time, it does not have a biological purpose and no defined end-point (Grichnik *et al* 1991, Ready *et al* 1992).

Despite the fact that acute and chronic pain may not necessarily represent distinct entities, but rather a continuum, due to possible transition of acute pain into chronic pain (Spacek 2006), it is important to recognise that the nature of the stimulus generates different types of pain. These different types represent separate clinical entities, in terms of their aetiology, pathophysiological processes, clinical manifestations, and, thus, recognition (Langford *et al* 2010)

and treatment options (McKune *et al* 2015, Fitzcharles *et al* 2021). Nociplastic pain, for instance, will generally display lower responsiveness to peripherally directed therapies such as nonsteroidal anti-inflammatory drugs (Fitzcharles *et al* 2021).

As the present thesis focusses on the measurement of postoperative acute pain in dogs, concepts related to methods for recognition and quantification of other types of pain will not be covered.

1.3. Methodology for the recognition and quantification of acute pain in animals

The broad intricacy of the patient comfort, which incorporates socio-economic, cultural, cognitive, affective, and provider-related components (Johnston *et al* 2021), is commonly assessed in human medicine with the use of questionnaires and scales like the Likert scale (made up of numbers, being in fact a set of ordered categories). These scales represent one of the most common tools to score pain and discomfort in the postoperative period (Johnston *et al* 2021). In this review, the authors discuss how pain is linked to multiple aspects of the postoperative experience, and that no single metric can be used alone to assess patient comfort. The overall patient experience and satisfaction are in fact better captured by integrating multiple standardised endpoints (Delphi consensus), which range from pain intensity (at rest and during movement), pain at 24 hours postoperatively, nausea and vomiting, completion of quality-of-recovery scales (QoR-15), time to gastrointestinal recovery, time to mobilisation, to sleep quality. Measurement of all these outcomes is based on self-assessment though, which poses a further challenge in quantifying pain in patients and species incapable of self-reporting. Animals indeed are non-verbal patients, thus leaving the assessment of their pain experience to interpretation of body language, facial expressions, behavioural changes, and changes in objective measures such as physiological variables, all of which carry species and individual variability.

1.3.1. The use of objective measures

Examples of objective measures are physiological variables such as heart rate, respiratory rate, pupil diameter, and blood pressure, of which the association with pain has been studied extensively. None of these parameters have been found reliable in isolation, despite their inclusion in some recent multi-dimensional scoring systems like the University of Melbourne pain scale (Firth and Haldane 1999), 4A-VET (Holopherne-Doran *et al* 2010), UNESP-Botucatu multidimensional composite pain scale in cats (Brondani *et al* 2011, 2013). This might be the result of the potential influence of other confounders such as stress, anxiety and fear on these physiological variables, especially in a hospital environment (Hansen 1997, Kyles *et al* 1998, Holton *et al* 1998b).

A case in point is the interrelation between physiological stress and pain, which activate distinct responses in the body, yet they share considerable physiological overlapping effects. Both acute stress and pain induce activation of the sympathetic nervous system, with release of epinephrine and norepinephrine from the adrenal medulla, resulting in increases in heart rate, systolic and diastolic blood pressure, and diversion of blood flow to the brain and muscles (Brotman *et al* 2007, Burton *et al* 2016). Acute stress is also characterised by a neuroendocrine response which involves stimulation of the hypothalamic-pituitary-adrenal (HPA) axis: the hypothalamus releases corticotropin-releasing hormone (CRH), which is responsible for the secretion of the adrenocorticotrophic hormone (ACTH) from the anterior pituitary gland. ACTH ultimately targets the zona fasciculata of the adrenal cortex resulting in the secretion of glucocorticoids, particularly cortisol (Ahmad *et al* 2015). In contrast, there is no clear evidence that acute pain activates the HPA resulting in cortisol release (Abdallah 2017). However, both acute pain and surgery can induce a stress response (Fox *et al* 1994, Srithunyarat *et al* 2016), thus resulting in superimpositions in the changes in physiological variables and hormonal products of the neuroendocrine response.

Plasma concentrations of cortisol have been measured in association with physiological stress induced by surgery or other painful procedures, and correlations have been investigated between changes in these variable and pain scores. However, in some studies plasma cortisol concentrations have been shown

to poorly correlate with pain scores in dogs (Srithunyarat *et al* 2016 and 2017), rabbits (Keating *et al* 2012), horses (Rietmann *et al* 2004), lambs (Molony *et al* 1997), calves (Tschoner 2021) and dairy cows (Des Roches *et al* 2017).

Srithunyarat *et al* investigated the correlation between multiple objective measures with pain scores and stress behaviour in healthy dogs undergoing ovariohysterectomy (2016) and dogs with bone fractures (2017). The objective measures investigated were physiological variables (temperature, heart rate, and respiratory rate), plasma cortisol concentrations, and vasostatin and catestatin. These latter two are measurable bioactive epitopes of chromogranin A, a glycoprotein co-released with catecholamines from the adrenal medulla following activation of the sympathetic nervous system. Findings from the study conducted in 2016 revealed significant differences in all these variables before and after surgery, but none of them demonstrated a reliable correlation with pain scores measured either with the Glasgow Composite Measure Pain Scale – Short Form (GCMPS-SF) or the Visual Analogue Scale (VAS) in any of the two studies.

Another work conducted in rabbits undergoing tattooing (Keating *et al* 2012) investigated the effects of application of EMLA cream on the changes in cardiovascular responses, serum cortisol concentration, and behavioural and facial expressions of pain. The authors couldn't find any significant correlation between physiological and serum cortisol responses and acute pain in rabbits that received sham and tattoo treatments with or without EMLA cream, while facial expressions did appear more reliable to assess acute pain.

In contrast, somewhat conflicting findings emerged from a study in an equine experimental orthopaedic pain model on 18 otherwise healthy horses (Bussières *et al* 2008) conducted to develop a composite pain scale (CPS) for acute orthopaedic pain in horses. In this research, acute inflammatory pain was associated with synovitis induced by injection of intrasynovial amphotericin B. Heart rate, respiratory rate, bowel sounds, rectal temperature, non-invasive systemic arterial blood pressure (NIBP), serum glucose and cortisol were investigated in conjunction with behavioural signs of pain. Despite pointing out the relatively low sample size and the non-generalizability of their results due to the specific type of pain considered, the authors did find the correlation between plasma concentrations of cortisol and pain scores to be moderate. Amongst the

physiological variables assessed, correlation with pain scores was overall poor for rectal temperature and bowel sounds, only moderate for heart rate and respiratory rate, whilst it was excellent for mean NIBP demonstrating both high specificity and sensitivity as an indicator of acute orthopaedic pain in a controlled experimental setting.

Serum or plasma concentrations of cortisol have been studied extensively also in the farm animal literature, especially in pain research in calves (Tschoner 2021), frequently in association with other objective measures such as changes in acute phase proteins like haptoglobin, serum and milk amyloid A. However, while changes in acute phase proteins have been linked more specifically to inflammation (Eckersall *et al* 2001, 2010) and phases of the disease process (Des Roches *et al* 2017) than to pain, measurement of cortisol concentrations demonstrated a predominant role as indicator of acute stress (Molony *et al* 1997, Des Roches *et al* 2017, Tschoner 2021).

1.3.2. Facial expressions

As previously mentioned, the pain experience encompasses sensory and emotional components. While aspects of the sensory component can be recognised and quantified by tools such as sensory testing, the detection and evaluation of the emotional component is more subjective and largely relies on verbal communication in adult human patients (Machado *et al* 2020) or on the use of behavioural scales in non-verbal patients. Despite being widely accepted that animals are capable of exhibiting facial expressions of other emotional states (Langford *et al* 2010) and to process them (Tate *et al* 2006), systematic and reproducible methods for the evaluation of the emotional component of the pain experience can be problematic in non-human mammalian species due to their inability to self-report (Flecknell 2010).

Changes in facial expressions related to common emotions have been characterised and coded in humans using the “action units” based on facial muscle groups of the facial action coding system (Ekman *et al* 1978). Similar scales have been successfully developed and adapted to assess clinical pain and response to

analgesia in human patients with absent or impaired verbal communication, such as infants and people affected by cognitive impairments (William 2002). The first researchers to study and code, using a method analogue to the facial coding system in humans, facial expressions of pain in veterinary medicine were Langford *et al* in 2010, who developed a Grimace Scale in the laboratory mouse. The authors first identified five relevant features of the mouse's face which displayed relevant and consistent changes in association with acute pain: orbital tightening, nose bulge, cheek bulge, ear position, and whisker change. The intensity of the change of each of these features was then scored on a three-point simple descriptive scale (SDS) (absent, moderate, severe). Finally, the collected images of mice exhibiting facial expressions of different levels of pain were ordered to create the Mouse Grimace Scale (Fig. 1.1).

Since then, facial expressions have interested researchers across multiple species, and they still have considerable value as identified in composite pain scales (for example the Feline GCMPS) and many Grimace scales are available in cats (Holden *et al* 2014, Evangelista *et al* 2019), horses (Dalla Costa *et al* 2014, Gleerup *et al* 2015), donkeys (Orth *et al* 2020), rabbits (Keating *et al* 2012), rats (Sotocinal *et al* 2011), sheep (McLennan *et al* 2016, Hager *et al* 2017), lambs (Guesgen *et al* 2016), ferrets (Reijgwart *et al* 2017), dairy cattle (Gleerup *et al* 2015), piglets (Di Giminiani *et al* 2016, Viscardi *et al* 2017), sows (Navarro *et al* 2020) and harbour seal pups (MacRae *et al* 2018). In fact, especially in laboratory species, reliable, reproducible and non-time consuming methods for pain assessments represent a valuable aid to ensure animal welfare and test the efficacy of new drug interventions (Flecknell 2010, Langford *et al* 2010).

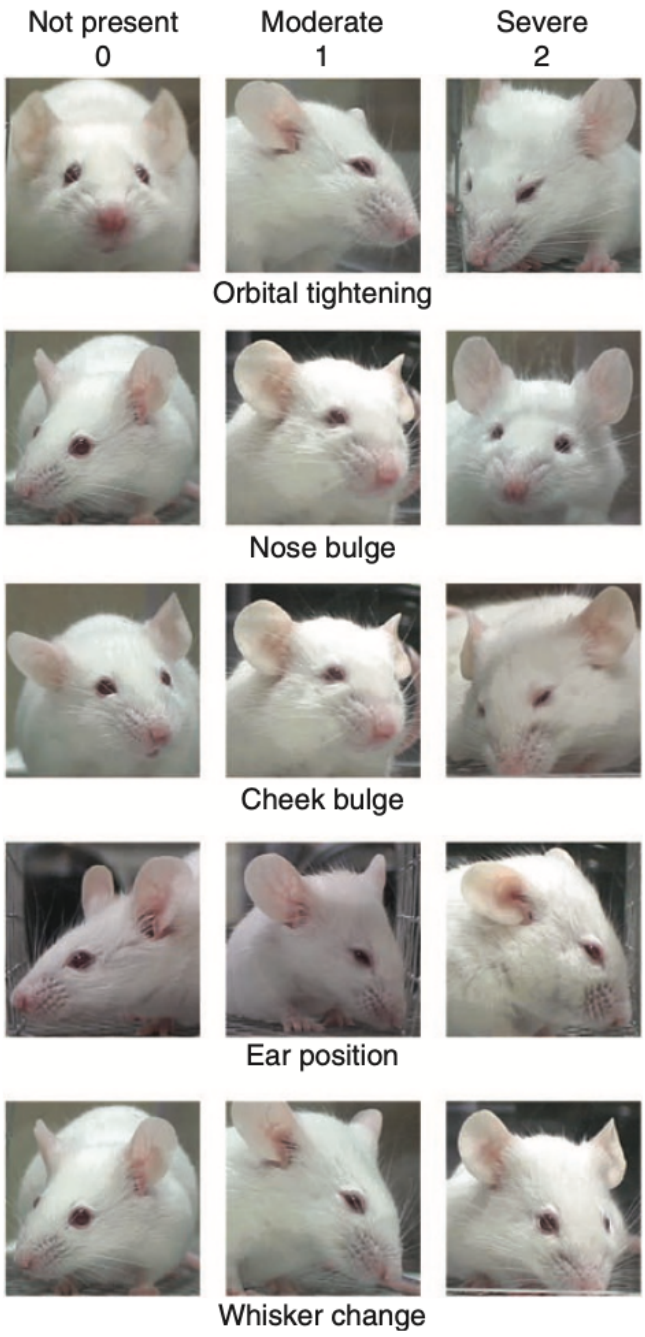

Fig. 1.1 Mouse Grimace Scale. Reprinted from Langford *et al* 2010.

1.3.3. Behaviour-based pain scoring systems

To try to capture the complex, multi-dimensional experience of pain in its entirety, an alternative multimodal approach involving the use of behaviour-based pain scoring systems has been developed and refined over the years (Holton *et al* 2001).

1.3.3.1. Unidimensional pain scoring systems

The first behavioural tools developed in the veterinary literature were simple unidimensional scales utilised to score the intensity of pain experienced by the patient (Holton *et al* 1998a). Examples are the Simple Descriptive Scale (SDS), the Visual Analogue Scale (VAS), and the Numeric Rating Scale (NRS) (Fig. 1.2). A refinement of the VAS is represented by the Dynamic Interactive Visual Analogue Scale (DIVAS), which adds a dynamic and interactive assessment of the patient involving observation from a distance, interaction with the patient and palpation of the wound/painful area.

The SDS is typically composed of 4 or 5 descriptors (no pain, mild, moderate, severe, very severe), thus being highly subjective and lacking in sensitivity for the detection of small changes (Downie *et al* 1978).

The NRS uses numbers instead of descriptors to score pain, typically from 0 to 10, where 0 represents no pain and 10 the worst possible pain. On one hand, this structure improves discrimination between categories and consequently the performance of this scale for comparative purposes (Downie *et al* 1978). On the other hand, it is discontinuous compared to the VAS and it can have unequal weight between the categories (Hansen 2003).

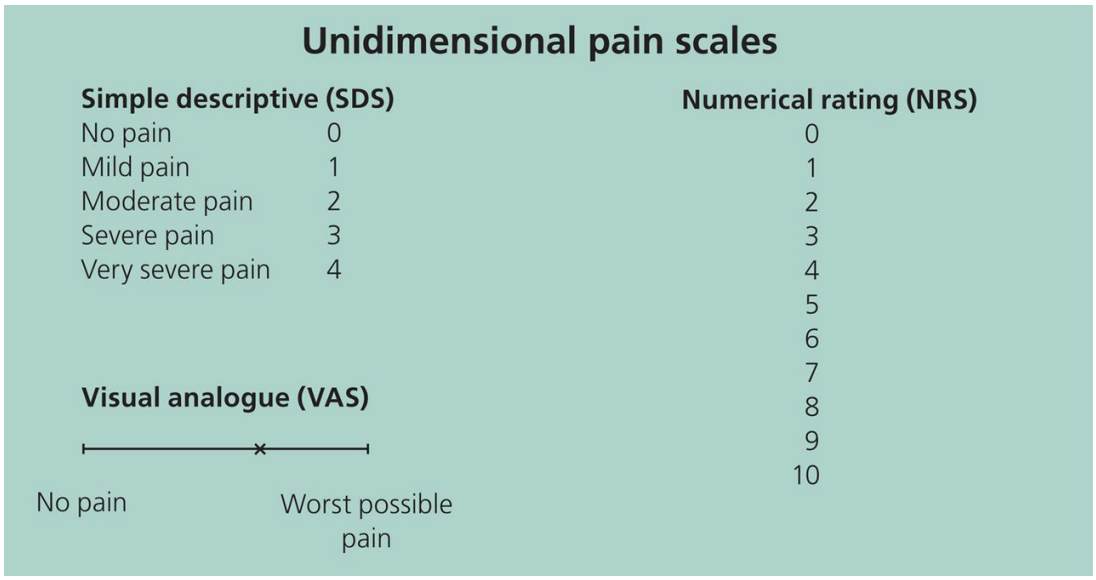

Fig. 1.2 Examples of mono-dimensional pain scoring systems. Reprinted from Reid *et al* 2013.

With the VAS, the assessor is asked to place a straight vertical mark on a direct line, conventionally 100 mm-long, in a position between the two extremes which subjectively corresponds to the level of the patient's pain. The two extremes are usually anchored with 'no pain' or 'the worst possible pain' although in some studies 'the worst possible pain for that procedure' is used. The two main problems associated with the use of the VAS are the considerable training required to reduce intra- and inter-observer variation and the linearity of the scale (Fox *et al* 2000, Hansen 2003, McKune *et al* 2015). This latter aspect is of relevant importance when choosing a proper statistical method to analyse the data and when comparing results from different studies (Chapman 1976, Mantha *et al* 1993, Holton *et al* 2001).

Overall, these scales are very easy to use, but they all lack sensitivity, are prone to a great inter observer variability and are influenced by observer-specific related factors, such as experience, age, gender, personal experience, training, and personal health (Price *et al* 2002).

1.3.3.2. Composite pain scoring systems

Whilst unidimensional scales measure only one dimension of the pain experience, namely its intensity, the most recent attempts have been focused on creating multifactorial metrology instruments which also consider the sensory and affective components of pain, integrating various aspects of the patient resting, interactive behaviours and overall attitude.

They are more complex, made up of multiple different domains and associated categories, each one of them composed of several sub-category expressions, which are scored separately and assigned their own weight. These scales include observation of spontaneous behaviour from a distance, assessment of interactive behaviours at rest and during specific movements and palpation of specific areas. Cumulative scores of each category would then form the final pain score assigned to the patient, with a positive relationship between the total score and the level of pain.

An example of this approach is the multifactorial numerical rating equine composite pain scale (ECPS) (Bussières *et al* 2008), developed for acute orthopaedic pain in horses utilising a multifactorial NRS. This scale incorporates three main domains (behaviour, physiologic data, response to treatment), each one of them encompassing several categories. Behavioural categories include appearance (reluctance to move, restlessness, agitation, and anxiety), sweating, posture (weight distribution, comfort), kicking at abdomen, pawing on the floor, head movement, and appetite. Physiologic parameters comprise heart rate, respiratory rate, digestive sounds, and rectal temperature, while response to treatment consists of interactive behaviour and response to palpation of the painful area. Each category is assigned a list of descriptors weighed between 0 (normal) and 3 (worst deviation from normality), for a maximum possible cumulative score of 39.

A different approach is exemplified in the small animal literature by the Colorado State University (CSU) Scale for Acute Pain (Mich *et al* in 2010) (Fig 1.3). This scale is based on a multifactorial SDS, as indicated by the generic 0-4 scale displayed on the left side (which corresponds to increasing levels of pain from the top to the bottom), and it includes visual aids to assist in the patient evaluation: it is colour-coded for different levels of pain and each level also features a drawing which adds a further visual cue and encourages the assessor to observe the overall patient's pain behaviour without focusing solely on a specific area. Importantly, what differentiates the CSU pain scale from a SDS is the multifactorial approach adopted, as it takes into consideration three main domains: psychological and behavioural signs of pain (vocalisation, attention to wound/painful area, attention to surroundings, interactive behaviour, overall attitude), response to palpation, and body tension, and each domain contains multiple descriptors of the behaviours under evaluation.

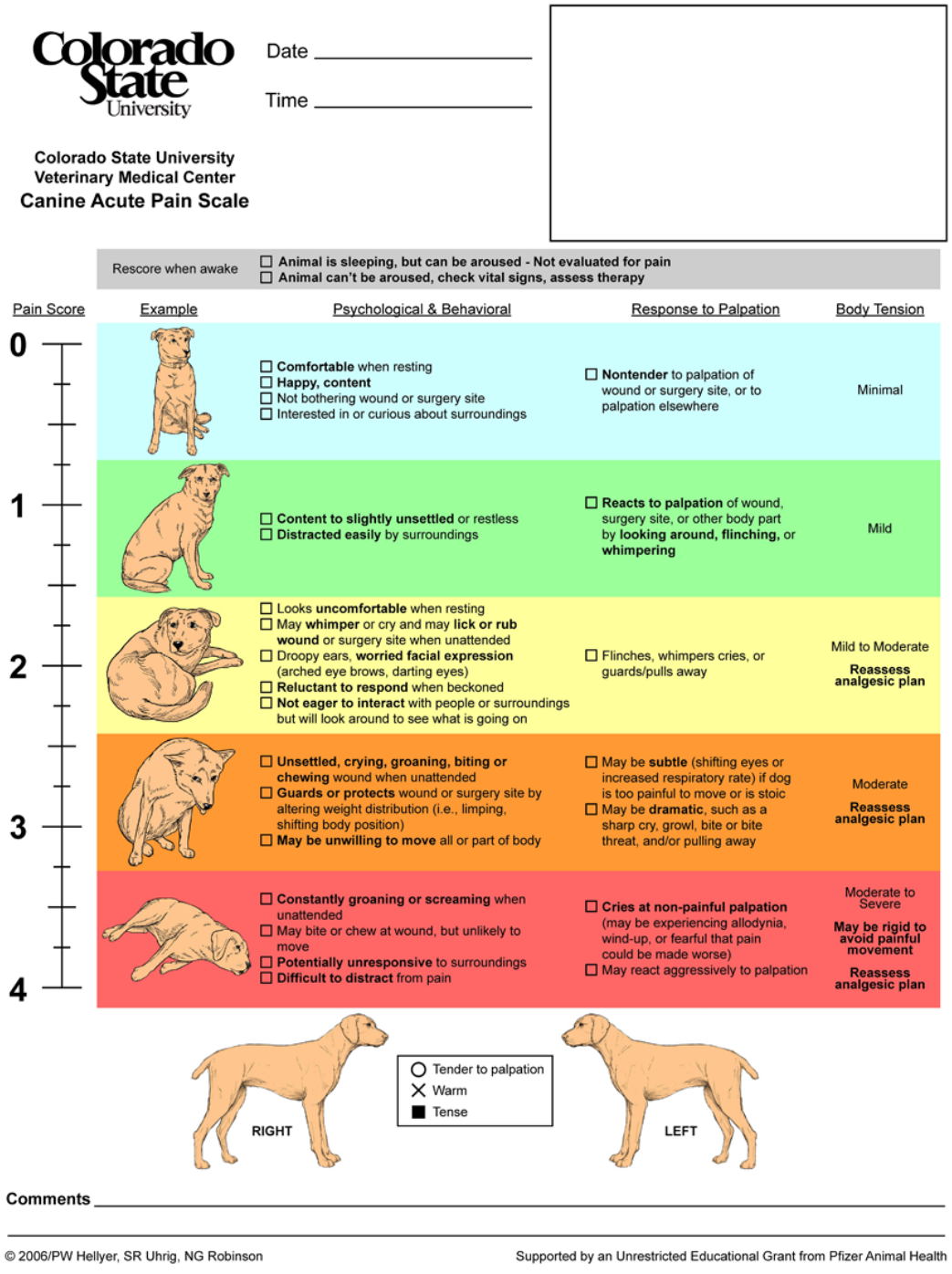

Fig. 1.3 CSU Scale for acute pain in dogs. Reprinted from Mich *et al* 2010.

In all these scales, the inclusion of descriptors minimises the interpretation required for each category, and this represents an advantage in terms of higher sensitivity and specificity (Guillot *et al* 2011) and reduced inter-observer variability (Morton *et al* 2005, Murrell *et al* 2008) compared to the unidimensional instruments previously described. Morton *et al* (2005), for example, included a full list of detailed definitions for all the expressions used in Glasgow Composite Measure Pain Scale (GCMPS), stating that this approach would reduce the training required to use the scale, the method for completion would be clearly understood, and the evaluator would be left in no doubt in the choice of a descriptor when interpreting a patient's pain behaviour. Furthermore, the cumulative score provides a more accurate and comprehensive picture of the animal's pain experience compared to the subjective evaluation of its intensity alone provided by unidimensional pain scoring systems (Holton *et al* 1998a, Price *et al* 2002).

However, when conceiving a multifactorial behavioural scale to assess pain, essential elements of its foundation are constituted by the validity of its contents and the criteria to select them (Holton *et al* 2001). A plethora of behaviours have been observed in the postoperative period associated with pain by many researchers involved in the development of such tools (Morton *et al* 1985, Sanford *et al* 1986, Conzemius *et al* 1997, Hellyer *et al* 1998, Firth *et al* 1999). For example, Fox *et al* (2000) identified an extensive list of 166 behaviours associated with pain in the postoperative period in both groups of bitches that underwent ovariohysterectomy, either assigned to receiving butorphanol or placebo.

Refining the list of all possible behaviours to a selected number of items that are reliably and consistently associated with pain requires robust methodology. In many of these scales, detailed criteria for inclusion of items were not given and the validity of their contents was not tested by clinical studies (Holton *et al* 2001).

The most popular composite behavioural pain scales in use to measure acute pain in dogs and their points of difference are summarised in Table 1.1.

Table 1.1. Most popular composite behavioural pain scales to measure acute pain in dogs and their points of difference.

Scale	**Behavioural categories**	**Objective measures (physiological variables)**	**Principles of development**	**Intervention level for provision of additional analgesia**	**Validated**
GCMPS-SF	Yes	No	Psychometric approach	Yes	Yes
Colorado State University Canine Acute Pain Scale	Yes	No	Multifactorial SDS	No	No
University of Melbourne Pain Scale	Yes	Yes	Multifactorial NRS	No	Yes
4AVET	Yes	Yes	Multifactorial NRS	No	Yes

1.4. Concepts underlying the scientific development of pain scales using robust methodology

The development of a reliable instrument to measure an intangible construct like pain represents a challenge both in veterinary and in human patients. In the human literature, this challenge has been addressed in psychiatry by applying psychometric methods to measure attributes like intelligence, anxiety, quality of life, and depression, using formally assessed structured questionnaires (Guyatt *et al* 1992, Streiner *et al* 1995). As the word 'pain' refers "not to a specific sensation which can vary only in intensity, but to an endless variety of qualities that are categorized under a single linguistic label" (Melzack *et al* 1971), Melzack and Torgerson applied the same methodology in 1971 to develop a 'language of pain', which subsequently formed the basis of the McGill pain questionnaire (Melzack 1975), designed to provide quantitative assessment of clinical pain that could be treated statistically. The original method was subsequently modified, but many versions are still in use today, proving the validity of their content and their criteria (Holton *et al* 2001). In the veterinary literature, the psychometric approach utilised for the construction of the McGill pain questionnaire was adopted by Holton et al in 2001 during the development of Glasgow Composite Measure Pain Scale (GCMPS).

1.4.1. The psychometric theory

Well-established psychometric methods should be applied during the construction of a pain measurement composite scale to ensure that the content of the resultant instrument is valid and actually measures the property of interest (Morton *et al* 2005). This concept is fundamental both in the clinical context (considering the wide use of these scales and, as such, their impact as clinical decision-making tools) and in research, in light of the importance of a valid measurement scale in quantitative and translational studies of analgesia.

The approach used to create psychometric instruments to measure pain comprises a number of discrete stages (Streiner *et al* 2008).

Phase 1 involves the initial collection of words and expressions associated with pain. Notably, while the target population conveying the initial list is frequently represented by self-reporting pain sufferers, recognition and description of behavioural signs of pain rely on independent observers in veterinary patients and humans unable to communicate (Morton *et al* 2005). A refining process is then applied to the collected words to select a list of items for possible inclusion in the instrument following specified criteria. For example, replacing expressions characterised by the same meaning but worded slightly differently with a single expression, or substituting recurrent expressions related to specific causes with generalised expressions ("rubbing ear" and "rubbing side" substituted with "rubbing painful area") (Holton *et al* 2001).

In phase 2, the refined words and expressions are validated and categorised by a pool of experts into domains (for example behavioural signs, physiological signs, and response to treatment) and associated categories (demeanour, posture, and mobility are examples of categories related to behavioural signs). An instrument is developed comprising the selected domains and categories with associated expressions, and consideration is given to layout, descriptors of expressions, and instructions for use. The resulting prototype is initially tested by a group of target respondents to ensure ease of use.

In phase 3, the instrument is tested by clinical studies to assess its psychometric properties: validity, reliability, and responsiveness.

1.4.2. Validation process

The assessment of validity is an essential part of the development of a measurement scale. Validity is "an overall evaluative judgment, founded on empirical evidence and theoretical rationales, of the adequacy and appropriateness of inferences and actions based on test scores" (Messick 1986).

This is a more specific definition, that can be translated as the effectiveness with which a test or scale measures the categories under investigation.
The validation process consists of multiple steps, which involve addressing different types of validity. Most commonly, the three types of validity tested are the content, criterion, and construct validity.

1.4.2.1. Content validity

Content validity is the prerequisite for other types of validity (Zamanzadeh *et al* 2015), as it determines the ability of an instrument to actually measure the property it is intended to measure. For example, a pain scale lacks content validity, thus lacking validity, if it demonstrates effective and reliable measurements, but fails to measure specifically pain (addressing levels of anxiety instead) (Reid *et al* 2018). As such, for a pain scale, content validity, as it relates to the ability of the instrument to measure specifically the construct it was designed to measure, can be regarded as the 'specificity' of the instrument.

In the development process of an instrument to measure pain, the first step is to convey all the items that describe pertinent aspects of this construct without including any extraneous features (such as descriptors attributable to stress or anxiety). The second step is to assess the appropriateness and completeness with which the included items fully cover all aspects of pain in the categories and sub-categories within the scale (Morton *et al* 2005).

Qualitative assessment (also called face validity) represents the simplest and traditional way to determine content validity and relies on the opinion of a panel of experts to assess the appropriateness of items within the scale (Frayers *et al* 2002). More recently, quantitative methods have been introduced in the human literature to provide evidence of content validity by computing a content validity index (CVI) (Polit *et al* 2006). With this method, the ratings provided by a pool of experts on the relevance and clarity of the items within the instrument are used to calculate an item-CVI (I-CVI) and a scale-CVI (S-CVI), which are then used by researchers to confirm, revise, or delete items. This latest approach has also been

adopted in the veterinary literature by Noble *et al* (2018) during the validation process of a feline health-related quality-of-life questionnaire.

1.4.2.2. Criterion validity

Criterion validity establishes the effectiveness of the scale's measurement comparing it with an existing gold standard (Morton *et al* 2005). This can be achieved by establishing predictive validity, that can be assessed by testing the ability of a scale to predict future change, or concurrent validity, which involves testing simultaneously a new instrument with a validated standard that measures the same concept or criterion. The correlation between the two measures is then calculated to assess how effectively the new instrument predicts the validated standard's results, with a higher correlation coefficient suggestive of higher criterion validity.

1.4.2.3. Construct validity

When a gold standard is not available, validity can be determined by testing construct validity. In a "known-groups" approach to construct validation, a hypothesis is first created, then is supported or discredited through experiment (Reid *et al* 2018). In the case of pain scales, examples of hypotheses are the prediction of changes in pain scores following administration of proven analgesics or over time following surgery, or the ability of the scale to discriminate between different severities of pain inflicted by different surgeries (Morton *et al* 2005). Construct validity can also be examined formally by the use of statistical testing (factorial validity), adopted for instance by Holton *et al* in 2001 to examine formally the internal structure of the Glasgow Composite Measure Pain Scale (GCMPS). Hierarchical agglomerative cluster analysis (factor analysis) is a statistical technique used to identify correlations between responses to the items of an instrument (for instance the sub-category expressions of the GCMPS), in order to determine whether it is possible to cluster them into smaller groups called "factors". This process of reduction of a large number of variables into a

smaller number of factors produces a factor model, which demonstrates factorial validity when it successfully describes the construct that the instrument was created to measure (Reid *et al* 2018).

1.4.3. Reliability

An instrument can be tested in clinical studies to determine whether it produces a score that is repeatable (intra-rater reliability) and reproducible (inter-rater reliability). Repeatability is demonstrated when the same score is assigned to an unchanging subject at two different time points by the same assessor, while reproducibility refers to the ability of the instrument to generate the same score when two different assessors evaluate the same subject at the same time (Streiner *et al* 2008). Alternatively, a form of reliability assessed by a statistical method called Cronbach's Alpha can be used to investigate the internal consistency of results across items of an instrument (Reid *et al* 2018). This test is considered the most appropriate to test internal consistency also in the human literature (Tavakol *et al* 2011, English *et al* 2015, Green *et al* 2016), as it establishes how closely related a set of items in the scale are as a group, and was used to assess the internal consistency of the GCMPS.

The Cronbach's Alpha coefficients obtained for each category during the validation of the GCMPS (Holton *et al* 2001), for instance, demonstrated inadequate internal consistency for only two of the categories explored, 'demeanour' and 'response to people', thus guiding the revision of these items within the questionnaire. When these two categories were combined, the measure of consistency improved and demonstrated high reliability in all the categories.

1.4.4. Responsiveness

Another important element of an evaluative scale, together with validity and reliability, is the sensitivity (also referred to as responsiveness) with which it detects change in the attribute being investigated (Kazis 1989, Wright *et al* 1997).

The level of responsiveness required from a scale depends on many factors, amongst which the construct under evaluation and its pattern of change, the patients' population, and the condition (Prasad 1996); the level of responsiveness of a scale that measures complex clinical constructs, like intelligence, depression, and pain, should also be adequate to capture the minimal amount of change that is considered clinically important by the patient or the clinician (Morton *et al* 2005).

This last concept is of particular relevance when applying a measurement tool in a clinical setting. In fact, while the 'sensitivity to change' refers to the ability of an instrument to detect "signal out of noise", the 'sensitivity to minimal clinically important difference' (MCID) (further discussed in paragraph 8.15 of this chapter) refers to the ability to detect the "smallest meaningful signal" (Prasad 1996). The ability to measure the MCID emphasizes the importance of testing responsiveness during the development of a measurement instrument for clinical use (Reid *et al* 2018).

Although responsiveness may be quantified by using several statistical indices (Wright *et al* 1997), a study conducted on the measurement of pain in infants (Barr 1998) suggested that responsiveness of a scale could be estimated by applying an intervention of known efficacy and measuring the magnitude of change. For a pain scale, the administration of an analgesic drug of known efficacy would represent an appropriate intervention, and the index of responsiveness would be derived from the difference in pain scores before and after treatment when compared to the within-subject variation.

1.4.5. Utility

Utility refers to the ease of use of the instrument without the requirement for lengthy training (Reid *et al* 2018). For self-reporting respondents, the questionnaire should be quick and easy to complete, and it should be easy to administer, score, and interpret for clinicians assessing a non-verbal patient (Reid *et al* 2018). Simplicity and the time required to complete the questionnaire are essential elements to increase the utility of the instrument for routine clinical

use. These principles, for example, were applied during the development of the short form of the GCMPS (Reid *et al* 2007).

1.4.6. Interval level measurement

The measurement in a scale may have nominal, ordinal, interval, or ratio scale properties depending on the nature of the response options to an item in the questionnaire.

If the response options to an item are binary such as yes/no, the information provided by the resulting nominal level measurement will simply indicate into which category a response falls.

Questionnaires evaluating composite constructs require a higher level of precision and usually have more complex response options, such as ordinal or continuous (Streiner *et al* 2008).

An ordinal scale, such as an NRS, possesses a higher precision of measurement, although it is discontinuous and may lack sensitivity and responsiveness if the ordered categories are broad (Reid *et al* 2018).

An interval level scale, an example of which is the graduated sight tube of a flowmeter displaying the millilitres of oxygen delivered to the patient per minute, provides continuous and more precise measurement (Morton *et al* 2005). As the change in pain intensity is assumed to lie on a continuum, it is important that its measurement possesses interval scale properties to ensure minimum loss of information and to minimise error (Morton *et al* 2005, Streiner *et al* 2008). Measuring an attribute like pain using nominal or ordinal level scales results in loss of information due to the lack of continuous measurement, and to unequal intervals between categories, thus not accurately reflecting the level of pain experienced by the patient or differences in pain scores between groups of patients (Morton *et al* 2005). Furthermore, categorical and ordinal information restricts the type of statistical tests that can be used to analyse the scores to non-parametric analysis (Morton *et al* 2005).

An interval scale can be created by applying a scaling model, of which two main types are commonly used in psychometrics. The theoretical (or subjective estimate) model is based on the subjective estimation of the appropriate weights to be assigned to descriptors given by the investigator. In the empirical (or discriminant) model, first described by Thurstone in 1928, the relationships between items in a scale are first investigated via a group of experts, then analysed statistically to derive a number (weight) for each descriptor that appropriately quantifies the category investigated. Such models determine how weights are assigned to each item and how they are then combined to produce an overall score. For instance, the concept of equal intervals between consecutive points on a scale means that "double the pain score" truly translates into "twice as painful", thus allowing two different scores to be readily interpreted and compared (Morton *et al* 2005). An empirical scaling approach based on the Thurstone model was adopted for the development of the GCMPS by Morton *et al* in 2005, and the application of these statistically derived weights to the word descriptors is one of the elements that distinguish the GCMPS from other composite measure pain scales.

Ratio scales have the properties of interval level scales, although in the former the zero score reflects the absence of the attribute. In contrast, the zero score of an interval level scale is arbitrary (Morton *et al* 2005). The sensory and emotional components of the pain experience can be conveyed by non-verbal subjects (such as animals) through expression of pain behaviours. It is therefore possible that levels of pain insufficient to cause manifestation of such behaviours are not captured by the assessor, which in turn may assign a score of zero, although this does not necessarily indicate complete absence of pain (Morton *et al* 2005).

1.5. Application of scientific methodology to the development of the Glasgow Composite Measure Pain Scale (GCMPS) for measurement of acute pain in dogs

1.5.1. Development of the first prototype scale

The first prototype of the Glasgow Composite Measure Pain Scale (GCMPS) was a multi-item behavioural pain assessment tool that was developed by Holton *et al* in 2001 using the same psychometric approach utilised for the construction of the McGill pain questionnaire.

In phase 1, 69 practicing veterinary surgeons identified 279 expressions or words associated with acute pain in the dog. A refining process was applied to the collected words applying specified criteria, examples of which are the replacement of expressions characterised by the same meaning but worded slightly differently with a single expression, or the substitution of recurrent expressions related to specific causes with generalised expressions ("rubbing ear" and "rubbing side" substituted with "rubbing painful area"). The refined list encompassed 47 expressions, 39 of which associated with behavioural and 8 with physiological signs of pain.

In phase 2, five experts in the assessment and treatment of acute pain in dogs categorised the list of expressions into nine behavioural categories and one category for physiological signs. A separate group of 75 practicing veterinary surgeons scored the expressions within the categories using a 100 mm VAS and ordered them according to associated increasing levels of pain.

The validation process consisted initially of multiple different statistical methods, including cluster analysis to test for factorial validity and Cronbach's alpha to test for internal consistency.

Subsequently, the group of five experts in pain management assessed the clinical validity of the changes to the content within categories suggested by statistical analyses. In this phase, it was decided to remove the category associated with physiological signs, due to their poor value as indicators of pain in a hospital

setting (Holton *et al* 1998b), and the category 'response to food', due to the difficulties in assessing this response in a hospital setting (Holton *et al* 2001). Also, the categories 'demeanour' and 'response to people' were combined.

The final version of the prototype comprised seven behavioural categories (posture, activity, vocalisation, attention to wound area, demeanour, mobility, and response to touch), each associated with several expressions; detailed definitions of each expression were provided to minimise subjective interpretation of descriptors. The tool was presented as a questionnaire which first involved observation of spontaneous behaviours from a distance, then assessment of interactive behaviours at rest and during specific manipulations.

1.5.2. Development of the prototype into an interval level scale

The application of an empirical scaling model to the prototype GCMPS by Morton and colleagues was designed to create interval level measurement, which retains substantial importance in quantitative clinical and research studies of analgesia, because the difference between two points on the scale can be readily interpreted and compared (Morton *et al* 2005). Furthermore, the concept of equal intervals between two consecutive items on the scale was deemed highly appropriate for the measurement of an attribute like the intensity of pain, assumed to lie on a continuum, as this approach would minimise error and loss of information (Morton *et al* 2005, Streiner *et al* 2008).

The development of the interval level GCMPS was conducted in three phases.
The first study involved the application of the Thurstone model of matched pairs (Thurstone 1928) to the GCMPS to create an interval level measurement. Sixteen veterinary surgeons were asked to assign pain intensity values to pairs of expressions within each category, presented to them in a randomised order, and the relationships between items were then analysed statistically to derive a weight for each descriptor that would quantify appropriately the items and the categories investigated.

The other two phases assessed the validity of the tool by testing its performance in a clinical context. The second study was conducted at the University of Glasgow Veterinary Hospital and included 80 dogs divided in four groups: 20 dogs underwent orthopaedic surgery, 20 dogs soft tissue surgery, 20 dogs were hospitalised for medical conditions, and 20 dogs (owned by staff of the University) represented the control group, being judged clinically normal. Pain was scored in the enrolled participants by five veterinary surgeons with experience of veterinary practice, who were postgraduate students at the University of Glasgow Veterinary School. The five scorers, who were not familiar with any of the dogs enrolled and did not participate in the development of the GCMPS, assessed each dog independently of each other and were individually explained the examination procedure and the use of the scale before making their assessments. The observers were unaware of the treatment allocation of the dogs, and they were not provided with the definitions of the scale items. In this study, the test constructs used to assess validity were the presence or absence of surgical intervention, the group in which the patient was enrolled, and the perceived severity of pain associated with the surgical procedures or medical conditions.

The last study was designed to further test validity and to assess sensitivity of the scale. In this phase, 77 dogs (different from those enrolled in the second study) that had undergone orthopaedic or soft tissue surgery at the University of Glasgow Veterinary Hospital were included. Concurrent criterion validity was assessed by comparing pain scores assigned in the postoperative period by one observer (the first author of the paper) with the use of the prototype first and then an 11-point NRS. Further construct validity was assessed by testing the hypotheses that post-surgical pain would decrease with time and that orthopaedic surgery would be associated with higher pain scores than soft tissue surgery.

Results of the various statistical methods utilised during the clinical phases of this study indicated that the methodology used in the design of the GCMPS (Fig. 1.4) supported its face validity, content validity, and responsiveness, reliably measuring pain in the clinical context in which it was tested. The creation and validation of a scale to measure pain in dogs that provided continuous, interval level measurement as described by Morton and colleagues represent the first work using this approach in the veterinary literature.

The Glasgow Composite Measure Pain Scale

The questionnaire is made up of a number of sections each of which have several possible answers. Please tick the answer that you feel is appropriate to the dog you are assessing. Approach the kennel, ensure you are not wearing a laboratory coat or theatre 'greens' as the dog may associate these with stress and/or pain. While you approach the kennel look at the dog's behaviour and reactions. From outside the dog's kennel look at the dog's behaviour and answer the following questions.

Look at the dog's posture, does it seem...
- Rigid
- Hunched or Tense
- Neither of these

Does the dog seem to be...
- Restless
- Comfortable

If the dog is vocalising is it...
- Crying or Whimpering
- Groaning
- Screaming
- Not vocalising/none of these

If the dog is paying attention to its wound is it...
- Chewing
- Licking or Looking or Rubbing
- Ignoring its wound

Now approach the kennel door and call the dog's name. Then open the door and encourage the dog to come to you. From the dog's reaction to you and behaviours when you were watching him/her assess his/her character.

Does the dog seem to be...
- Aggressive
- Depressed
- Disinterested
- Nervous or Anxious or Fearful
- Quiet or Indifferent
- Happy and Content
- Happy and Bouncy

Now look at the dog's response to stimuli. If the mobility assessment is possible then open the kennel and put a lead on the dog. If the animal is sitting down encourage it to stand and then come out of the kennel. Walk slowly up and down the area outside the kennel. If the dog was standing up in the kennel and has undergone a procedure which may be painful in the perianal area, ask the animal to sit down.

During this procedure did the dog seem to be...
- Stiff
- Slow or Reluctant to rise or sit
- Lame
- None of these
- Assessment not carried out

The next procedure is to assess the dog's response to touch. If the animal has a wound, apply gentle pressure to the wound using two fingers in an area approx. 2 inches around it. If the position of the wound is such that it is impossible to touch, then apply the pressure to the closest point to the wound. If there is no wound then apply the same pressure to the stifle and surrounding area.

When touched did the dog...
- Cry
- Flinch
- Snap
- Growl or Guard wound
- None of these

Fig. 1.4 GCMPS. Reprinted from Reid *et al* 2007.

1.5.3. Development of the Short Form of the GCMPS (GCMPS-SF) from the interval scale prototype and derivation of an intervention score for provision of rescue analgesia

The GCMPS-SF (Reid *et al* 2007) was developed by refining the interval scale designed by Morton *et al* in 2005 to create a 'user friendly' questionnaire with cut-off points for provision of additional analgesia, with the principal driving aim to improve the usefulness of the instrument for routine clinical use. To fulfil this aim, the authors identified in the modification of the length of the scale and the derivation of an intervention level linked to the pain score the two key objectives to improve it. In fact, the length of the GCMPS was a determinant limiting its applicability in a busy practice environment, thus underlining for the need of a less pleonastic, simple to use questionnaire that would include fewer steps to complete. Furthermore, it was considered that the usefulness of the pain measurement instrument would be markedly improved if the pain score was linked to an intervention score associated with the requirement for provision of additional analgesia.

1.5.3.1. Development of a short form of the GCMPS

In order to shorten the questionnaire with the primary aim to reduce the time taken for completion, the following strategy was adopted. According to the authors' clinical judgement and/or feedback from more than 500 practicing veterinary surgeons, the categories and items within the scale were first reviewed in an attempt to reduce them where possible, and the number of items within each category was then balanced by combining or splitting the associated word descriptors where appropriate. No details on selection criteria for veterinarians and response types were given in this study.

In terms of measurement properties, the questionnaire was converted to an ordinal scale by ranking the items numerically within each category in accordance with their related pain severity. The final layout of the questionnaire was

reconfigured to improve its utility and comprised three sections encompassing a total of six behavioural categories, each associated with descriptive expressions (items), and instructions for completion were provided at the beginning of each section (Fig. 1.5).
In detail, section A comprises two behavioural categories (vocalisation and attention to wound/painful area) with 4 to 5 associated descriptors each and relates to the observation of the patient from a distance. Section B and C involve interaction with the patient: the former is dedicated to mobility, and it might not be carried out in case of spinal, pelvic limb or multiple limb fractures, where assistance is required for locomotion or when drug intervention impedes this assessment (for example epidural injection of local anaesthetics); the latter is response to touch and entails gentle palpation of an area approximately 5 centimetres around the site. Section D comprises two behavioural categories, posture and activity, to try and capture the overall attitude of the patient in respect to its surroundings, to stimulation, and to body language. Descriptors for each category are placed in an increasing order of pain intensity and weighed accordingly. The maximum total cumulative score can be either 24, when all the categories can be assessed, or 20, when mobility assessment cannot be carried out.

Conversion of the scale from interval to ordinal in nature, obtained by substituting a rank number for the calculated weight, implied a decrease in level of precision of the instrument. In fact, changing a scale from interval, where the level of pain can potentially assume every value within the scale (as it lies on a continuum) to ordinal, thus assigning defined scores to descriptors ordered within pre-set categories, inevitably decreases the level of precision with which a pain response can be quantified. Estimation of this reduction would be possible by comparing the pain scores assigned by a single observer using the GCMPS as designed by Morton *et al* (2005) and the GCMPS-SF as designed by Reid *et al* (2007) and evaluating on a large sample of dogs the mean difference in pain scores assigned to the same dogs at the same time points with the two scales. However, this calculation was not done, so the decrease in the level of precision cannot be quantified. Nevertheless, while more precise measurement is required for research purposes, the level of precision achieved with an ordinal scale was deemed adequate for clinical purposes (Reid *et al* 2007).

The underlying rationale of balancing the number of items within each category was dictated by an attempt to minimise bias. In the interval level GCMPS, the category ‘demeanour’ comprises seven descriptors, which would have been assigned a score from 0 to 6 with the introduction of the ranking system. By comparison, the category ‘comfort’ encompasses two descriptors, which would have ranked from 0 to one. Consequently, the category ‘demeanour’ would have had a significant greater weight than ‘comfort’ in the final score, although it is not established whether it actually retains a greater importance in the expression and measurement of pain (Holton 2000).

Enhancement of simplicity and ease of use was not only achieved through the process elucidated above, but also by omitting the reference to the list of definitions from the final design of the questionnaire. Despite the lack of clear and specific definitions might translate into an increase in inter-observer variability, the authors decided to omit them with the consideration that all words and expressions were of common use and had dictionary definitions (see Appendix 1 for the rationale of this omission as stated by the authors).

SHORT **F**ORM OF THE **G**LASGOW **C**OMPOSITE **P**AIN **S**CALE

Dog's name _____

Hospital Number _____ **Date** / / **Time**

Surgery Yes/No (delete as appropriate)

Procedure or Condition_____

_____

In the sections below please circle the appropriate score in each list and sum these to give the total score.

A. Look at dog in Kennel

Is the dog?

(i)		(ii)	
Quiet	0	Ignoring any wound or painful area	0
Crying or whimpering	1	Looking at wound or painful area	1
Groaning	2	Licking wound or painful area	2
Screaming	3	Rubbing wound or painful area	3
		Chewing wound or painful area	4

In the case of spinal, pelvic or multiple limb fractures, or where assistance is required to aid locomotion do not carry out section **B** and proceed to **C**
Please tick if this is the case ☐ then proceed to C.

B. Put lead on dog and lead out of the kennel.

When the dog rises/walks is it?

(iii)	
Normal	0
Lame	1
Slow or reluctant	2
Stiff	3
It refuses to move	4

C. If it has a wound or painful area including abdomen, apply gentle pressure 2 inches round the site.

Does it?

(iv)	
Do nothing	0
Look round	1
Flinch	2
Growl or guard area	3
Snap	4
Cry	5

D. Overall

Is the dog? (v)		*Is the dog?* (vi)	
Happy and content or happy and bouncy	0	Comfortable	0
Quiet	1	Unsettled	1
Indifferent or non-responsive to surroundings	2	Restless	2
Nervous or anxious or fearful	3	Hunched or tense	3
Depressed or non-responsive to stimulation	4	Rigid	4

Total Score (i+ii+iii+iv+v+vi) = _____

Fig. 1.5 GCMPS-SF. Reprinted from Reid *et al* 2007.

1.5.3.2. Derivation of an intervention level for provision of additional analgesia

The intervention level was defined as the pain score at which a dog would display sufficient pain behaviours to be judged in need of analgesic therapy by the assessing clinician.

A hundred dogs that had undergone orthopaedic or soft tissue surgical procedures at three different teaching referral hospitals (43 at University College Dublin, 43 at the University of Glasgow, and 36 at North Carolina State University) were included in this analysis with no restrictions on age, sex, breed, type of surgery, and anaesthetic protocol. Dogs were pain scored in the post-operative period by the veterinary surgeon carrying out routine postoperative examinations, thus encompassing multiple observers. Pain scorers were instructed to first complete the GCMPS-SF and then establish whether the patient needed analgesic treatment based on their expertise and clinical judgement. Descriptive statistics expressed as mean ± SD, median, range and interquartile ranges were initially used to gather information on how the pain scores differed between dogs that did and did not require analgesia (analgesia groups) as judged by the veterinary surgeon. To define the intervention level linear discriminant analysis, a linear model for classification and dimensionality reduction, was then used to identify the pain score that would include the maximum possible number of dogs in the correct analgesia group as allocated by the clinician. This analysis was separately conducted to derive two intervention scores, one indicative of requirements for additional analgesia when all the categories are assessed (maximum cumulative score of 24), and one when mobility assessment cannot be carried out (maximum cumulative score of 20). Results of this analysis produced an analgesic intervention level of 6/24 or higher, with 84% of dogs correctly classified in their analgesic-requirement group and misclassification rates of only 16%. When the same analysis was conducted to derive the intervention level for a possible maximum score of 20, it produced a cut-off score of 5/20 or higher.

1.5.3.3. Validation

As the Short Form was derived from the GCMPS validated in 2005 without addition of any new items, it retained the content validity of the original scale. Construct validity was established through a 'known-groups approach' by testing the hypothesis that pain scores would differ between dogs requiring or not requiring analgesia. Results of the field study showed that the median pain score for dogs requiring analgesia was 7, whereas it was 3 for those not requiring analgesic treatment, and that this pattern was consistent across all the three hospitals. Moreover, as the clinical study was carried out in a mixed population of dogs, undergoing a variety of surgical procedures, and pain scored by multiple observers, these results supported construct validity, reliability, and utility of the simplified questionnaire.

More recently, a pilot study conducted by Tait *et al* (2011) in dogs with painful medical and surgical conditions investigated the changes in pain scores assigned with the GCMPS-SF before and after administration of analgesic treatment compared to the clinician's perception of the change in the level of pain. Results of this investigation supported construct validity and responsiveness of the scale, providing further evidence that its validity for the measurement of acute pain in dogs was not limited to acute post-operative pain.

To retain the validity of the scale, it should be used as it was originally described and validated, thus preserving its measurement properties. To stress the importance of this concept, a few studies have been conducted to ensure the validity of the metrology instrument was preserved also when used in different contexts and in a different language. Murrell *et al* (2008) tested and validated a modified version of the scale in a veterinary teaching hospital in the Netherlands, with a maximum total score of 10, with the consideration that it was applied to a different clinical environment (implying a different surgical case load and different analgesic interventions adopted), and where English is not the first language. The clarity, intelligibility, and appropriateness of the categories and related descriptors were considered crucial to retain the conceptual content of the scale also in a work carried out by Della Rocca *et al* in 2018, where an accurate

process of linguistic validation was undertaken to create and validate the Italian version of the GCMP-SF (ICMPS-SF).

1.6. Effects of confounding factors on the use of behavioural pain scoring systems

Concerns have been raised by many researchers across the veterinary literature about the influence of factors other than pain on the final score assigned to a patient using various behavioural pain scoring systems, amongst which species and temperament (Mathews *et al* 2014, Ijichi *et al* 2014, Buisman *et al* 2017, Lush *et al* 2018, Elwood *et al* 2022), the effects of sedative and analgesic drugs (Guillot *et al* 2011, Rialland *et al* 2012, Buisman *et al* 2016), and the number and experience of the evaluators assessing patients (Holton *et al* 1998a, Carsten *et al* 2008, Guillot *et al* 2011, Barletta *et al* 2016, Hofmeister *et al* 2018). No studies have been conducted yet on the influence of age and dog breed on postoperative behavioural expressions of pain, although these two factors might influence pain behaviours (Mathews *et al* 2014) and breed was perceived by dog owners and veterinarians to play a role in the sensitivity to painful stimuli (Gruen *et al* 2020).

The patient's demeanour, for example, has been shown to have the potential to remarkably influence the score assigned to some categories, increasing the final score assigned to the patient. This aspect was evaluated in a study conducted in cats (Buisman *et al* 2017), where post-operative pain was scored with two validated scales, the UNESP-Botucatu multidimensional pain scale and the revised Composite Measure Pain Scale – Feline for acute pain in cats. The authors observed a strong negative correlation between demeanour and eating behaviour in a hospital setting during the post-operative period, and that high pre-operative demeanour scores (shy/aggressive) could significantly increase post-operative pain scores if demeanour was not accounted for during pain assessment.

Other studies investigated the effects on pain scores assigned by multiple observers using unidimensional behavioural pain scoring systems. Holton *et al* in 1998 (1998a) compared the use of three different pain scoring systems (SDS, NRS

and VAS) in 50 dogs recovering from surgery. A maximum of four veterinarians pain scored the dogs at different time points after surgery, and results demonstrated a significant variability among observers with the use of all three scales. The authors concluded that comparative analysis of pain score data obtained in analgesia studies must take into account observer variability when more than one assessor is used.

1.7. Effects of confounding factors and potential sources of bias on the GCMPS-SF scores

The GCMPS-SF is still amongst the only few validated scales to assess acute post-operative pain in dogs and importantly it is the only scale at present linked to an intervention level for provision of additional analgesia. Despite the marked increase in accuracy and reliability compared to other behavioural pain scales due to the scientific approach adopted during its development, there are still potential sources of bias and confounding factors that could play a role in the interpretation of the patient's pain behaviour, thus altering the final score.

1.7.1. Observer-related factors

Observer-related factors have been investigated in multiple studies, which highlighted the potential for the experience of the individual and the number of observers to influence the final score, thus introducing a source of bias.

1.7.1.1. Experience of the individual

A study conducted by Barletta *et al* (2016) investigated the effects of experience on pain assessment by comparing scores allotted by first- and second- year veterinary students without training and experienced anaesthesiologists. All assessors were asked to watch 90-second videos and to score pain using the DIVAS and the GCMPS-SF in 13 client-owned dogs that had undergone a variety of surgical

procedures. Significant differences emerged between scores assigned by the two groups of assessors, with a tendency for students to assign higher pain scores in dogs that were deemed less painful by anaesthesiologists and lower pain scores to patients that were given higher scores by anaesthesiologists. The authors postulated that these differences, although possibly associated with pain assessment performed on videos, thus being related to the lack of interaction with the animal and to the quality/duration of the videotapes, could also reflect differences in training and ability to interpret dogs' behaviour, especially considering the significant differences between groups.

1.7.1.2. Number of observers

The agreement between multiple evaluators performing pain assessment has also been object of various studies comparing pain scores assigned with the simultaneous use of different behavioural scales (Carsten *et al* 2008, Guillot *et al* 2011, Hofmeister *et al* 2018).

In the study conducted by Carsten and colleagues, two trained observers used the GCMPS-SF and the VAS to score acute pain in dogs undergoing radiation therapy, while pain induced by bone marrow aspirates (Guillot *et al* 2011) was assessed with the use of two indices (an inactivity index (IAI) and a normal behavioural index (NBI), constructed from automated video analysis) and two scales (the GCMPS-SF and the 4A-VET) by two veterinary surgeons with different experience in behavioural pain assessment. Despite demonstrating the best inter-rater reliability among the pain assessment tools compared, some degree of variation was also observed with the use of the GCMPS-SF in both studies.

Recently, Hofmeister *et al* (2018) tested the agreement amongst six experienced anaesthesiologists when asked to pain score videos of 31 post-operative dogs using three different assessment tools – the VAS, the NRS and the GCMPS-SF – all together and at a three, six-, and nine-month interval separately. Findings of this study suggested that intra- and inter-observer variability was fair to excellent for all the scales, with the best agreement amongst evaluators achieved with the GCMPS-SF.

1.7.2. Patient-related factors

1.7.2.1. Anxiety

During the study conducted by Hofmeister *et al* in 2018 discussed in the previous paragraph, concerns were raised in relation to some descriptors within the GCMPS-SF which did not manage to capture accurately the behaviour of the patient under evaluation and some other descriptors that were not considered signs of pain by many evaluators. Notably, as pointed out in the discussion of this paper, "nervous or anxious or fearful" adds 3 points to the total score of the GCMPS-SF, but it was not considered a relevant indicator of pain by the anaesthesiologists participating in this study.

The same concern was raised in an observational clinical trial conducted on seven dogs undergoing curative intent radiation therapy for neoplasia of the forelimb (Carsten *et al* 2008) which aimed to investigate the correlation between the degree of skin damage and pain scores assigned with the VAS and the GCMPS-SF. The GCMPS-SF scores were noted to be high at the beginning of the treatment, with a progressive decrease over the first days until radiation fraction number six, and these initially higher pain scores were not in agreement with the level of pain assigned with other assessment methods, like physical examination and VAS. The authors noted the same pattern in the levels of anxiety and nervousness, which subsided as the dogs were acclimatising to the daily routine at the radiotherapy facility, and attributed the initially higher GCMPS-SF scores to these behavioural components unrelated to pain.

The role of pre-operative and post-operative anxiety on the total GCMPS-SF score was specifically investigated in a recent work conducted by Ellwood *et al* (2022). Eighteen dogs undergoing surgical management of stifle joint disease were evaluated pre- and post-operatively for pain, anxiety, and sedation scores. Anxiety was assessed with the use of a behaviour-based scoring system (REF, Reactivity Evaluation Form) and a VAS, sedation was scored with a 0-3 SDS, and pain was evaluated with the GCMPS-SF. Dogs were divided into groups based on their baseline anxiety scores, and all the assessments were carried out simultaneously by the same observer. Overall, there was a significant difference

between median baseline and post-operative pain scores [2 (0-3) *vs* 3 (2-5), respectively, *p*=0.0032], although none of the scores reached the threshold for administration of additional analgesia. When divided into groups based on anxiety scores determined either with the REF or the VAS, no significant correlations between groups were found on any of the following relationships investigated: pre-operative and post-operative pain scores between groups, pre-operative anxiety and pain scores, pre-operative anxiety and post-operative pain scores, and pre-operative anxiety scores and the change in pain scores. The authors concluded that, despite other factors interfering with pain assessment in a hospital environment, baseline anxiety did not seem to have a relationship with pain scores assigned with the use of the GCMPS-SF.

1.7.2.2. Temperament

The influence of personality on behavioural expressions of pain has been investigated in multiples species, included humans (Harskin *et al* 1989, Ramirez-Maestre 2004, Soriano *et al* 2012), horses (Ijichi *et al* 2014), cats (Buisman *et al* 2017), and dogs (Lush *et al* 2018), and strong positive correlations have been found.

The influence of personality is being discussed separately from anxiety in the present thesis, as they represent distinct traits of an individual behavioural expression. Personality is defined as 'individual differences in behaviour that are consistent over time and across contexts' (Koolhaas *et al* 1999), while anxiety is a 'response in anticipation of a specific stimulus or situation' (Ellwood *et al* 2022), thus being time and context sensitive.

Studies investigating the relationships between personality and pain expression in humans found that extrovert and neurotic people express their experiences of pain remarkably different (Lush *et al* 2018). Neuroticism has been associated with a high emotional stress response to pain and a low degree of emotional stability (Lush *et al* 2018), whereas extroversion is characterised by a clear expression of the pain experience, even when this is less intense (Harskin *et al* 1989, Ramirez-Maestre 2004, Soriano *et al* 2012).

In horses, Ijichi *et al* (2014) provided evidence of a positive correlation between extroversion and behavioural expressions of pain, and between neuroticism and reduced tolerance to pain.

Correlations between extroversion and neuroticism with behavioural expressions of pain in dogs were explored in a study conducted by Lush *et al* in 2018. Seventeen male dogs scheduled for a single standardised elective procedure (castration) were enrolled in the study, that was conducted in two veterinary practices in the UK. Dogs' personality traits 'extroversion' and 'neuroticism' were measured using the validated Monash Canine Personality Questionnaire – Revised (MCPQ-R), which characterises extrovert dogs as active, excitable, and restless, while neurotic dogs as fearful, submissive, and timid, and these traits were compared with post-operative pain scores assigned with the GCMPS-SF. A single observer, blind to individual personality scores, retrospectively scored all the patients from 3-minute videotapes; as this modality excluded the possibility of interaction with the patient, the maximum possible total score with the GCMPS-SF was 15. Results from this study demonstrated a strong positive correlation between extroversion and peak pain scores (Spearman: $r_s = 0.558$, $p = 0.031$), while no correlation was found between pain scores and neuroticism (Spearman: $r_s = 0.107$, $p = 0.703$). The authors concluded that extrovert personality was associated with more prominent behavioural indicators of pain, as dogs with higher scores for extroversion were assigned higher peak GCMPS-SF scores, despite all dogs had a similar degree of tissue trauma. However, analysis of mobility and palpation of wound/painful area of the GCMPS-SF were not carried out in this study, and it is therefore possible that the lack of interaction with the patient accentuated the effect of personality on pain assessment (Ellwood *et al* 2022).

1.7.3. The effect of sedative/analgesic drugs

The effects of sedation on the performance of the GCMPS-SF have been evaluated in the veterinary literature and have produced contrasting results.

Guillot *et al* (2011) conducted a pilot experimental study comparing two pain scales (the GCMPS-SF and the 4A-VET) and two indices [an inactivity index (IAI) and a normal behavioural index (NBI), both constructed from automated video analysis] in 16 healthy beagle dogs undergoing bone marrow aspiration. Dogs were divided into groups based on the puncture site [sternal (stern) or iliac crest (iliac) bone] and on the administration of sedative and analgesic medications before the procedure (a combination of medetomidine and hydromorphone) or no medications (sed and no-sed groups); all dogs were administered deracoxib in the periprocedural period. Levels of sedation and pain were scored at baseline (before procedure and sedation), after 20 and 50 minutes, and 24 hours post-procedure. Pain scores increased from baseline at the first two time points in all groups with the use of both pain scales, confirming their responsiveness in distinguishing between different levels of pain. No correlation was observed between pain scores assigned with the 4A-VET and sedation scores, and both the indices utilised were effective in differentiating sedated from non-sedated dogs. In contrast, the GCMPS-SF scores did appear markedly affected by the medications administered, with the stern-sed group being assigned significantly higher scores than the stern-no-sed group. This represents a somewhat surprising finding, as the influence of sedation is expected to blunt behavioural responses, thus potentially reducing pain scores assigned with a pain behavioural assessment tool. In fact, especially at the first post-procedural time-point assessments, the influence of sedation, diminishing the level of consciousness and altering motor function responses (Pereira-Morales *et al* 2018), will tend to prevent/minimise both spontaneous and evoked behaviours. However, sedation scores achieved in this study were not reported for any group or time point, thus making possible inferences on the correlation between pain and sedation speculative.

Different findings emerged from a previous study (Murrell *et al* 2008) whose aim was to test and validate a modified form of the GCMPS-SF in a veterinary teaching hospital in the Netherlands. This research was conducted in 60 dogs undergoing a variety of orthopaedic and soft tissue surgeries, and included patients with multiple ASA status. In this study, assessments were completed by one observer with the use of the modified GCMPS-SF for pain scores and a SDS for sedation scores, with the first post-operative pain assessment carried out at six hours postoperatively. Although 27% of dogs with an ASA status of 3 or 4 were still

showing mild to moderate degrees of sedation at the first pain assessment, no significant correlation was detected between pain scores and sedation scores. Furthermore, dogs with an ASA status of 3 or higher might have had depressed mentation as a result of the systemic condition or disease, yet none of these effects did confound pain assessment (Murrell *et al* 2008). In the authors' opinion, these findings suggested that the modified GCMPS-SF, despite being constituted by several behavioural descriptors that may be influenced by sedation, was effective in differentiating pain from other factors potentially affecting the dogs' behaviour, and provided evidence of the content validity of the pain scale.

1.8. The GCMPS-SF and acute pain study methodology

Since its development, the GCMPS-SF has had a robust and consolidated impact in numerous different contexts. Being freely downloadable from the internet, thousands of downloads have been tracked both in veterinary practice and industry. However, once downloaded the questionnaire can be copied so the actual total usage is likely to be underestimated.

It has been widely adopted as a clinical standard for measurement of acute pain in dogs (Calvo *et al* 2014).

It has also been utilised in industry trials by pharmaceutical companies to test the efficacy and obtain approval of new drugs. Some examples of trials that based their results on the use of the GCMPS-SF are Merial Ltd, which obtained approval from the US Food and Drug Administration (FDA) for "Previcox" (firocoxib) chewable tablets in dogs following orthopaedic surgery in 2008; Novartis Animal Health US Inc had approval from the FDA for "Deramaxx" (deracoxib) chewable tablets in dogs following orthopaedic surgery in 2011; "Recuvyra" (fentanyl transdermal solution) obtained approval from the European Medicines Agency in dogs for orthopaedic procedures in a study conducted by Elanco Animal Health in 2011, and from the FDA in dogs for soft tissue surgery in a large trial conducted by Nexycon Pharmaceutical Inc in 2012.

Finally, it has been widely adopted to measure pain in research clinical trials investigating the effect of drugs and interventions on perioperative acute pain.

Research clinical studies may be complex and challenging to conduct, requiring careful consideration of multiple factors such as group sizes, statistical power, control groups, pain measurement instruments, rescue analgesic provision, and data analysis (Hofmeister *et al* 2007, Slingsby 2010).

In the following sections we first consider the statistical basis upon which clinical trials are built, such as hypothesis testing, statistical power, significant level, power analysis, and confidence intervals. We then consider how trial design may influence these factors hence influencing the probability of drawing correct/erroneous conclusions.

1.8.1. Hypothesis testing

Hypothesis testing is a formal statistical procedure used to evaluate the strength of evidence provided by the data collected in order to establish how reliably the observed findings in the sample investigated can be extrapolated to the larger population the sample was drawn from (Davis *et al* 2006). The first step in hypothesis testing is the formulation of a specific hypothesis, which must be stated in the form of a null hypothesis (H_0) and an alternative hypothesis (H_1). The null hypothesis is the prediction of no difference between groups or no relationship between variables, while the alternative hypothesis represents the initial research hypothesis of a difference between groups or a correlation between variables. Hypothesis testing always starts from the assumption that the null hypothesis is true. Data collected are then treated statistically to assess the likelihood of obtaining the study's results under this assumption, and the outcomes of the statistical tests determine with reasonable probability whether the null hypothesis can be rejected in favour of the alternative hypothesis.

1.8.2. Statistical power

The statistical power, or sensitivity, represents the likelihood of detecting a true effect in the population investigated if there is one. Having enough statistical

power positively correlates with the probability of drawing accurate conclusions about results of a study, as a higher power indicates a higher probability of detecting a true effect and, consequently, a lower risk of a false negative result, referred to as Type II statistical error, or β-error. Statistical power is usually set at 80% or higher, thus reflecting the probability of 80% or higher that the statistical tests applied will detect a true existing effect, and can be expressed as 1-β. In other words, researchers accept a probability of 20% or lower that the study will not detect a true difference between groups or an existing correlation between variables. On one hand, a low power means that statistical tests won't be sensitive enough to detect a true effect at all; on the other hand, overly increasing the power increases the sensitivity to very small effects, that may be statistically significant but not clinically relevant.

1.8.3. Power analysis and sample size estimation

Sample size estimation is used to determine the minimum number of patients needed in a study to detect a predetermined effect of defined magnitude. Power analysis is a calculation that can be used to estimate the sample size, and is constituted by four components:

- Statistical power: the likelihood of detecting a true effect if there is one.
- Sample size: the minimum number of patients needed in a study to be able to detect an effect if there is one.
- Significant level (alpha): the likelihood of finding an effect when there is no actual true effect.
- Effect size: the magnitude of the expected results, often based on similar studies or a pilot study, that is considered of relevance.

Power analysis allows calculation of any of these variables when the other three are known or estimated, thus offering different perspectives when used in clinical research. An *a priori* sample size calculation is used during the planning stage of a research project to determine the minimum number of patients required for a

set power and level of significance to detect the predetermined effect size. This calculation reasonably ensures that the number of subjects enrolled confers enough power to the study to draw accurate conclusions should no statistical difference be found between groups or no correlation between variables (Hofmeister *et al* 2007). Increases in the sample size will confer higher power to the study, although not indefinitely: over-increasing the sample size will add only marginally increases to the power, at the expense of increased time, costs, and, especially in experimental trials, increase number of subjects unnecessarily enrolled. *Post hoc* power analysis is a retrospective analysis that might be utilised to estimate the power achieved by a study based on the sample size used and the effect size detected, and can therefore provide information on the likelihood that the lack of significant difference between groups or lack of correlation between variables may be imputable to insufficient power (Hofmeister *et al* 2007). In other words, *a priori* power analysis can be used to determine the sample size, while *post hoc* power analysis can be used to determine the power achieved.

1.8.4. Significance level (alpha) and *p*-value

The significance level, or alpha (α), is the threshold for statistical significance established during the planning stage of a study and represents the maximum risk researchers are willing to take to reject a true null hypothesis, which would generate a false positive result (Type I statistical error, or α-error). The significance level is commonly set at 1 (or 0.01) or 5% (or 0.05), thus reflecting the probability of 1% or 5%, respectively, that the results are obtained merely by chance under the null hypothesis; differently worded, it defines the Type I statistical error rate.

Statistical tests applied to the research data produce the *p* (probability) value, a quantitative measure that is compared to the pre-set significance level and provides information about the statistical significance of a finding to estimate the probability of rejecting the null hypothesis:

- A *p* value equal or higher than the significance level (≥ 0.01 or ≥ 0.05) indicates that the results lack statistical significance, and the null hypothesis is not rejected.
- A *p* value lower than the significant level (< 0.01 or < 0.05) indicates that the results are statistically significant, and the null hypothesis is rejected. The lower the *p* value, the higher the statistical significance, because it reflects a decreased likelihood of occurrence of a difference between groups or a correlation between variables if there were no true effect.

Hypothesis testing can never prove the null hypothesis, because a lack of statistical significance does not necessarily mean that absolutely no effect exists; however, the choice of the significance level will provide the researcher with relevant information on how to interpret the study results in terms of the probability of rejecting the null hypothesis in favour of the alternative hypothesis. A more stringent significance level implies that an effect has to be larger (more meaningful) to be considered statistically significant, whereas increasing the significance level increases the probability of finding a statistically significant difference or correlation but at the expenses of accuracy, as it increases the likelihood of an effect occurring merely by chance.

1.8.5. Confidence Interval (CI)

In clinical research, results of summary or test statistics represent an estimate, because data analysed are collected from a sample out of the population of interest. The confidence interval (CI) provides an indication of the degree of uncertainty around the estimate, as it defines the probability for a variable under investigation to fall between a range of values should the experiment be repeated or the population re-sampled in the same way. The confidence level expresses the percentage of times the estimate is expected to fall between the lower and upper limit of the range constituted by the CI, and it is set by the alpha value ($1 - \alpha$). Thus, if the alpha value is set at 0.05 as the threshold for statistical significance, the confidence level is 0.95, or 95%, meaning that there is a 95% probability that the value of a studied parameter lies within this range (Fig. 1.6).

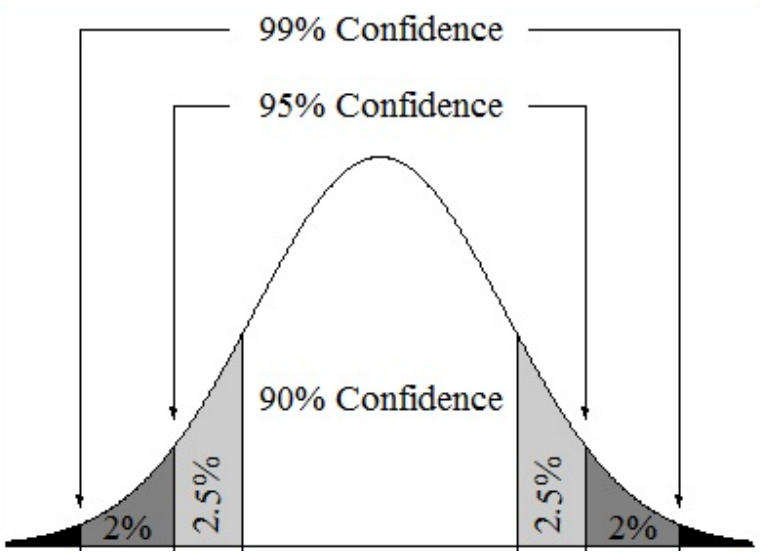

Fig. 1.6 Example of different values of confidence level determining the width of the CI. Reprinted and adapted from www.365datascience.com.

1.8.6. Group sizes

Unequal group sizes may be found in research studies and may be the result of the randomisation technique, planned imbalances between groups or study dropouts.

On one hand, simple randomisation reduces remarkably the potential for selection bias, as it ensures complete randomness in group allocation; on the other hand, while this technique generates group sizes indicative of random variation (Schulz *et al* 2002), it may also lead to inequalities in the number of subjects allocated to each group.

Reasons to plan imbalances between groups may include accessibility problems, financial costs, and differences in variability between groups. In a large clinical trial investigating the efficacy of a new drug on postoperative pain inflicted by a standardised procedure, for example, the control group, that has received an intervention of known efficacy, retains a lower variability compared to study group in which the effects of the intervention are yet to be determined; in this scenario, it is not uncommon to set a ratio between the study and the control group of 2:1 (Wagner *et al* 2008, Gruet *et al* 2013).

Study dropouts are frequently encountered in analgesia studies for a number of different reasons, of which some examples are listed below:

- Human errors: wrong medication administered in the pre- or post-operative period, incomplete recording of the study variables.
- Patient factors: unexpected side effects that warrant patient withdrawn from the study, additional patient analgesic requirements that no longer meet with the inclusion criteria.
- Technical glitches: malfunction of computers, software, or analysers used to process samples if objective measures of pain are investigated.

Irrespective of the cause, unequal group sizes may lead to a decrease in statistical power, with a positive correlation between the magnitude of inequality between groups and the decrease in statistical power, and to a higher rate of Type I statistical error, also positively correlated with the magnitude of group size imbalance (Rusticus *et al* 2014). Furthermore, dropouts may lead to missing data

completely at random or *at random* (Bhaskaran *et al* 2014), which, despite the possible confusing terminology, represent distinct scenarios. Missing data completely at random means less data included in statistical analysis due to exclusion of similar proportions of participants from all study arms, thus affecting the overall power of the study but not creating a bias in the recorded data. In contrast, missing data at random refers to the disparity in the distribution of missing data between groups, thus introducing an uneven variability that must be taken into account when implementing imputation techniques to treat the data in an unbiased and statistically valid way (Bhaskaran *et al* 2014).

While a discrepancy in group sizes *per se* does not necessarily represent a problem (Schulz *et al* 2002), as the likelihood of predisposing to Type I and Type II statistical error increases in a linear fashion with the magnitude of group size imbalance (Rusticus *et al* 2014), uncertainty remains on the difference in group sizes that is considered sufficient to influence the validity of the study outcomes (Keppel *et al* 2004).

1.8.7. Variability

Amongst the many factors that can affect the ability of acute pain clinical trials to detect a meaningful effect and producing valid results (referred to as the 'sensitivity' of the trial), researchers must consider the various sources of variability pertaining to the patient and the procedure used to assess and quantify pain when planning a study.

Patient's factors include age, ongoing chronic painful conditions, previous treatments, psychological distress, the severity of tissue trauma (Gilron *et al* 2019), breed (Mathews *et al* 2014), and individual variations. To this respect, a considerable decrease in variability would be achieved with a matched case-control study design, as it eliminates confounding (Rose *et al* 2009). Matching in case-control studies is frequently utilised in public health and medical literature (Rose *et al* 2009, Stuart 2010), and although it can be encountered in veterinary research (examples of which are Alford *et al* 2001, Bartlett *et al* 2010, Robinson *et al* 2020), it is not common in the veterinary pain literature.

The vast majority of clinical trials investigating treatments of acute pain have been conducted on acute postoperative pain induced by a surgical procedure, both in human (Gilron *et al* 2019) and animal research (Hansen 2003). The utilisation of a single standardised procedure performed in healthy, relatively homogeneous populations, examples of which are molar extraction in human (Gilron *et al* 2019) or ovariohysterectomy in canine patients (Hansen 2003), has been found advantageous in terms of high sensitivity and reproducibility of findings, due to standardised conditions that reduce variability, and accelerate timelines, due to the routine and ubiquitarian implementation of the selected procedures. However, a potential implication of this study methodology concerns the external validity of trials, because the restrictive criteria applied to reduce patient and environmental variability may negatively impact the generalisability of findings, thus limiting their relevance in a broader context (Gilron *et al* 2019).

Both patient and environmental variability are negatively correlated with the power of a study, as power decreases with increasing levels of variation (Rusticus *et al* 2014, Gilron *et al* 2019). In fact, the addition of heterogeneous elements may confound the study results and may adversely impact the ability of the statistical tests applied to the data to detect a meaningful effect.

1.8.8. Effect size

The effect size refers to the magnitude of a difference between groups or a correlation between variables and indicates the clinical relevance of a research outcome. Differently from statistical significance, the effect size is independent from the sample size, as it is calculated only from the data.

Depending on the type of comparisons under investigation, different indices can be used to estimate the effect size. Cohen's *d* or odds ratio (OR) are common indices utilised to calculate the magnitude of the effect size between groups (for example the magnitude of reduction in pain scores from one intervention *vs* another), while the measure of association between variables is calculated with the Pearson's *r* correlation (to measure the degree of linear correlation between

two quantitative variables) or the r^2 coefficient of determination (to measure the proportion of variance in one variable depending on changes in the other variable) (Sullivan *et al* 2012).

Actual effect sizes are calculated after completion of the study from the data collected, and this aspect is of particular relevance in meta-analyses, because it represents a standardised set of data that is quantitatively and readily comparable across different studies on a single topic (Sullivan *et al* 2012).
Desired effect sizes calculated in the planning stage of a prospective clinical research project are also a valuable resource and should ideally be based on previously published studies of similar methodology (Levine *et al* 2001). Incorporating a desired effect size in the sample size calculation, researchers can estimate the minimum sample size required to confer the study enough power to detect a pre-determined effect of defined magnitude on the variable of interest. In fact, the detectable effect size is inversely correlated with the study power: while high-powered large clinical trials have the ability to detect a wide range of clinically relevant effects, small trials, characterised by a lower power, will be sensitive enough only to detect large effect sizes.

1.8.9. Control groups

The choice of the control group is fundamental in analgesia clinical studies investigating acute pain. While a positive control group receive an analgesic intervention of known efficacy in the species investigated, the negative control group receive a placebo, which means complete absence of analgesic effect contextually. As summarised in an editorial by Slingsby published in 2010, studies incorporating placebo control groups are able to determine whether:

- The species under investigation demonstrate pain responses.
- The procedure itself elicits a pain response.
- The pain scoring system utilised is sensitive enough to detect a pain response.
- Other drugs administered interfere with behavioural pain responses, for example sedative or dissociative drugs (Fox *et al* 2000).

The use of negative controlled groups in dogs and cats demonstrated the existence of acute and chronic pain, that pain or its effects can be measured with a variety of methods, and the efficacy of analgesic intervention on modulation of pain responses. In negative controlled trials, the study intervention represents the only appreciable variable between the two groups, as its effects in the study group are compared to complete absence of intervention in the negative control. In fact, the investigation of the effects of a single analgesic intervention compared to no intervention on a standardised pain inducing procedure decreases the number of variables that might influence the response to the intervention. This decreased variability consequently generates results more reliably imputable to the intervention and less affected by unrelated factors, thus conferring a higher power to the study (Lipsitch *et al* 2010). Furthermore, the effect size observed will be the result solely of the intervention, which will provide information on the magnitude of attenuation of the pain response pertinent to that analgesic on a specific procedure (Moser 2019).

In the editorial author's opinion (Slingsby 2010), "Negatively controlled studies in dogs and cats with adequate rescue analgesia protocols may provide more believable data (e.g. showing that the scoring system worked) and can cause less animal suffering than a poorly designed positive controlled study". However, while negative controlled studies demonstrate the potential to yield more sound results, many recent clinical trials have utilised positive controls advocating the ethical implications of leaving pain untreated. Positive controls are expected to have a known effect of already established magnitude, thus allowing researchers to demonstrate that the study protocol is sensitive enough to detect the desired effect. This group is then compared to the study group in order to test the difference in the effects of the study intervention (Moser 2019).

1.8.10. Pain measurement instrument

Distinct properties of the pain scoring system utilised to detect and measure the effects of an intervention may affect considerably the interpretation of findings of analgesia studies.

As discussed in section 4 of this chapter, amongst the essential requirements of a scientifically developed measurement instrument researchers must consider validity, thus ensuring it actually measures the property of interest, reliability, thus providing consistent results across multiple users, and responsiveness, thus demonstrating sensitivity to change. All these attributes, when applied to a pain scoring system, translate into the ability to identify and measure pain, differentiate between animals that are either treated or not treated with analgesics, and to determine analgesic requirements (Slingsby 2010). These concepts are of fundamental importance not only for clinical purposes, but also in the research context, where studies rely on this ability to determine whether to reject the null hypothesis in favour of the research hypothesis and to draw appropriate conclusions.

In addition, consideration of the properties of the pain metrology instrument used will allow imputation of appropriate statistical testing to analyse the data (Nair *et al* 2020).

Finally, while an interval level scale provides more precise measurement compared to an ordinal scale (Morton *et al* 2005, Reid *et al* 2007), a scientifically developed and validated ordinal scale provided with a cut-off score indicative of the requirement for additional analgesia still offers reliable pain measurements with the advantage of increased utility and universality of the scale, and consistency in the provision of rescue analgesia (Reid *et al* 2007).

1.8.11. Rescue analgesic provision

Rescue analgesia refers to the administration of a selected analgesic drug/s, other than the test intervention/s, at any time during the study a patient is deemed in pain as determined by the pain measurement instrument. It represents a fundamental requirement in analgesia studies to reduce animal suffering (as well as in clinical practice) and to obtain ethical approval, which mandates provision of details of the rescue plan and the indicators for exit from the trial.

Main considerations related to the administration of rescue analgesia pertain to the choice of the rescue analgesia regimen (Slingsby 2010, Singla *et al* 2017), the use of licensed drugs, and the time to (Slingsby 2010) and the consistency of (Reid *et al* 2007, Hofmeister *et al* 2018) rescue analgesia administration. When designing a study, the choice of the rescue analgesic should be a drug proven to be effective in the species investigated and should be a licensed drug for a specific condition in that species, as the intervention is often the only unlicensed drug permitted for obtainment of ethical approval of a research project (for example in the UK) (Slingsby 2010). In addition, it should have rapid onset of action due to the implications on animal welfare.

A relevant impact on the study outcomes is determined by the rescue plan adopted in the placebo and treatment arms. A conceptual review that analysed studies conducted in human patients undergoing first metatarsal bunionectomy (Singla *et al* 2017) revealed considerable variability in the rescue regimen between studies, an overall high proportion of subjects in the studies included requiring rescue in both study arms, and a significant variability in the mean number of doses of rescue drugs administered in the postoperative period. Analysis of the data collected indicated that the study outcomes could be significantly influenced by both liberal use of rescue analgesia, which may negatively impact the assay sensitivity, and stringent use of rescue, which may lead to an unacceptable increase in patients withdrawing or being removed from the study.

The time to administration of rescue analgesia also provides researchers with valuable information on the test drug. When additional analgesia is required shortly after administration of the intervention, researchers might infer that the test drug is not effective, or that it is characterised by ultra-short duration of action (as it might be the case for remifentanil for example), or that the animals were in the negative control group. Instead, the administration of the rescue drug after a period of analgesia might provide information on the duration of action of the intervention.

As the provision of additional analgesia relies on the assessment that an animal is in pain, a standardised pain score level that can be used to indicate these additional requirements represents an advantage in terms of consistency of

provision of rescue analgesics, decreased inter-observer variability, and improved comparability of findings between studies. The value of the intervention level score was confirmed during the validation process of the GCMPS-SF (Reid *et al* 2007), as it was found to be consistent across a varied population of dogs undergoing a variety of surgical procedures, and when pain was scored by multiple observers in three referral teaching hospitals.

In contrast, a more recent study investigating the agreement amongst six ACVAA diplomates (referred to as 'evaluators') when pain scoring videos of dogs recovering from anaesthesia showed different results (Hofmeister *et al* 2018). In this study, the evaluators assessed pain with the simultaneous use of three scoring systems - the VAS, the NRS, and the GCMPS-SF – presented to them in a random order. The pre-assigned cut-off scores used to indicate the requirement for additional analgesia were VAS ≥ 4/10, NRS ≥ 4/10, and GCMPS-SF ≥ 6/24. For each case, if the evaluator's clinical recommendation to administer or not administer additional analgesia was in disagreement with the cut-off score, the case was referred to as a 'conflict'. Interestingly, the evaluators' assessment of analgesia requirements was likely to generate more conflicts with the use of the GCMPS-SF than with the other two scales. In fact, the recommendation to administer rescue analgesia agreed with cut-off scores in 51.6%, 77.4%, and 71% of the evaluations with the use of GCMPS-SF, the NRS, and the VAS, respectively ($p = 0.0076$). Based on these findings, the authors suggested that the score provided by the GCMPS-SF, being a number generated from multidimensional data, may not represent the most appropriate way to make the complex clinical decision of the need for analgesia. Differently, the VAS and NRS scores are more influenced by the observer perception, as the evaluator will place a mark on the scale above the cut-off score if an animal is considered in pain. This might explain the better agreement between clinical decision and cut-off VAS and NRS scores observed in this study. However, these findings highlight an existing controversy on the appropriate use of the intervention level of the GCMPS-SF and its association with the clinical decision to administer additional analgesia.

1.8.12. Data analysis

The type of variable and distribution of the data will dictate the appropriate statistical tests to be used for analysis.

Quantitative variables contain numeric data that represent real amounts, to which all mathematical operations can be applied, and include *discrete* (e.g. number of patients hospitalised) and *continuous* (e.g. blood pressure, temperature) data. While the former can only be expressed as integers within a given range, the latter can potentially assume infinite values, including decimals, within the range selected.

Qualitative (categorical) variables contain data that represent groups, and can be expressed as *binary* (yes/no outcomes), *nominal* (attributes defining a category without a rank or order between them, i.e. number of cat breeds), or *ordinal* (representing attributes in a category ranked in a specific order) data.

Depending on the measurement properties of the instrument that generates them, pain scores can be statistically treated as different types of variables. Pain scores, describing the level of pain, are categorical variables that may be recorded as descriptors ("mild", "moderate", severe" of the SDS) or as numbers (for example pain scores assigned with the NRS, the VAS, and the GCMPS-SF). Yet, these numbers are descriptive of a category, rather than representing actual amounts, and as such must be treated as a categorical variable. In particular, as pain scores (irrespective whether recorded as descriptors or numbers) are used to describe the rank position within a category, they can be more specifically expressed as ranked categorical data, also termed ordinal data. Pain scores generated with the GCMPS-SF, which is an ordinal scale, are an example of ordinal data, being classified into categories that are ordered according to the level of severity. However, a number of transformations can be applied to the data in order to convert an ordinal into a quantitative variable, as this will allow imputation of a wider range of statistical techniques. Pain scores for example can be expressed as a ratio that indicates the amount of decrease following administration of an intervention, thus obtaining an actual number to be used for analysis (Nair 2020).

The data collected are first inspected to determine the frequency of distribution of values. *Normally distributed* data are characterised by most values lying around

a centre and symmetrically tapering off toward the tail ends (Fig. 1.7), while a *skewed distribution* may have most values lying asymmetrically toward one end or the other (Fig. 1.8).

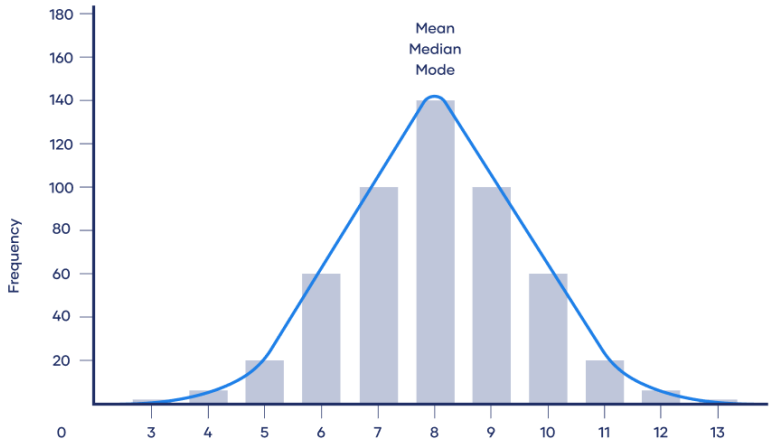

Fig. 1.7 Example of normal distribution. Reprinted and adapted from www.scribbr.com.

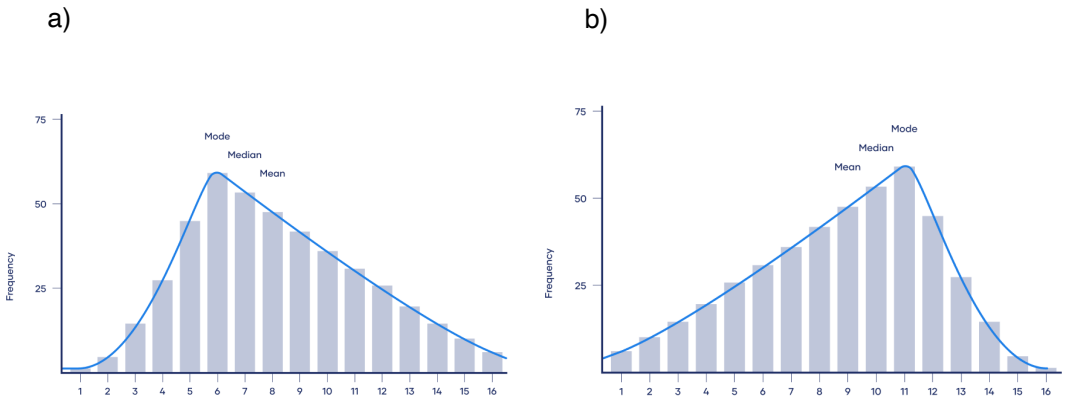

Fig. 1.8 Examples of positive (a) and negative (b) skewed distribution. Reprinted and adapted from www.scribbr.com.

This information is important to calculate the most appropriate **descriptive (summary) statistics** to summarise the data set, which include measures of central tendency and variability.

Measures of central tendency represent the average value within the sample, expressing the overall tendency of the data. Depending on the distribution of the data, this central tendency is better expressed as the *mean* (calculated as the sum of all the values divided by the total number of values), *median* (calculated by ordering the values in the sample in an ascending or descending order and selecting the number in the middle of the range) or *mode* (displaying the most represented value/s in the sample) (Fig. 1.7 and 1.8). Calculation of the mean may be the most appropriate measure for normally distributed data, although it is important to note that mean, median, and mode will correspond to the same value in this case, while mode and median better express the central tendency for values following a skewed distribution (McCluskey *et al* 2007).

Information concerning the spread of data in the sample is obtained by calculation of measures of variability. The *range*, calculated subtracting the lowest from the highest value, simply gives an idea of the interval between the extreme values, thus reflecting the spread of the whole data set, and it is the measure that provides the least information as it can be considerably influenced by outliers (McCluskey *et al* 2007). In skewed distribution, more information about variability within the sample is provided by the *interquartile range (IQR)*, which reflects the spread of the middle half of a data set, thus not being subjected to the influence of extreme outliers. Quartiles are a type of percentile, which is a value that defines the percentage of data falling below it. For example, the first quartile (Q1) corresponds to the 25^{th} percentile, indicating that 25% of values falls below the first quartile (Fig. 1.9). The whole data set is divided by three quartiles into four parts, each containing an equal number of values, of which the second and the third part constitute the IQR.

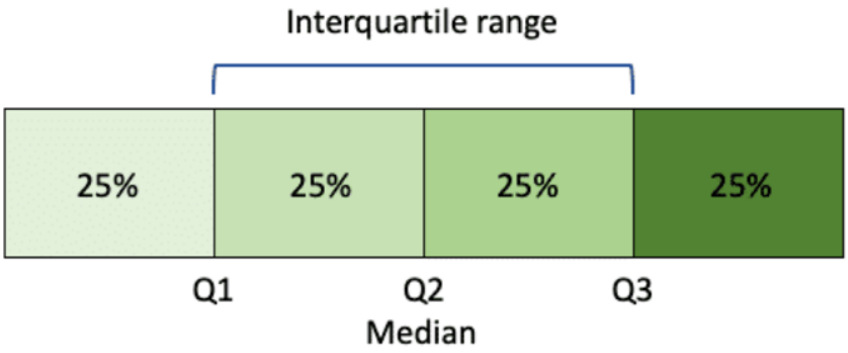

Fig. 1.9 Quartiles and Interquartile range (IQR). Reprinted from www.scribbr.com.

In normal distribution, measures of variability that best describe the data are the standard deviation and the variance. The standard deviation (SD) represents the average distance between each value and the mean, and reflects the average amount of variability of the data set: the larger the SD, the higher the variability is present in the sample. The variance, calculated as the square of the SD, provides information on the degree of spread within the dataset: the larger the variance, the more spread the data is in relation to the mean.

While descriptive statistics describe the data in the sample, **inferential (test) statistics**, which represent the next step in hypothesis testing, formally test the research hypotheses to assist in the decision on whether to reject the null hypothesis in favour of the alternative hypothesis. The research question and the study design will dictate the type of statistical tests to be used:

- Comparison tests assess the probability of an existing difference between groups.
- Regression tests assess relationships between variables with the assumption of cause and effect.
- Correlation tests assess relationships between variables without the assumption of causation.

Determination of the appropriate statistical tests also depends upon the type of variables under investigation and whether the data meet prespecified common statistical assumptions (Nimon 2012):

1. *Normality*: it assumes normal distribution of the continuous variables to be included in the analysis.
2. *Independence:* it assumes observations/variables to be tested are independent of each other.
3. *Equality of variance*: it assumes homogeneity of variance across different groups or samples.
4. *Linearity*: it assumes a linear relationship between pairs of continuous variables.

Parametric statistical tests are suitable when assumptions of normal distribution and homogeneity of variance are met, while nonparametric statistical tests do not make any assumptions about data distribution. In order to use more powerful

parametric statistical tests, non-normally distributed data (like pain scores) can undergo logarithmic transformation to be normalised (Sedgwick 2012). If the data do not meet the assumption of independence (for example multiple pain score assessments performed at intervals in the same subject during the postoperative period), a test that accounts for structure may be the most appropriate choice (for example repeated-measures tests, like the within-subjects ANOVA or ANOVA for correlated samples). While detailed description of parametric and nonparametric statistical tests is beyond the scope of this thesis, common tests and their use are summarised in Table 1.2.

Table 1.2 Summary of a few common statistical parametric tests and their nonparametric equivalents.

	Parametric test	Aim	Equivalent nonparametric test	Aim
Regression	**Simple linear regression**	Estimate the effects of a continuous variable on another variable		
	Multiple linear regression	Estimate the effects of two or more continuous variables on another variable		
	Logistic regression	Estimate the effects of one or more continuous variables on a binary outcome		
Correlation	**Pearson's *r***	Estimate the correlation between two continuous variables	**Spearman's *r* or rank correlation coefficient**	Estimate the correlation between two quantitative variables
			Chi square test	Estimate the correlation between two categorical variables
Comparison	**Paired t-test**	Compare differences amongst means of two groups that belong to the same population (dependent)	**Wilcoxon Signed-rank test**	Compare magnitude and direction of difference between distribution scores of two dependent variables
	Independent t-test	Compare differences amongst means of two groups that belong to different populations (independent)	**Wilcoxon Rank-Sum Test (Mann- Whitney U test)**	Compare sum of rankings of scores between two independent variables
	ANOVA	Compare differences amongst means of two or more groups	**Kruskal-Wallis H test**	Compare mean rankings of scores between 3 or more samples
	MANOVA	Estimate the effects of independent categorical variables on multiple continuous variables	**ANOSIM**	Estimates the effects of three or more categorical variables on two or more quantitative variables

1.8.12.1. Analysis of data after administration of rescue analgesia

Data arising from patients after provision of rescue analgesic therapies can be handled in various ways, which differ depending on the inclusion or exclusion of these data in further analyses.

When data acquired after rescue administration are included, descriptive summary statistics can be calculated as a proportion, which indicates the fraction of patients within a group that received rescue, and will be expressed as a percentage, or can be calculated as a mean, indicating the average number of doses of rescue analgesics administered in a study group (Singla *et al* 2017). Inclusion of patients that received rescue medications in the statistical analysis may lead to results heavily dependent on the rescue regimen adopted: if liberal, it may negatively impact the assay sensitivity; if stringent, it may lead to an unacceptable increase in patients withdrawing/being removed from the study (Singla *et al* 2017).

When data arising from patients after provision of additional analgesia are excluded from statistical analysis (or patients are removed from the study), there are a few considerations to mention.

Exclusion of these subjects from analysis may lead to missing data at random, because there will be systematic differences between the missing and observed values explained by the provision of rescue, as this event won't be equally distributed through the study arms; consequently, this eventuality may introduce bias due to the disparity in the distribution of missing data between groups (Bhaskaran *et al* 2014). With the assumption that data are missed completely at random, thus being equally distributed between groups, no bias would be introduced in the interpretation of results. However, although not introducing bias, this eventuality may lead to exclusion of a substantial proportion of the original sample with appreciable decrease in precision and power of the study (Sterne *et al* 2009).

In an attempt to preserve the study power, a variety of approaches exist to deal with imputed data, and each method entails different implications with respect to interpretation of study results.

Patients with incomplete data may be included with the use of intention-to-treat analyses, which encompass various methods, although implementation of new and more reliable techniques is constantly being evaluated (Singla *et al* 2017). One form of intention-to-treat analysis is the Last Observation Carried Forward (LOCF), a single imputation technique that can be used with longitudinal data, as are the repeated pain scores acquired from a patient at subsequent time points. The last observed value before dropout is used for that subject for all the remaining time points, therefore making the assumption that the patient's response remains constant at the last observed value. However, this approach does not account for the uncertainty regarding the missing values, which can arise from measurement errors (for example inter-observer variability in assigning pain scores) and from the fact that pain levels vary in continuous time, while values are usually measured in discrete time (pre-determined intervals). All these factors may introduce a degree of bias that will be proportional to the degree of measurement error (Moreno-Betancur *et al* 2018). Furthermore, this technique was demonstrated to be unbiased only when data were missing completely at random and when the distribution of the last observed values was exactly equal to the distribution of imputed values; since this latter condition can never be proven, the use of single LOCF has been discouraged by multiple authors (Sterne *et al* 2009, Lachin 2016, Single *et al* 2017).

A more recent approach involves the use of a multiple imputation technique, the windowed LOCF. In this case, pain assessment scores obtained after administration of rescue analgesia, as considered artificially lowered, will be ignored, and the pre-rescue pain score is carried forward for a window of time equivalent to the expected duration of action of the rescue therapy. This approach, often utilised in studies that select rescue regimens of known efficacy, allows to account for the impact of rescue administration on pain scores in the negative control group, thus presenting the advantage of minimising missing data. However, bias is still potentially introduced if the choice of the rescue regimen is

based on previous studies that employed liberal rescue protocols and calculated results with the use of single imputation techniques (Singla *et al* 2017).

1.8.12.2. Survival analysis

Survival analysis is a branch of statistics that analyses the expected duration of time (survival time) until an event occurs, for example it can be used to estimate the time to administration of rescue analgesia. Survival times cannot usually be analysed with standard statistical techniques, because data are often 'censored', which means they describe an event that either has not occurred yet or is not known to occur: if a patient arrives at the end of the study without receiving any rescue therapy, or is removed from the study before the end of data collection for reasons unrelated to the study protocol, their survival times would be considered censored.

Survival analysis can be used in several ways:

- To describe the survival times of subjects within a group
    - Survival function
    - Hazard function
    - Kaplan-Meier survival method
- To compare the survival curves of two or more groups
    - Log-rank test
- To compare the difference between survival times of two or more groups
    - Cox's proportional hazards model (Cox regression)

1.8.13. Statistical errors

In hypothesis testing, results of statistical tests run on data collected during an experiment assist researchers in the decision to support or reject the null hypothesis in favour of the research hypothesis. However, since these decisions are based on probabilities, there are distinct risks of drawing the wrong conclusions, namely the risk of false positive (Type I statistical error) and false negative (Type II statistical error) results (Table 1.3).

Table 1.3 Summary of statistical errors.

Type I and Type II statical error

Null hypothesis is	True	False
Rejected	**Type I statistical error** False positive result Probability = α	Correct decision True positive Probability = 1 - β
Not rejected	Correct decision True negative Probability = 1 - α	**Type II statistical error** False negative Probability = β

1.8.13.1. Type I statistical error

The Type I statistical error represents the probability of rejecting the null hypothesis when it is actually true, thus reflecting the probability of a false positive result.

It is strictly correlated with the significant level (α) and the *p* value. The significance level determines the Type I statistical error rate, representing the maximum risk researchers are willing to take to reject a true null hypothesis in the planning stage of the research, while a *p* value generated after data analysis indicates the actual probability of rejecting a true null hypothesis. For example, assuming researchers set the significance level at 0.05 and statistical tests give a *p* value of 0.028, results would be considered statistically significant and the null hypothesis would be rejected; however, these results still indicate a probability of 2.8% of rejecting a true null hypothesis.

In graphic visualisation, the probability curve of the null hypothesis distribution illustrated in fig. 1.10 shows the set of all possible results if the null hypothesis were actually true. If results of a study fall in the shaded area represented by α (also called 'critical region' in statistics), a true null hypothesis would be rejected leading to false positive conclusions.

Factors that influence the Type I error rate include:

- Unequal group sizes: this may lead to a higher rate of Type I statistical error, which will be positively correlated with the magnitude of group size imbalance (Rusticus *et al* 2014).
- Increases in the number of outcome measures: this practice may lead to an increase in the Type I error rate, as it increases the probability that at least one of the endpoints reaches significance merely by chance (Oyama *et al* 2017).

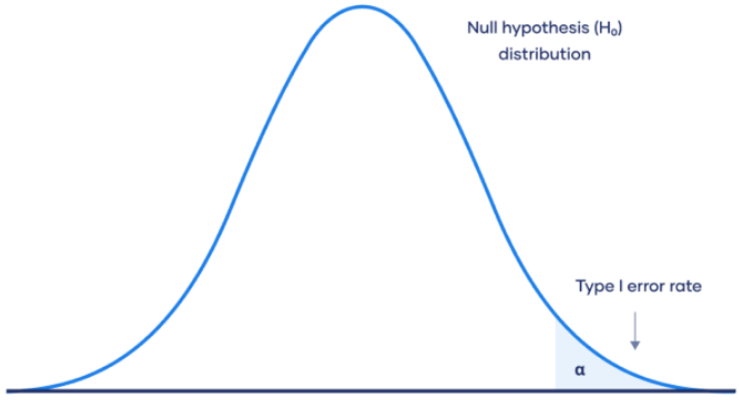

Fig. 1.10 Null hypothesis probability curve and probability of Type I statistical error. Reprinted and adapted from www.scribbr.com.

1.8.13.2. Type II statistical error

The Type II statistical error represents the probability of not rejecting the null hypothesis when it is actually false, thus reflecting the probability of a false negative result.

It is strictly correlated with the statistical power of the study (β), which indicates the probability of detecting a true effect in the population investigated if there is one. If the statistical power is set at 80%, researchers accept a probability of 20% that the study will not detect a true difference between groups or an existing correlation between variables thus leading to a false negative conclusion. Consequently, increasing the power increases the probability of detecting a true effect and, thus, decreases the risk of a false negative result.

In graphic visualisation, the probability curve of the alternative hypothesis distribution illustrated in fig. 1.11 shows the set of all possible results if the null hypothesis were actually false. If results of a study fall in the shaded area represented by β, the study fails to reject a false null hypothesis leading to false negative conclusions.

Factors that affect power will inversely affect the Type II error rate:

- Unequal or small group sizes are associated with a decrease in statistical power (Rusticus *et al* 2014).
- Increased variability is negatively correlated with the power of a study, as power decreases with increasing levels of variation (Rusticus *et al* 2014, Gilron *et al* 2019).
- The detectable effect size is inversely correlated with the study power: trials with a low power will be sensitive enough only to detect large effect sizes.

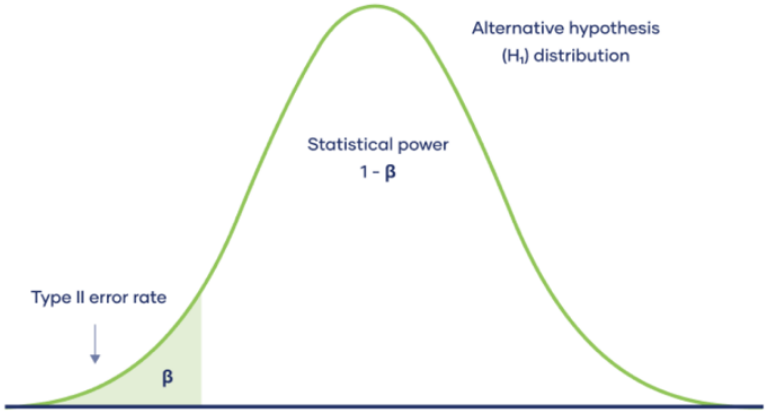

Fig. 1.11 Alternative hypothesis probability curve and probability of Type II statistical error. Reprinted and adapted from www.scribbr.com.

1.8.13.3. Interplay between Type I and Type II statistical errors

The probability of incurring Type I and Type II statistical errors are interdependent: the choice of the significant level (Type I error rate) set at the beginning of the study influences the statistical power, which is inversely related to the Type II error rate, as well as the choice of the power level affects the sensitivity of the assay, thus influencing the Type I error rate.

In graphic visualisation, the probability curves of the null and alternative hypothesis distribution illustrated in fig. 1.12 show the interdependence between the Type I and Type II error rates displayed as the overlap of the two distributions, represented by the shaded areas, where Type I and Type II statistical errors occur.

As deducible from the graph, decreasing the significant level (α) decreases the Type I error risk, but increases the Type II error rate. If researchers decrease α from 0.05 to 0.01, this indicates a reduction in the probability of a false positive result from 5% to 1%, thus also meaning that only results with a *p* value below or equal to 0.01 will be considered statistically significant; clearly, this restricted threshold for significance decreases the power of the study to detect an effect. Increasing the power level (β) decreases the probability of Type II statistical error, but increases the Type I error risk. In fact, an increase in power from 80% to 90% reflects a correspondent increase in the essay sensitivity to detect a true effect (reducing the Type II error rate by 10%), but also increases the essay sensitivity to very small random effects.

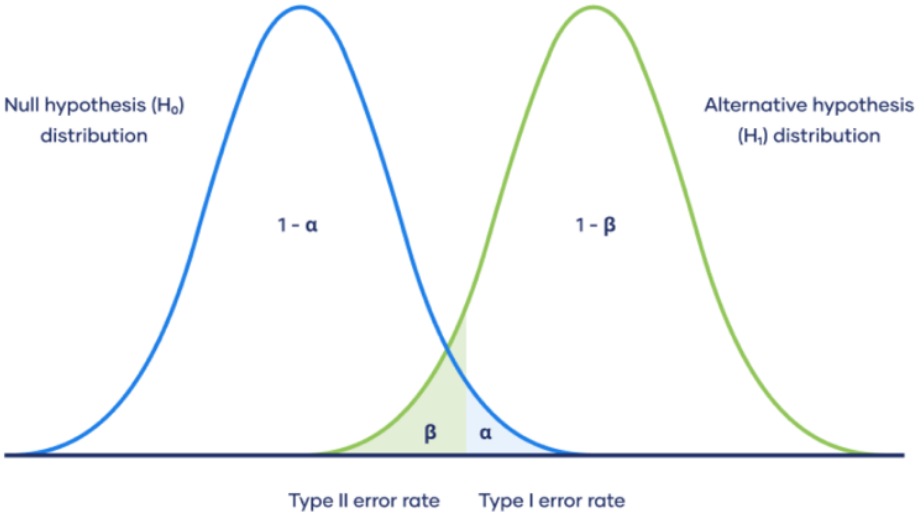

Fig. 1.12 Interplay between Type I and Type II error rate. Reprinted and adapted from www.scribbr.com.

1.8.14. Study design

1.8.14.1. Hypotheses

In randomised controlled trials, sample size estimation, methods used for analysis of primary outcomes, and interpretation of results are all dependent on clear specification of the type of hypothesis to be tested (Wang *et al* 2017).

1.8.14.1.1. Superiority design

The research hypothesis states that the intervention under investigation (treatment group) is expected to be superior to a comparative intervention (positive or placebo control group) (Wang *et al* 2017). Sample size estimation and definition of the CI are determined as described in section 7 of this chapter.

1.8.14.1.2. Equivalence design

The research hypothesis states that the difference between the interventions in two or more groups is expected to be negligible, thus clinically unimportant. In statistics, it is not possible to prove that different interventions are exactly equal, so in this trial design a tolerance range is established which defines what would be considered the minimum important difference between treatments (Lesaffre 2008). To reject the null hypothesis, and declare equivalence between treatment and control groups, the mean difference of two or more groups must fall within the tolerance range (Wang *et al* 2017).

1.8.14.1.3. Non-inferiority design

The research hypothesis states that the intervention is not inferior to a gold standard or a treatment of known efficacy. Careful consideration of the following

elements of non-inferiority design is essential to ensure robust research conduct and to mitigate against the possibility of drawing inaccurate conclusions (Rehal *et al* 2016):

- Clear justification of the non-inferiority margin: this is used as a reference for conclusion of non-inferiority, as it is the value that defines the maximum clinically acceptable extent to which the intervention can be less effective than the control while still providing evidence of an effect.
- The set alpha level should be consistent with the CI, as these allow to declare statistical significance when compared against the margin. Inconsistencies between the pre-determined values of these parameters may lead to biased inferences, for example CIs stricter than the alpha level may predispose to false positive results.

1.8.14.2. Controlled *versus* Observational

In **Observational clinical studies** researchers merely observe the results of a diagnostic test, the effect of a risk factor, a treatment, or other interventions in a group of subjects that share similar characteristics (cohort studies), without interferences or manipulation of the research participants. An example of a cohort study in the veterinary literature is an investigation conducted by Lush and colleagues in dogs on the correlation between personality and pain scores (Lush *et al* 2018).
This research design may involve observations of two groups (case-control studies), although it is important to notice that both groups are chosen based on attributes of interest they already possess (for instance investigation of the occurrence of lung disease in two groups, smokers and non-smokers).
Despite observational studies are devoid of ethical implications, the lack of standardised conditions makes inferences challenging due to the presence of multiple confounding factors.

Controlled clinical studies include a comparison between two or more groups, where the treatment group/s receive the intervention/s while the control group receive either a placebo or an intervention of established efficacy. Control of confounding factors reduces variability, which allows researchers to make

stronger inferences regarding a cause-and-effect relationship between variables (Hariton *et al* 2018). Inclusion of characteristics such as blinding and randomisation to controlled studies further reduces bias, which has led to consider randomised, blinded, controlled trials the gold standard in clinical research due to the higher quality of evidence they produce (Hofmeister *et al* 2007, Jones *et al* 2015, Hariton *et al* 2018). Quality of results of such trials is also related to power and sample size, which, amongst other factors, are based on the study design.

In a within-subjects design (cross-over trials), where each participant receives all interventions in a randomised order with a washout period in between treatments, inter-subject variation does not unevenly affect the outcomes of different treatments; as such, fewer participants are required in this design to provide the same level of power (Safarkhani *et al* 2017).

In a between-subjects design (longitudinal trials, or parallel-group design), each participant receives only one treatment, and multiple subjects are allocated to each treatment. Therefore, as this design is characterised by a greater individual variability, a larger sample size is required to establish relationships between variables. Bias potentially associated with uneven distribution of variability between groups can be minimised by random group allocation.

1.8.14.3. Blinding

Blinding, or 'masking', refers to the concealment of assigned interventions from individuals involved in a trial who may potentially be influenced by this information (Day *et al* 2000). In human trials, these individuals may be represented by participants, healthcare providers, data collectors, and outcome evaluators. Lack of blinding of participants to group allocation may introduce bias by influencing the responses to treatment due the awareness of administration of a placebo or a new intervention, by influencing the compliance to treatment, and by the risk of increased dropout rates of participants in placebo arms (Moher *et al* 2010). While this aspect of blinding does not apply to veterinary trials, blinding of evaluators, data collectors, and data analysts is as fundamental as in human trials, especially when the outcome measure involves subjective assessments of the response (for example alleviation of pain) (Day *et al* 2000). In fact, masking at multiple levels prevent bias associated with evaluation of responses, patient

management, decision of treatment success/failure, and the choice of analytical strategies (such as selection of favourable outcomes or time points) (Moher *et al* 2010).

Blinding, unlike treatment concealment, may not always be possible, for instance in some cases when the aim is to assess levels of pain induced by two surgical procedures with distinct traits that would make group allocation obvious.

Correct use and detailed reporting of the blinding strategies adopted are key components to define the quality of controlled trials, as there is sustained body of evidence in the literature that failure to use appropriate levels of blinding and/or to report items with a risk of bias is associated with finding exaggerated effect sizes (Day *et al* 2000, Rufiange *et al* 2019).

1.8.14.4. Randomisation

Randomisation refers to the method used for treatment assignment of participants at trial entry, and constitutes a fundamental component of high quality randomised controlled trials (RCTs) (Moher *et al* 2010). Successful randomisation hinges on two equally important steps: generation of an unpredictable allocation sequence and concealment of this sequence from investigators and evaluators until assignment is completed.

Random allocation has three main advantages: first, it reduces selection bias at trial entry by balancing both known and unknown sources of variability between groups (Rufiange *et al* 2019). Secondly, it permits the use of probability theory, which expresses the likelihood that differences in outcomes between treatment groups are determined merely by chance (Di Girolamo *et al* 2017). Finally, it may facilitate blinding the identity of treatments to researchers and evaluators, which further reduces bias after treatment allocation (Moher *et al* 2010).

A proper randomisation procedure, by preventing selection bias, proffers support to the internal validity of the study (Di Gerolamo *et al* 2017, Rufiange *et al* 2019).

1.8.14.5. Single *versus* multicentre

Single centre clinical trials, where patients are enrolled in a single practice, hospital, or centre, are characterised by a limited sample size compared to multicentre trials, although they present the advantages of being less expensive and easier to conduct.

Multicentre clinical trials represent a collaborative effort of two or more independent centres in the enrolment of participants and data collection. On one hand, these trials offer several advantages: increased sample sizes, a larger variety of population characteristics (for example in terms of variety of breeds represented), representation of different contexts (multiple surgeons and evaluators, different geographical locations, different hospital procedures). Input of these heterogeneous factors in the study will result in increased generalizability of findings to the overall population (Furr 2012). On the other hand, these studies are more difficult and more expensive to plan and conduct.

Multiple meta-analyses in the human literature have explored the difference in results obtained in single versus multicentre trials and its association with elements of the study design (Bafeta *et al* 2012, Unverzagt *et al* 2013), and overall larger effect sizes were observed in single centre compared to multicentre trials. Amongst the various underlying mechanisms that may explain this finding [different mechanisms of patient selection, higher expertise of teams in single centre trials, standardised interventions (Bafeta *et al* 2012)], major differences were identified in methodological/reporting quality. Smaller trials showed a tendency to increased risk of bias related to deficiencies in implementation and reporting of strategies such as random sequence generation, allocation concealment, and blinding (Bafeta *et al* 2012, Unverzagt *et al* 2013).

1.8.14.6. Clinical *versus* experimental

Experimental studies involve the use of experimental (laboratory) animals owned by a research facility. These studies are characterised by a low degree of patient

and environmental variability due to strict standardised controlled conditions and selection of purpose-bred subjects that share similar characteristics with no attempt to randomly sample a study population. While this study design excels in reproducibility and internal validity, it lacks external validity, as findings may not be generalised beyond specific methods, participants, and conditions (Fiske *et al* 2005).

Clinical studies involve the use of client-owned animal patients admitted to a clinic, hospital, or centre that can be enrolled in a controlled clinical trial (interventional study) or in an observational study (Bertout *et al* 2021). As these studies are characterised by a higher degree of environmental and patient variation compared to the laboratory setting, they may hold a decreased level of precision (Fiske *et al* 2005) due to the possible influence of confounding factors on study results. However, reproducibility and internal validity are achieved through good research conduct that involves proper implementation of such steps as randomisation, blinding, and estimation of an adequate sample size in the planning stage. Moreover, the higher variability in the clinical conditions and the natural variation within the sample population provide external validity (generalizability of findings) to these studies.

In particular, the veterinary subject diversity is among the factors that confer translational value to naturally occurring companion animal models (Klinck *et al* 2017). Cross-species research extends from comparative studies in oncology (Kol *et al* 2015), to orthopaedics (Meeson *et al* 2019), immunology (Bilgic *et al* 2019), aging (Hoffman *et al* 2018), neurologic diseases (Steinmetz *et al* 2013, Golubczyk *et al* 2019), as well as to pain research (Klinck *et al* 2017, Robertson-Plouch *et al* 2019).

1.8.15. Minimum clinically important difference (MCID)

The minimum clinically important difference (MCID), defined as the smallest difference in scores in the domain of interest perceived as relevant by the patient (verbal patients) or the clinician (non-verbal patients) and which would mandate changes in patient's management, was introduced to differentiate between

statistical significance and clinical relevance (Jaeschke *et al* 1989). While a statistically significant difference detected in a research clinical trial might be relevant, very small statistically significant effects (that might be detected especially in large trials) may not have clinical relevance. In addition to its clinical utility, the MCID assumes a fundamental role also in the planning stage of research clinical studies. In fact, during sample size estimation, it can assist researchers in determining the minimum relevant effect the essay should be sensitive to in order to estimate an adequate sample size (Olsen *et al* 2017).

Many conflicting findings do exist in the human literature regarding the assessment of the MCID for pain scores, as it can be influenced by a variety of factors including the scoring system and baseline pain levels (Olsen *et al* 2017, Bahreini *et al* 2020). A clinical trial in human patients with break-through cancer-related pain (Farrar *et al* 2000) studied the pain scores obtained with several pain scales and integrated the patient's assessment for requirements for additional doses of transmucosal fentanyl citrate to evaluate cut-off points. Results of this study indicated that a reduction of 33% or 2 points on an 11- point NRS was considered clinically important and that this was consistent in this population of cancer patients with a variety of painful conditions. However, more recent studies report a considerable variability irrespective of the pain measurement instrument used (Bahreini *et al* 2020). A recent meta-analysis of human trials that assessed the MCID in acute pain in the emergency department reported absolute values for the MCID ranging from 8 to 40 mm on a 100-mm scale and relative changes from baseline pain scores ranging from 13 to 85% (Olsen *et al* 2017).

Paucity of information can be retrieved on the MCID in pain scores in the veterinary literature. McKune and colleagues (2014) suggested a minimum clinically relevant difference in GCMPS-SF scores of 2.6, based on the previous work of Morton and colleagues in 2005. However, in the cited publication (Morton et al 2005) the authors found a mean ± SD GCMPS score of 2.6 in dogs that had undergone a surgical procedure compared to 1.2 ± 1.2 in dogs that had not undergone surgery. The median pain scores for the same two groups (surgery *vs* no surgery) were 2.4 and 0.9, respectively. Therefore, the value of 2.6 reported as the MCID in the work conducted by McKune and colleagues (2014) was not indicative of the difference between groups in the original publication, nor it was

regarded as the minimum clinically relevant difference. To the best of this thesis' authors' knowledge, no formal assessment of the MCID exists in the veterinary literature.

1.9. The use of evidence synthesis to evaluate research conduct

Evidence synthesis can be defined as the review of what is known from existing literature using systematic, rigorous, and transparent methods in order to clarify the evidence base (Gough *et al* 2020). Using these principles of rigour and transparency, various methods exist to conduct such syntheses and to address the research questions, including systematic and scoping review approaches. Key characteristics that differentiate scoping and systematic reviews are summarised in Table 1.3.

Table 1.3 Characteristics of traditional literature reviews, scoping reviews, and systematic reviews. Reprinted from Munn *et al* 2018.

	Traditional Literature Reviews	Scoping reviews	Systematic reviews
A priori review protocol	No	Yes (some)	Yes
PROSPERO registration of the review protocol	No	No[a]	Yes
Explicit, transparent, peer reviewed search strategy	No	Yes	Yes
Standardized data extraction forms	No	Yes	Yes
Mandatory Critical Appraisal (Risk of Bias Assessment)	No	No[b]	Yes
Synthesis of findings from individual studies and the generation of 'summary' findings[c]	No	No	Yes

[a]Current situation; this may change in time. [b]Critical appraisal is not mandatory, however, reviewers may decide to assess and report the risk of bias in scoping reviews. [c]By using statistical meta-analysis (for quantitative effectiveness, or prevalence or incidence, diagnostic accuracy, aetiology or risk, prognostic or psychometric data), or meta-synthesis (experiential or expert opinion data) or both in mixed methods reviews

1.9.1. Systematic review

A systematic review is a comprehensive search of the literature and evidence that meets pre-set specified criteria in order to address a well-defined research question/s using a systematic approach that minimises bias (Higgins *et al* 2011).

Key features of systematic reviews are (Arksey *et al* 2005, Peters *et al* 2015, Munn *et al* 2018):

- Clearly stated specific research question or series of questions addressing the effectiveness, meaningfulness, feasibility, or appropriateness of a treatment or practice relevant to end users.
- Rigid set of eligibility criteria for a relatively narrow range of quality studies.
- Explicit, rigorous, transparent, thus reproducible methodology.
- Systematic search of the international evidence to identify all studies that would meet the eligibility criteria.
- Inclusion of steps to reduce error and increase reliability (for example multiple reviewers).
- Structured data extraction and presentation.
- Assessment of the risk of bias and of the validity of findings of the included studies.
- Systematic presentation and synthesis of findings from the included studies.

These reviews may be considered the gold standard of evidence-based medicine (Munn *et al* 2014), as they may inform on whether current practice is based on relevant evidence, on the quality of that evidence, and may produce statements to guide clinical-decision making and to develop trustworthy clinical guidelines (Munn *et al* 2018).

1.9.2. Scoping review

According to Colquhoun and colleagues (2014), a scoping review can be defined as ‘a form of knowledge synthesis that addresses an exploratory research question

aimed at mapping key concepts, types of evidence, and gaps in research related to a defined area or field by systematically searching, selecting and synthesizing existing knowledge'.

Key features of scoping reviews are (Colquhoun *et al* 2014, Peters *et al* 2015, Munn *et al* 2018, Peters *et al* 2021):

- Clearly stated set of objectives (research questions), that can be broader and beyond those related to effectiveness compared to systematic reviews.
- Specified inclusion criteria for all relevant literature regardless of study design and quality of the studies included.
- Explicit, rigorous, transparent, thus reproducibile methodology.
- Systematic search of the evidence to identify all studies that would meet the eligibility criteria.
- Inclusion of steps to reduce error and increase reliability (for example multiple reviewers).
- Structured data extraction and presentation.
- Assessment of the risk of bias and of the validity of findings of the included studies may not be carried out.
- Systematic presentation of findings from the included studies.
- Synthesis of findings from individual studies or as a summary may not be required.

1.9.3. Indications for the conduct of scoping reviews

In general, a key difference between systematic and scoping reviews relies on the research question/s: a scoping review tends to have a broader aim than systematic reviews with associated extensive inclusion criteria (Munn *et al* 2018). Being designed to answer broader research question/s and having less narrow eligibility criteria than that of a systematic review, scoping reviews are particularly useful in disciplines with emerging evidence, where a comprehensive high-quality body of literature does not yet exist (Colquhoun *et al* 2014).

One of the main purposes of scoping reviews is to identify and map the types of available evidence on a relevant topic or field in terms of its nature, features, and volume (Peters *et al* 2015), through a systematic, unbiased, and exhaustive summary of the literature (Arksey *et al* 2005). This approach can be adopted to map a selected body of literature with relevance to time, location (for example country), source (for examples peer-reviewed or grey literature, unless pre-determined filters are applied), and origin (for example clinical trials in the academic, industrial, or private sector fields) (Peters *et al* 2015).

In addition, scoping reviews have a great value not only in mapping the research available, but also in examining how this research has been conducted, as this approach is suitable for the investigation of study design and conduct of research on a particular topic (Peters *et al* 2015, Munn *et al* 2018).

Finally, when the researchers' interest is the identification, reporting and/or discussion of specified characteristics in papers or studies, a scoping review approach is appropriate to identify these key characteristics and to identify and analyse gaps in the research knowledge (Munn *et al* 2018), thus also allowing to make recommendations for future research (Peters *et al* 2015).

1.10. Aims and Objectives of this Project

A relevant finding documented in the literature is the varied methodology used in veterinary acute pain research studies, with inconsistencies resulting in the potential to affect study outcomes.

Of particular concern is the possibility of both Type I and Type II statistical error across the veterinary literature. Oyama *et al* (2017) underlined a recent tendency toward multiple endpoints, which is attractive to investigators because increasing the number of outcome measures decreases the required sample size. The authors pointed out how this practice, though, increases the possibility that at least one of the endpoints reaches significance merely by chance. This would increase the probability that a truly null hypothesis would be declared significant, thus leading

to the possibility of Type I statistical error, unless the threshold for significance for each primary outcome was made more stringent than 0.05 (Oyama *et al* 2017). A work conducted by Hofmeister *et al* in 2007, a web-based search review of veterinary analgesia studies that declared "no difference between treatments", concluded that 77% of these studies did not have sufficient power to detect a small (20%) difference in treatment effect between groups. In the authors' opinion, the possibility of Type II statistical error in this high proportion of the veterinary analgesia literature analysed in this review was indicative of major deficiencies in clinical research planning and determination of sample size. The authors stressed the fundamental role of prospective sample size calculations, the sensitivity of the scoring system used, and the consideration of the degree of treatment effect that is deemed clinically significant, and concluded highlighting the importance of "methodologically sound, prospective, randomised, blinded, controlled clinical trials" (Hofmeister *et al* 2007) due to the greater impact they have within evidence-based clinical decision making.

Despite the apparent popularity of the GCMPS-SF in acute pain trials in the dog, to the best of the authors' knowledge the conduct of research studies involving the instrument has not been assessed. In this thesis we propose a series of investigations using scoping review methodology with the overall **aims** of mapping the available evidence and examining the appropriateness of study design and clinical metrology use in acute postoperative pain studies in dogs. We propose 4 major **objectives**:

1.10.1. Popularity

Properties of validity, reliability, utility, and responsiveness of the GCMPS-SF, demonstrated across multiple contexts and surgical procedures during the validation process and subsequent studies, have led to a wide use of the scale within the analgesia literature since its development. The first **objective** of the present thesis is to describe the use, and to document the impact, of the GCMPS-SF in acute postoperative pain studies in the dog via a systematic web-based search of the literature to extrapolate variables describing the publications that employed it. We **hypothesise** that the use of the scale increases over time since

its development in 2007, both in terms of demographic spread and the number of research studies in which it has been employed.

1.10.2. GCMPS-SF use

The scientific methodology underlying the development of the GCMS-SF is based on robust and detailed criteria of items selection and assignment of weight to descriptors, which were also applied to derive the intervention level for provision of additional analgesia. In order to retain its validity, the scale must be used as originally developed, as modifications made to either the scale or the intervention level will alter its measurement properties. The **Objective** of the present work is to document whether the use of the GCMPS-SF in postoperative acute pain studies in dogs is appropriate via review of any alterations to the scale or the intervention level, and to document the statistical approaches used to analyse measured data as reported by the authors of the studies that employed it. We **hypothesise** that the majority of authors will have correctly used the scale in a recent cohort of postoperative analgesia studies.

1.10.3. Study design and power

Many factors related to the study design may significantly affect the ability of the assay to detect a statistically significant difference and/or may result in biased conclusions. Factors such as blinding and randomisation, the type of control group, the number and size of groups, the hypotheses, the primary and secondary outcomes selected, and the power of the study to detect a desired effect size require careful consideration in the planning stage of a research. Both the veterinary and human analgesia literature have documented the importance of conducting an *a priori* power estimation and adherence to guidelines for reporting of clinical trials (Moher *et al* 2010) to minimise the possibility of Type I and Type II statistical error. To test the **hypothesis** of whether research conduct in clinical trials employing the GCMPS-SF is appropriate, two strategies are implemented in the present thesis with the following **objectives**:

a. Review of
    I. Factors such as primary and secondary outcomes, group sizes, hypotheses, type of control group, and use of an *a priori* sample size calculation to investigate the adequacy of the study design.
    II. Completeness of reporting of those strategies which mitigate against error (such as correct implementation of sample size estimation and full reporting of all the elements that constitute it) to determine whether these studies are adequately powered.
b. Implementation of univariable analysis and subsequent multivariable logistic regression to investigate the association of study design factors with the detection of a statistically significant difference.

1.10.4. Actual effect size

Determination of statistical significance does not necessarily equate with clinical relevance. Especially in large clinical trials, detection of very small statistically significant differences may not translate into a clinical difference relevant to the patient or the clinician that would mandate changes in patient management. Although the desired effect size may be incorporated in the sample size calculation of trials that ultimately declare statistical significance, if the study is powered for another primary endpoint than the difference in the GCMPS-SF scores, the actual effect size of the GCMPS-SF may not be of clinical significance. Hence, a final **objective** of this thesis is to determine whether statistically significant differences appear clinically relevant by comparing actual effect sizes across different procedures with suggested minimum effect sizes.

1.10.5. Summary of the objectives and how they will be covered in this thesis

As described above, this project aims to address 4 objectives (popularity, use of the scale, study design and power – which also encompasses the further analysis

on association of study design and the finding of a statistically significant difference -, and effect size), which will be addressed as follows:
Chapter 2 aims to address the popularity of the scale, the use of the scale, and the study design and power of the trials that employed it (thus encompassing 3 objectives).
Chapter 3 aims to investigate the association of study design factors with the finding of a statistically significant difference.
Chapter 4 focuses on the investigation of the desired and the actual effect size.

For more clarity, the four objectives explored in this thesis will be covered as described below:

1) Popularity (section 1.10.1.). Covered by chapter 2
2) Use of the scale (section 1.10.2.). Covered by chapter 2
3) Study design and power (section 1.10.3.a and 1.10.3.b.)
    a. Review of the literature. Covered by chapter 2
    b. Association of factors of the study design with the finding of a statistically significant difference between groups. Covered by chapter 3
4) Effect size (section 1.10.4.). Covered by chapter 4

PREFACE TO CHAPTER 2

This chapter is an actual published paper [Testa B, Reid J, Scott ME, Murison PJ and Bell AM (2021) The Short Form of the Glasgow Composite Measure Pain Scale in Post-operative Analgesia Studies in Dogs: A Scoping Review. Front. Vet. Sci. 8:751949. doi: 10.3389/fvets.2021.751949].

As the present thesis was written in an alternative format, this chapter represents an accurate duplicate of the published work, which has been edited solely in terms of font and layout to adapt it to the requested style of an MVM(R) thesis.

However, it is appropriate to report a few clarifications in this introductory section to chapter 2.

Presentation of Fig. 2.2 has been slightly modified compared to the original version, although the content, type of graph, and data presented are unaltered. During the reviewing process of the present thesis, the trendline superimposed on the bar chart was deemed inappropriate (as publications are not related, hence there is no trend *per se*), and a decision was made to remove it. Accordingly, also the mention to the trendline originally present in the title of this figure has been removed.

We also acknowledge that some of the terminology utilised in the published paper might benefit from further explanation, with the consideration that papers are often more concise than a standard thesis and that some expressions are not in common use. Therefore, in order to improve readability of this chapter, we include a brief glossary to help the reader navigate through the terms and concepts expressed.

GLOSSARY

Change relative to baseline	Pain scores are not reported as absolute values, data are given instead as the delta from pre-treatment to post-treatment values.

Cross-over design	Controlled clinical trial characterised by a within-subjects design, where each participant receives all interventions in a randomised order with a washout period in between treatments.
Guiding of rescue	Pain scores are used to determine whether an animal requires additional analgesia beyond what is mandated by study protocols. This analgesia is termed ‘rescue’ and administration is triggered by a pain score above a certain threshold.
Pooled into classes	Continuous data are categorised, grouped into broader classes transforming them into categorical data. For example, pain scores of 0-8 might be termed ‘mild pain’, 8-16 ‘moderate pain, and 16-24 ‘severe pain’.
Research instrument	A tool used to collect, measure, and analyse data related to a research subject. Research instruments can be tests, surveys, scales, questionnaires, or even checklists.
Submission checklists	Checklists to be completed [in accordance with the CONSORT (Consolidated Standards of Reporting Trials) guidelines (Moher *et al* 2010)] by authors and reviewers to ensure that all fundamental aspects of a randomised controlled trial have been addressed and reported with transparency (for example *a priori* sample size estimation, study design, statistical tests to be used, etc).

CHAPTER 2
THE SHORT FORM OF THE GLASGOW COMPOSITE MEASURE PAIN SCALE IN POST-OPERATIVE ANALGESIA STUDIES IN DOGS: A SCOPING REVIEW

2.1. Abstract

The measurement and treatment of acute pain in animals is essential from a welfare perspective. Valid pain-related outcome measures are also crucial for ensuring reliable and translatable findings in veterinary clinical trials. The short form of the Glasgow Composite Measure Pain Scale (GCMPS-SF) is a multi-item behavioural pain assessment tool, developed and validated using a psychometric approach, to measure acute pain in the dog. Here we conduct a scoping review to identify prospective research studies that have used the GCMPS-SF. We aim to describe the contexts in which it has been used, verify the correct use of the scale, and examine whether these studies are well-designed and adequately powered. We identify 114 eligible studies, indicating widespread use of the scale. We also document a limited number of modifications to the scale and intervention level, which would alter its validity. A variety of methods, with no consensus, were used to analyse data derived from the scale. However, we also find many deficiencies in reporting of experimental design in terms of the observers used, the underlying hypothesis of the research, the statement of primary outcome, and the use of *a priori* sample size calculations. These deficiencies may predispose to both type I and type II statistical errors in the small animal pain literature. We recommend more robust use of the scale and derived data to ensure success of future studies using the tool ensuring reliable and translatable outcomes.

2.2. Introduction

The translational value of natural companion animal models of pain has recently been highlighted (Kol *et al* 2015, Klinck *et al* 2017). Acute pain is common in veterinary practice and valid measurement of this abstract construct is crucially important as a fundamental prerequisite to effective pain management (Mathews *et al* 2014, Epstein *et al* 2015). Translational and veterinary clinical research designed to demonstrate the efficacy of analgesic interventions also relies on the use of valid pain outcome measures (Reid *et al* 2018). However, this can be challenging as pain is an unpleasant multi-dimensional experience with sensory and emotional components, which, by its nature, is not directly measurable in animals as they are unable to self-report.

Historically, acute pain in animals has been measured using behavioural observation quantified with simple tools such as the simple descriptive scale (SDS), numerical rating scale (NRS) and the visual analogue scale (VAS) (Holton *et al* 1998a). However, these tools are associated with a high level of inter-observer variation and their unidimensional nature may not adequately capture complex constructs like pain (Holton *et al* 1998a, Holton *et al* 1998b). The Glasgow composite measure pain scale (GCMPS) is a multi-item behavioural pain assessment tool, developed using a psychometric approach, to measure acute pain in the dog (Holton *et al* 2001, Morton *et al* 2005). The short form of the scale (GCMPS-SF) was developed for routine clinical use and comprises six behavioural categories with associated descriptors: vocalization, attention to wound, mobility, response to touch, demeanour and posture/activity (Reid *et al* 2007). The GCMPS-SF has been validated for the assessment of acute post-operative pain and importantly the score is linked to an intervention level, which guides the requirement for additional analgesia. To retain the validity of the scale, it should be used as it was originally described and validated, thus preserving its integral measurement properties.

As one of the few validated instruments for acute pain measurement in dogs, the GCMPS-SF has been adopted widely in research studies investigating the effect of drugs and interventions on perioperative acute pain. Research studies of this type may be complex and challenging to conduct, requiring careful consideration of

factors including group sizes, statistical power, control groups, pain measurement instruments, rescue analgesic provision, and data analysis (Slingsby 2010, Hofmeister *et al* 2018). Of particular concern is the finding that many studies of this type may be underpowered to detect a clinically significant difference (Hofmeister *et al* 2007).

Here we conduct a scoping review of the literature to identify prospective research studies that have used the GCMPS- SF to measure acute perioperative pain in the dog. The aim of this study was threefold: (i) describe the use of the GCMPS- SF in terms of the features of research studies in which it has been employed; (ii) determine if the GCMPS-SF has been adopted in an appropriate manner to give valid results; and (iii) establish whether the study design of clinical trials employing the GCMPS-SF is such that these studies are well-designed and adequately powered.

2.3. Methods

2.3.1. Literature search

A systematic search of PubMed, CAB abstracts, Web of Science and Google Scholar for papers published between 2007 and 2019 (inclusive) was performed (see Appendix 2). Searches were carried out on each platform using combinations of the following key words (and derivatives): dogs (dog, dogs), the Glasgow Composite Measure Pain Scale—short form (GCMPS- SF, GCMPS, CMPS, CMPS-SF, Glasgow Composite Measure Pain Scale, GCMPS short form, CMPS short form, GCPS), postoperative (post operative, post-operative, postoperative) and pain. We also used the citing articles search feature in Google Scholar and Web of Science to identify any articles citing the original paper describing the development of the GCMPS- SF (Reid *et al* 2007).

2.3.2. Inclusion criteria

Publications were included if they met the following criteria: (i) use of the Glasgow CMPS-SF to assess pain; (ii) investigating acute post-operative pain; (iii) prospective design; (iv) use of the English language; (v) published in a peer-reviewed journal; (vi) conducted in dogs, and (vii) available in full to the authors. Only English language studies were included because validated translations of the GCMPS-SF only recently became widely available (Della Rocca *et al* 2018). Foreign language versions of psychometrically developed scales may not be valid and any assessment of validity must take into account the cultural and linguistic aspects of the target language (Della Rocca *et al* 2018). We felt that the potential inclusion of foreign language versions of the scale would make the interpretation of any results difficult as these would not be comparable without validation.

2.3.3. Data extraction and appraisal

Data extraction and coding was performed by one reviewer (BT) with the coding for each article independently reviewed (AB). Any discrepancies or queries were resolved by discussion and consensus. Before performing the review, a data extraction form was developed to extract information from the studies to fulfil the aims of our investigation and the sections were as described below. All data were derived from the manuscripts themselves or noted as not specified if details of a variable were not given. Authors were not contacted to gather further details.

2.3.4. Variables describing the publications

The year and journal were recorded from the website of the publisher. The country of origin of the research was defined as that of the first author's institution. Pain inducing procedures were classified as soft tissue, neurological or orthopaedic surgeries. We also recorded whether cases enrolled in a given study underwent the same single surgical procedure, or whether multiple different procedures were used. Any intervention(s) used in the studies was coded into the

classes: analgesic drugs, surgical techniques, regional anaesthesia techniques or alternative therapies. The “regional anaesthesia techniques” category was used for studies which compared regional anaesthesia techniques exclusively to each other.

Any other metrology instruments used for the measurement of pain or nociception alongside the GCMPS-SF were recorded. Finally, we assessed whether the GCMPS-SF was intended as a primary outcome measure in the study. This was determined to be the case if pain assessment was a major aim specified in the title or if a stated hypothesis or aim involved pain measurement.

2.3.5. Variables describing the use of the GCMPS-SF and measured data

We determined whether any modifications to the scale had been made. Section B of the scale (locomotion) may be omitted if the animal requires assistance to ambulate and therefore this was not counted as a modification. As an analgesic intervention threshold for the scale has been derived (greater than or equal to a score of 6/24 or 5/20 if section B is omitted), we recorded whether the appropriate intervention level had been used, or if this had been modified. We also recorded details of the number, type, and experience of those using the instrument.

The trial design for each study was first classed as either observational, i.e., containing a single group where all animals were treated the same, or controlled, where comparisons were made between two or more groups. We then divided the controlled studies into groups based on their stated hypothesis. Those trials where multiple groups were compared with the aim of disproving the null hypothesis were termed superiority trials, in contrast to those stating they were specifically designed to evaluate either equivalence or non-inferiority. We recorded whether any transformations were applied to GCMPS-SF scores prior to statistical testing. We also noted how authors approached the scores arising from animals after any provision of rescue analgesia; specifically, we asked whether these scores were excluded from further statistical analysis and whether a last observation carried

forward (LOCF) methodology was used. For controlled trials, the statistical techniques used to compare GCMPS-SF scores between groups were classified into the following broad classes, each class potentially encompassing a number of different specific statistical techniques: (i) parametric testing; (ii) non-parametric testing; and (iii) categorical comparisons of GCMP-SF scores after grouping into classes. For non-inferiority/equivalence trials a fourth group was required to allow for those studies using a confidence interval-based approach to non-inferiority testing. When scores from the GCMPS-SF were used to guide rescue analgesic provision, we recorded the statistical techniques used to compare rescue analgesic use and whether these involved comparing the proportions of animals rescued between groups or the mean number/dose of rescue analgesics required. We also noted any use of survival analysis statistics to compare the time to rescue between groups.

2.3.6. Variables describing the study design

We recorded whether each study was conducted across single or multiple centres. When client owned dogs were used as subjects, we termed these publications clinical studies. Where client owned dogs were not used, we used the term experimental study. Among the controlled studies, we recorded whether the authors clearly stated if the trial was randomized and blinded. We did not however record any further details of these parameters such as methods of blinding or randomization. Controlled studies were also classified by the type of control group used, i.e., the group to which the animals receiving the intervention are compared. In studies with a positive control each group received an analgesic which was assumed to provide the same degree of analgesia, e.g., one non-steroidal anti-inflammatory *versus* another. We described studies as negatively controlled if no effective analgesia was present at the time of pain scoring. This may have been due to placebo administration, or in some cases where a short acting analgesic was given at premedication (e.g., pethidine or fentanyl). However, this dichotomous scheme did not satisfactorily classify some publications and hence a third descriptor was used, pseudo- negative. In these studies, all groups had some form of analgesia present at the majority of timepoints of pain scoring. However, in one group, the analgesic or combination

of analgesics will be potentially less effective. An example of such a study would be where a nerve block is compared to sham but all dogs in the study received an NSAID pre-operatively.

We recorded the number of groups in each study, alongside the mean group size for each study and whether there was a >20% discrepancy in group sizes. While a discrepancy in group sizes is not necessarily problematic (Schulz *et al* 2002, Shibasaki *et al* 2018), the 20% threshold was arbitrarily defined as a level that was considered significant before data collection.

We determined the number of studies that had conducted an *a priori* sample size calculation and, among those, we recorded if the number of cases required was declared and whether sufficient dogs were recruited. In studies where we had determined pain, as measured by the GCMPS-SF, to be a primary outcome, we noted whether a sample size calculation was based specifically on pain score data. In order to evaluate the quality of sample size calculations, we used established criteria from the Consolidated Standards of Reporting Trials (CONSORT) guidelines (Moher *et al* 2010). The following elements were required for a study to be categorized as complying with CONSORT sample size guidelines: (i) the clinically important target difference between the groups; (ii) the α (type I) statistical error level; (iii) the statistical power (or the β (type II) statistical error level); (iv) the standard deviation (SD) of the measurements; and v) the source of the standard deviation used in the sample size analysis. We calculated a score out of ten for each sample size calculation based on the information provided. Two points were allocated for appropriate details given for each of the five required elements. With respect to the source of SD values, we allocated a single point to studies using unpublished preliminary data, and two points where a published study was cited as the source.

Finally, we recorded whether a statistically significant difference was found in each of the controlled superiority studies and whether this reflected differences in absolute pain scores, the provision of rescue analgesia, or both.

2.3.7. Statistical analysis

All coded variables were recorded in Microsoft Excel (Microsoft, Washington, U.S.). Summary statistics were generated in Jamovi (The Jamovi project, Sydney, Australia). A Mann-Whitney U test was used to compare group sizes between subgroups (non- inferiority *versus* superiority and sample size calculation *versus* no sample size calculation) with the p-value for significance set at <0.05.

2.4. Results

We identified 2,763 records through the database search. Following removal of duplicates, screening and full text eligibility assessment, 114 studies were finally included in the scoping review (Figure 2.1 and Appendix 3).

2.4.1. Variables describing the publications

The numbers of studies employing the GCMPS-SF per year are shown in Figure 2.2. The journals in which the studies were published and the country of origin of the research are described in Table 2.1. Single soft tissue and orthopaedic surgeries accounted for the majority of pain inducing procedures in the eligible studies (Table 2.1 and Appendix 3). The most common interventions investigated were analgesic drugs (Table 2.1). In 43% of the studies, another metrology instrument that measured pain or nociception was used alongside the GCMPS- SF (Table 2.1). Furthermore, we established that the GCMPS-SF represented a primary outcome measure in 73% of the studies included in this review.

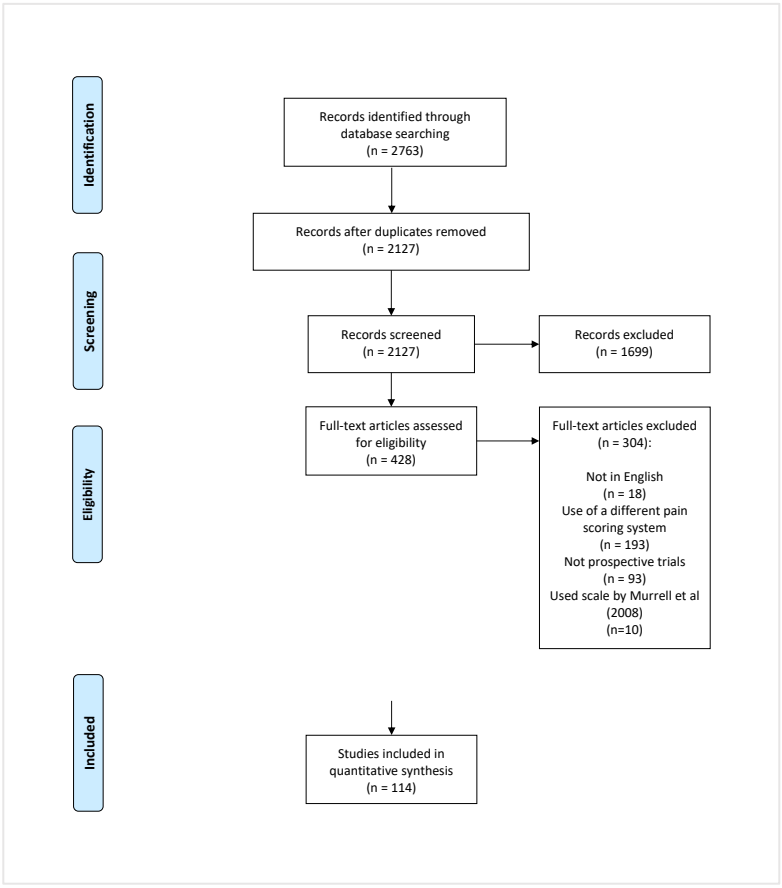

Fig. 2.1 Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flowchart showing the number of studies included in each stage of the review.

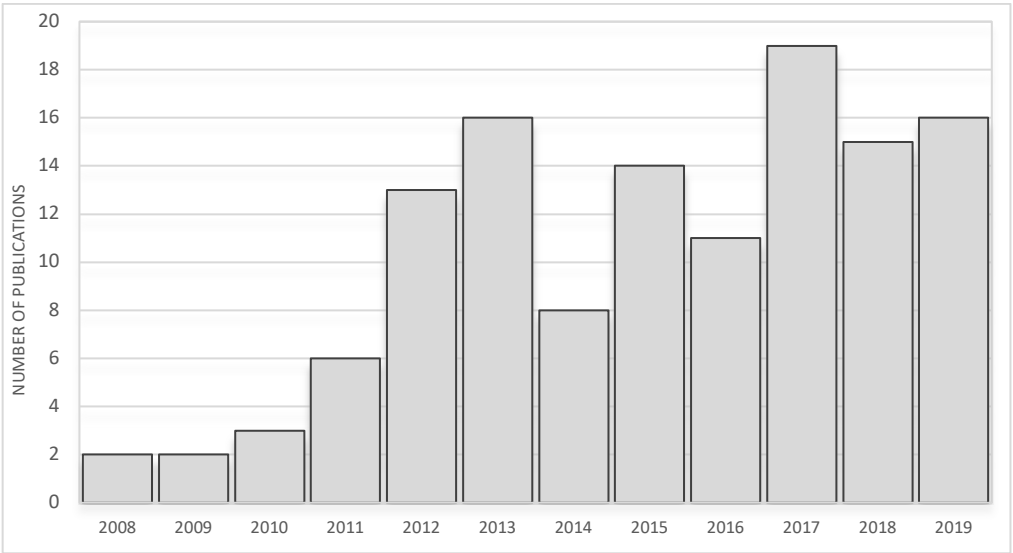

Fig. 2.2 The number of publications using the GCMPS-SF by year of publication between 2007 and 2019.

Table 2.1 Variables describing the publications included in the review. In the section describing other metrology instruments, the counts of studies in italics are not mutually exclusive. *Note: text in Italics denotes subgroups.*

Variable	Category	n =	%
Journal	Veterinary Anaesthesia & Analgesia	29	25%
	Journal of the American Animal Hospital Association	11	10%
	Veterinary Surgery	10	9%
	The Veterinary Journal	7	6%
	Journal of Small Animal Practice	7	6%
	BMC Veterinary Research	6	5%
	American Journal of Veterinary Research	5	4%
	Journal of Veterinary Pharmacology and Therapeutics	3	3%
	Journal of Veterinary Internal Medicine	3	3%
	Journal of Veterinary Behaviour	3	3%
	Veterinarni Medicina	3	3%
	Journals with fewer than 3 articles (n=21)	27	24%
		114	100%
Country of origin	USA	36	32%
	UK	18	16%
	Italy	12	11%
	Spain	7	6%
	Canada	5	4%
	Ireland	4	4%
	Switzerland	4	4%
	China	3	3%
	Countries with fewer than 3 studies (n=19)	25	22%
		114	100%
Pain Inducing Procedure	Single Soft Tissue Procedure	48	42%
	Single Orthopaedic Procedure	20	18%
	Mixed Soft Tissue Procedures	15	13%
	Mixed Procedures	12	11%
	Single Neurological Procedures	11	10%
	Mixed Orthopaedic Procedures	8	7%
		114	100%
Analgesic Intervention assessed	Analgesic Drugs	79	69%
	Surgical Techniques	15	13%
	Regional Anaesthesia Techniques	14	12%
	Alternative Therapies	6	5%
		114	100%

(continued)

Table 2.1 Continued.

Variable	Category	n =	%
Other 'Pain' Metrology Instruments Used	No	65	57%
	Yes	49	43%
	VAS	*28*	
	Mechanical thresholds	*21*	
	Other Composite Pain scale	*9*	
	NRS	*7*	
	Gait Analysis	*3*	
	Electroencepalography	*2*	
	Serum biomarkers	2	
		114	100%
Was GCMPS a primary outcome measure?	Yes	83	73%
	No	31	27%
		114	100%

2.4.2. Variables describing the use of the GCMPS-SF and measured data

The great majority of studies included in this review did not modify the GCMPS-SF (Table 2.2). However, in 7% of the publications some modification was evident. During the course of our review, we found 10 studies which purported to use the GCMPS-SF but on closer inspection actually used a modified version of the scale as proposed by Murrell *et al* (Murrell *et al* 2008). These studies were excluded from our analysis (Figure 2.1). In most studies investigated, the intervention level for rescue analgesia used was clearly stated as recommended for the scale (Table 2.2). However, changes to the intervention level were described in around a third of the studies. In some cases, these increases were only by one point (12 of 30 instances), although the mean increase in the intervention level was to 38% of the maximum GCMPS-SF score (~9/24, with a range of 7–18). An intervention score in excess of 10/24 was used by 7 papers in this review. We also determined details of the observers who performed scoring in each publication (Table 2.2).

Of the 114 studies included in the review, 104 (91%) were controlled studies comparing two or more groups. We classed 85 (82%) of these as superiority and 19 (18%) as either equivalence or non-inferiority studies. A variety of statistical approaches were used in the controlled studies to prepare and analyse GCMPS-SF data, and these are summarized in Table 2.3. All trials compared absolute pain scores between groups in some manner, and a smaller proportion (42 of 85 superiority studies and 15 of 19 non-inferiority studies) compared the use of rescue analgesia.

Table 2.2 Variables from publications in the review describing how the GCMPS-SF was used.

Variable	Category	n =	%
Modifications to The Scale	No	106	93%
	Yes	8	7%
	Omit/alter section A	*5*	
	Omit section C	*2*	
	Combine with physiological data	*1*	
		114	100%
Intervention level (for non-modified scale)	≥ 5/20 or ≥ 6/24	57	54%
	Increased	30	28%
	Decreased	3	3%
	Not specified/Based on other metrology (e.g. VAS)	16	15%
		106	100%
Observer			
Number	Single	51	45%
	Not specified	18	16%
	Multiple	45	39%
	2	*21*	
	3	*3*	
	4	*1*	
	not specified	*20*	
		114	100%
Pain scoring experience	Experienced/trained	34	30%
	Inexperienced	3	3%
	Not specified	77	68%
		114	100%
Type	Veterinary surgeon	32	28%
	Nurse/Technician	5	4%
	Veterinary student	2	2%
	Mixed	4	4%
	Not specified	71	62%
		114	100%

Table 2.3 A summary of handling data from the GCMPS-SF and the statistical techniques used. 104 studies are included in this table and the 10 observational studies in the review omitted. *Where numbers of studies using survival analysis are given, these may also be accounted for in the other groupings for comparing rescue analgesia use.

Variable	Category	n =	%
Data transformed prior to statistical testing?	No/Not specified	80	77%
	Transformed to normal (e.g. log transform)	10	10%
	Area under curve	5	5%
	Percentage of possible max	4	4%
	Pooled into classes	4	4%
	Change relative to baseline	1	1%
		104	100%
Data excluded after rescue analgesia?	Yes	32	31%
	No	45	43%
	Not applicable - no rescue required	5	5%
	Both analyses performed	3	3%
	Not specified	19	18%
		104	100%
LOCF Stated as being used	Yes	7	20%
	No	25	80%
		32	100%
Statistics for superiority trials			
Comparing pain scores	Parametric	36	42%
	Non-parametric	33	39%
	Categorical	1	1%
	No formal statistical testing	2	2%
	Not specified	13	15%
		85	100%
Comparing rescue analgesia use	Proportions requiring rescue compared	26	62%
	Means of rescue analgesic administration compared	5	12%
	Both means and proportions compared	4	10%
	Survival analysis (time to rescue) conducted	*13**	*31%**
		42	100%
Statistics for non-inferiority trials			
Comparing pain scores	Parametric	6	32%
	Non-parametric	5	26%
	Non-inferiority confidence intervals	2	11%
	Categorical	1	5%
	Not specified	5	26%
		19	100%
Comparing rescue analgesia use	Proportions requiring rescue compared	9	60%
	Means of rescue analgesic administration compared	4	27%
	Survival analysis (time to rescue) conducted	*3**	*19%**
		15	100%

2.4.3. Variables describing the study designs and power

Details of the study designs used in this review are detailed in Table 2.4. Of the 104 controlled trials, 50 (48%) had conducted an *a priori* sample size calculation for any outcome measure. In 48 of the 50 cases, the total number of dogs required was declared and in 41 of those cases sufficient dogs were recruited. The sample size calculation was performed as per CONSORT guidelines in 12 (24%) of the studies and the median sample size calculation score allocated was 6 (range 2–10). The GCMPS-SF represented a primary outcome measure in 36 of the 50 studies with sample size calculations, and yet a sample size calculation related specifically to the GCMPS-SF in only 24 (67%) of these. During coding of the studies, we noticed larger group sizes in those with a non- inferiority vs. superiority design (50 ± 69 *versus* 21 ± 30 (mean SD), p = 0.017) and in those that included a sample size calculation compared to those without (36 ± 53 *versus* 18 ± 22, p = 0.001).

We restricted further analysis of study findings to the 85 controlled studies with a superiority hypothesis. In 38 (45%) of these studies, statistically significant differences were evident, and this occurred between absolute scores (n = 21, 55%), guiding of rescue (n = 4, 11%), and both measures (n = 13, 34%).

Table 2.4 Variables describing features of study design in the publications.

Variable	Category	n =	%
Study Design			
Centre	Single centre	101	89%
	Multi centre	13	11%
		114	100%
Setting	Clinical	104	91%
	Experimental	10	9%
		114	100%
Cross over design	Yes (all within experimental studies)	2	2%
	No	112	98%
		114	100%
Among controlled studies (n=104)			
Randomised	Yes	104	100%
	No	0	0%
		104	100%
Blinded	Yes	90	87%
	No	14	13%
		104	100%
Control	Positive	60	58%
	Pseudo-negative	31	30%
	Negative	13	13%
		104	100%
Number of groups	Two	75	72%
	Three	20	19%
	4 or greater	9	9%
		104	100%
Dogs per group	Mean +/- SD		27 +/- 41
	Median (range)		15 (5 - 251)
>20% size discrepancy?	Yes	7	7%
	No	97	93%
		104	100%

2.5. Discussion

In this review, we demonstrate the widespread international use of the GCMPS-SF in the canine post-operative analgesia literature. The scale has been applied broadly across investigations into the effect of many different analgesic interventions on pain induced by a variety of surgical interventions. This popularity is perhaps unsurprising given the properties of the scale; namely that it is one of only a few validated tools for the measurement of acute pain in the dog (Firth *et al* 1999, Rialland *et al* 2012, Della Rocca *et al* 2019), and that the scale has a high utility and a defined intervention level (Reid *et al* 2007).

Our results demonstrate a number of noteworthy issues relating to the appropriate use of the scale and the design of the trials in which it has been employed. These considerations have the potential to significantly affect the outcome of studies. Therefore, mitigating against potential shortcomings as described below will be vital to the success of future veterinary clinical research using the GCMPS-SF and its translational potential.

2.5.1. Appropriate use of the GCMPS-SF and derived data

The GCMPS-SF was developed using a psychometric approach and the validity is dependent on it being used as intended. Modifications to the scale, conducted without revalidation, change the measurement properties and should be avoided. Modifications were found in 7% of the papers in this review, and it is reassuring that this practice is rare. The defined intervention level is also no longer valid if changes are made. We documented a significant number of studies in which the intervention level had been altered. Some of these may simply have been due to poor reporting (stating “greater than” rather than “greater than or equal to”), however many changes were intentional, lacked supporting documentation and therefore were presumably based purely on author opinion. The intervention level was derived during a multi-centre clinical study at three separate veterinary hospitals, using animals that had undergone a variety of surgical procedures (Reid *et al* 2007). It is possible that in some other contexts the score may need to be

refined to better reflect the needs of a certain population (e.g., feral dog neutering), and novel data would ideally be presented in support of this. It does however seem unlikely that substantial changes in the intervention score (i.e., >10) would be appropriate in any context. Indeed, some of the increases, including an intervention level of 18, detected in this review raise ethical considerations, as animals in severe pain would not receive rescue analgesia.

An aspect of GCMPS-SF use that is poorly reported in the literature presented here, and has the capacity to significantly alter results, is the number and the experience of the observer(s) conducting the scoring. By using specific descriptors, the scale is designed to reduce respondent bias and decrease the interobserver variability that has been reported with unidimensional subjective pain scales (Holton *et al* 1998a). Among expert observers this would appear to be the case when scoring videos of painful dogs (Hofmeister *et al* 2018), although the use of inexperienced observers is not recommended as agreement may be poor (Barletta *et al* 2016).

We detected a lack of consensus regarding the statistical approach to absolute GCMPS-SF scores. The statistical test used should reflect the nature of the measurement, and the short form of the GCMPS is a non-interval level measure (Morton *et al* 2005). The choice of analysis may also need to be pragmatic to account for complexity of the data, such as repeated measures taken from the same individual. A number of different transformations have been applied to GCMPS-SF data prior to statistical testing, predominantly to normalize the data and utilize more powerful parametric statistics. Given the non-interval nature of GCMPS- SF data, pooling into classes (representing no pain, mild pain etc.) is a highly appropriate technique, but was only used in a minority of studies, perhaps as cut-off values are likely to be arbitrarily defined. A number of different approaches for dealing with scores arising after animals had received rescue analgesia were also evident, including whether imputation techniques such as LOCF were used. A lack of consensus in this regard also exists in the human acute pain literature (Singla *et al* 2017). A minority of studies in this review sought to evaluate equivalence, however very few of these used the most appropriate statistical approach to this, namely defining a non-inferiority margin and calculating confidence intervals (Rehal *et al* 2016).

2.5.2. The design of acute pain clinical trials

We also examined the trials using the instrument in terms of their design and statistical power. Appropriate blinding and randomization are crucial in clinical trials to prevent bias. Significant deficits in reporting have been shown in this respect in the veterinary literature (Di Girolamo *et al* 2017, Rufiange *et al* 2019). Consistent with this, we noticed during our coding of the data that authors would frequently state the trial was randomized and blinded without giving explicit details. More detailed assessment of these features, e.g., the extent of blinding, is a core part of risk of bias assessments. However, we chose not to conduct these assessments in detail during this review as our investigations centred on the use of pain scoring outcomes rather than establishing (via subsequent metanalysis) whether a particular outcome was well-evidenced across a number of studies.

A limited number of trials which used no effective analgesics in the control group (negative controls) were included in this review despite studies of this design often resulting in larger outcome effect sizes. This infrequency likely reflects the possibility of undertreatment of pain in placebo-treated participants and the ethical implications of this which are a significant consideration in veterinary medicine (Slingsby 2010) as in human medicine (Gilron *et al* 2019).

The number of animals enrolled per group in studies in this review seems relatively low and may be associated with a limited power to detect a significant difference. We observed significantly greater group sizes in non-inferiority trials which may be a reflection of the statistical approach required to demonstrate non-inferiority. We also show that group sizes are larger in studies where an *a priori* sample size calculation is carried out. Major deficits in the power of small animal analgesia studies were identified in literature from over 15 years ago (Hofmeister *et al* 2007). Although our methodology is different, our data would suggest that justification of adequate statistical power is still a significant issue in the small animal pain literature, and this issue is seen more broadly across veterinary clinical trials (Giuffrida 2014, Rufiange *et al* 2019). We also find other deficits in methodologies relating to statistical power. Many studies used pain measurement as a primary outcome, however in some cases a single sample size calculation was conducted for another primary outcome measure, such as anaesthetic

requirement. This could result in a study underpowered to detect differences in pain scores and a consequent Type II statistical error. Additionally, many of the published sample size calculations do not comprise sufficient information to judge their appropriateness. Publication of animal research is often dependent on the inclusion of a sample size calculation in order to satisfy ARRIVE (Animal Research: Reporting of In Vivo Experiments) guidance (Leung *et al* 2018). However, our data would suggest this guidance is not being universally applied.

Within pain outcome measures, absolute scores and the requirement for rescue appear to be used interchangeably as measures of efficacy in the canine pain literature. The use of multiple outcome measures to define analgesic success has been promoted recently (Johnston *et al* 2021). However, the use of multiple primary outcome measures without a multiple-comparisons adjustment of the threshold for significance may predispose to Type I statistical errors (Gewandter *et al* 2014, Oyama *et al* 2017), even if each component part is underpowered. As our review spans the period during which the GCMPS-SF has been in existence, it is conceivable that authors of earlier studies did not have access to preliminary data upon which to base a sample size calculation. However, now that a significant body of GCMPS-SF data is available across a number of contexts, this should not be the case. Promoting accessibility of GCMPS- SF data will be important to encourage appropriate experimental design using a priori sample size calculations in future.

2.5.3. Limitations

There are number of potential limitations to our findings. Firstly, despite using broad search terms, there is a possibility that we have not included some eligible publications that used the GCMPS-SF and did not mention it in a way that was captured by our search. Furthermore, a number of the coded variables (e.g., superiority vs. equivalence, or identification of primary outcome) were coded somewhat subjectively based on the information that was available and this may not have been as originally intended by the primary authors. This reflects deficiencies in reporting evident in some of the included studies and is mirrored more widely in the analgesia literature (Leung *et al* 2018, Gewandter *et al* 2019).

This is especially important as poor quality of reporting may be associated with finding exaggerated effects (Page *et al* 2016). A number of solutions to this problem have been proposed, including submission checklists (Han *et al* 2017, Gewandter *et al* 2019). Prior registration of clinical trials is also an essential requirement in human studies, and requires that primary outcome measures, hypotheses, sample size calculations and proposed statistical testing are declared before commencing the trial. Trial registries are in their infancy in veterinary medicine (Murphey 2019), but, based on our findings, are to be recommended to those conducting companion animal pain research.

2.5.4. Conclusions

In conclusion, this review demonstrates widespread use of the GCMPS-SF across the canine acute pain literature. For the most part, the scale has been adopted in a valid manner with only a few reported modifications to the scale and the intervention level. However, we document several deficiencies in experimental reporting and design which may predispose to both Type I and Type II statistical errors.

2.6. Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fvets.2021.751949/full#supplementary-material

CHAPTER 3
ASSOCIATION OF STUDY DESIGN FACTORS WITH THE FINDING OF A SIGNIFICANT DIFFERENCE

3.1. Introduction

The reason researchers expect a study to detect a significant difference is based on whether one truly exists. However, given the high risk of type II statistical errors in the small animal pain literature (Hofmeister *et al* 2007), we hypothesized that study design factors could underlie some of the variability seen in the ability to detect a statistically significant difference in pain scores between study groups. Using our existing dataset from the scoping review, we anticipated the following factors might affect the ability of a study to detect a difference between groups:

- Control type
  Negative controls decrease the number of variables potentially affecting results, which might favour the finding of a true difference (Lipsitch *et al* 2010). Furthermore, the decrease in confounding factors might also produce a bigger actual effect size *versus* a positive control, again potentially favouring the finding of a true difference (Moser 2019).
- GCMPS-SF scores as a primary or secondary outcome.
  The desired difference in pain scores between groups might be included in the *a priori* power estimation when pain scores are a primary outcome, hence favouring the finding of a true difference. When pain scores represent secondary outcome measures, instead, the study might not be powered enough to detect an existing difference in this endpoint.
- Modifications made to the CMPs-SF scale
  Modifications of the scale will alter its measurement properties. If modifications are introduced without consensus or a validation process, a bias might be potentially introduced resulting in an altered ability to find a difference in either direction.
- Implementation of an *a priori* power calculation.

Estimation of the sample size confers a higher power to the study, because it ensures that enough participants are enrolled to detect a difference of specified magnitude. Hence, adequate power translates into an increased probability of finding a true difference and decreased probability of Type II statistical error.

- Assigned CONSORT score in our analysis.

  Adherence to CONSORT guidelines (Moher *et al* 2010) ensures adequacy and completeness of reporting. Deficiencies in reporting might be associated with findings exaggerated effects (Page *et al* 2016), hence favouring the finding of a difference even if there is not one. This increases the probability of Type I statistical error.

- Changes to the intervention level

  Dependent on the type of changes made, a bias might be introduced in either direction. Increases in the intervention level for provision of additional analgesia might result in finding exaggerated effects, hence favouring the finding of a spurious difference and predisposing to Type I statistical error.

- Group size

  A small group size is associated with low power, which will decrease the ability of the study to detect a true difference and predispose to Type II statistical error (Hofmeister *et al* 2007). In contrast, imbalances between groups may introduce a bias in either direction, leading to either a higher rate of Type I or Type II statistical error (Rusticus *et al* 2014).

- Blinding

  Inadequate use and/or reporting of blinding strategies is associated with finding exaggerated effect sizes (Rufiange *et al* 2019), hence potentially increasing the probability of finding a spurious difference and a resultant Type I statistical error.

- Randomisation

  Lack of randomisation introduces imbalanced variability between groups, which may introduce a bias in either direction undermining the validity of the study (Di Gerolamo *et al* 2017, Rufiange *et al* 2019).

- Clinical *versus* experimental

Experimental studies, characterised by decreased variability, yield an increased probability of finding a true difference between groups (Fiske *et al* 2005).

- Study powered specifically for GCMPS-SF scores

  As highlighted above, implementation of a sample size estimation specifically aimed at detecting a pre-defined difference in the GCMPS-SF scores increases the probability of finding an existing difference.

- Single *versus* multicentre

  Mainly associated with deficiencies in implementation and reporting of strategies such as randomisation and blinding (Bafeta *et al* 2012, Unverzagt *et al* 2013), larger effect sizes are observed in single centre compared to multicentre trials, increasing the probability of finding a difference.

- Type of surgical procedure

  Utilisation of a single standardised surgical procedure reduces variability, which confers a higher power to the study hence favouring the finding of a true difference (Rusticus *et al* 2014, Gilron *et al* 2019).

- Observer

  The experience of the individual and the number of observers may influence the assigned pain scores introducing a higher variability, which may lead to a decrease probability of finding a true difference.

- Statistical techniques used

  The statistical techniques employed to analyse the data and to deal with data arising from patients after administration of rescue analgesia may introduce a bias in either direction (Singla *et al* 2017), thus possibly leading to higher rates of either Type I or Type II statistical errors.

3.2. Methods

In order to assess the influence of the factors described above on whether a significant difference occurred, we used multivariable logistic regression. We restricted our analysis to the 83 controlled (not observational) studies with a superiority hypothesis reported in the scoping review presented in chapter 2.

We investigated whether study design variable could predict the finding of a significant difference in pain outcomes using binomial multivariable logistic regression in Jamovi (The Jamovi project, Sydney, Australia). Factors were selected for inclusion in the final model if they approached significance ($p < 0.1$) in univariable tests. For this purpose, as our data were not normally distributed, Chi square test or Mann-Whitney U test were used as appropriate. In detail, to investigate the correlation between two categorical variables (type of control group, GCMPS-SF scores used as a primary/secondary outcome, modifications to the scale yes/no, power calculation yes/no, level of intervention for provision of additional analgesia modified yes/no, blinding, clinical versus experimental study design, power calculation done for GCMPS-SF scores yes/no, single versus multicentre study design, type of surgery, type of observer, statistical technique used) and the finding of a statistically significant difference between groups a Chi square test was used. A Mann-Whitney U test was used to compare the sum ranks of two independent groups (difference in median CONSORT scores and group size in studies that did and studies that did not find a statistically significant difference between groups).

We then used the sequential model builder in Jamovi to perform multivariable binomial logistic regression and evaluated model strength using the Akaike Information Criterion (AIC) and McFadden's R^2. In more detail, all variables determined to be significant at the univariable level were added as separate blocks and for each sequential model, AIC and McFaddens R^2 values were generated. The final model presented had the lowest AIC and highest R^2 of the sequence and hence all variables were retained and are presented here. The level of significance was set at $p < 0.05$.

3.3. Results

In the 83 controlled superiority studies with statistical testing, a significant difference between groups was not found in 45 (54%) of the studies.

The results of univariable tests showed an association between the finding of a significance difference and the type of control group [X^2 (2, n = 83) = 9.94, p = .007], pain scores being a primary or secondary outcome [X^2 (1, n = 83) = 8.10, p

= .004], modifications apported to the scale [X^2 (1, n = 83) = 7.66, p = .006], implementation of a sample size estimation [X^2 (1, n = 83) = 3.05, p = .081], and being the study single or multicentre [X^2 (1, n = 83) = 3.04, p = .081] (Appendix 4). When investigating the association between the completeness of reporting of elements of the sample size estimation (CONSORT score) and the finding of a statistical significant difference, results of a Mann-Whitney U test indicated that the difference between CONSORT scores in studies that did not find a difference (Median = 4.00, range 0.00-10.0) with studies that did (Median = 0.00, range 0.00-10.0) was statistically significant, U ($n_{\text{no diff}}$ = 45, n_{diff} = 38) = 636, p = .031 (Appendix 4).

In univariable analysis, the finding of a statistically significant difference showed no association with modifications implemented in the intervention level for provision of rescue analgesia [X^2 (1, n = 83) = 1.23, p = .267], blinding [X^2 (1, n = 83) = 0.876, p = .349], sample size calculation specifically powered to detect a difference in GCMPS-SF scores [X^2 (1, n = 83) = 1.14, p = .285], the type of surgery [X^2 (5, n = 83) = 3.58, p = .612], the type of observer assigning the pain scores [X^2 (3, n = 83) = 2.99, p = .392], clinical *versus* experimental study design [X^2 (1, n = 83) = 0.063, p = .801], and the choice of parametric *versus* nonparametric tests used for analysis of pain scores [X^2 (1, n = 83) = 1.25, p = .263] (Appendix 4). When the correlation of finding a significant difference was investigated with the group size, the difference between the group size in studies that did not find a difference (Median = 30.0, range 12-120) with studies that did (Median = 32.0, range 16-358) was not statistically significant, U ($n_{\text{no diff}}$ = 45, n_{diff} = 38) = 743, p = .307 (Appendix 4).

As all studies reported to be randomised, there was not sufficient variability to perform statistical testing.

Results of univariable tests with $p < 0.1$ were then included in the multivariable model (Table 3.1).

Table 3.1 Results of multivariate binomial logistic regression to estimate the odds of finding a statistical difference in GCMPS-SF scores or GCMPS-SF guided rescue analgesia *versus* not finding a difference. The analysis was conducted on the 83 controlled studies with a superiority hypothesis and formal statistical testing. The McFaddens R^2 and AIC for the model were 0.251 and 114, respectively.

Predictor	Reference	Estimate	SE	Z	p	Odds ratio	95% Confidence Interval	
							Lower	Upper
Intercept		-0.3144	0.537	-0.585	0.558	0.73	0.2548	2.092
CONSORT Score		-0.1436	0.15	-0.958	0.338	0.866	0.6457	1.162
Nature of Control group?								
Negative	Pseudo-negative	2.0496	0.905	2.264	0.024	7.765	1.3164	45.798
Positive	Pseudo-negative	0.4829	0.546	0.884	0.377	1.621	0.5556	4.728
Pain scores a primary outcome?								
Primary	Secondary	-1.3477	0.654	-2.062	0.039	3.849	1.0690	13.885
Scale Modified?								
Yes	No	2.5288	1.116	2.266	0.023	12.539	1.4064	111.786
Power Calculation conducted?								
Yes	No	0.5365	1.035	0.518	0.604	1.71	0.2247	13.012
Single vs Multicentre								
Single	Multi	-0.130	1.343	-0.096	0.923	0.878	0.0631	12.222

As shown in Table 3.1, the use of a negative control group was associated with a probability more than 7 times higher to find a difference between groups (p = .024; 95% CI, 1.31 to 45.79) compared to the use of pseudo-negative controls. For the modifications implemented in the scale, our model predicted that modifying the scale was correlated with a 12 times higher likelihood to find a difference in outcomes between groups (p = .023; 95% CI, 1.40 to 111.78). The finding of a difference was also significantly associated with the use of the GCMPS-SF scores as a primary outcome, with a probability approximately 4 time higher compared to the use of pain scores as secondary endpoints (p = .039; 95% CI, 1.06 to 13.88). Our model identified no statistically significant correlation between the finding of a difference and the presentation of a power calculation (p = .60), the quality of reporting of the elements that constitute it (p = .33), or single *versus* multicentre study design (p = .92).

When evaluating the strength of our logistic regression model, the McFadden's R^2 coefficient of determination was 0.251. Our results indicates that 25% of the variation in finding a difference was explained by the model, suggesting that methodological factors might actually affect study outcomes in our dataset.

3.4. Discussion

We evaluated the association of multiple study design factors with the probability of finding a statistically significant difference between groups in a cohort of 83 controlled studies with a superiority design that utilised the GCMPS-SF to measure pain. Results of our analyses indicated that some of these factors could underlie a considerable proportion of the variability seen in the ability of an assay to detect a statistically significant difference.

The use of a negative controlled group was positively correlated with the finding of a significant difference both in univariable and multivariable analysis. This finding is perhaps not surprising, and in line with the reported veterinary literature. In fact, the decreased number of factors potentially influencing the response to treatment reduces sources of variability, conferring a higher power to the study (Lipsitch *et al* 2010). Furthermore, as the magnitude of attenuation of

the pain response will be the result solely of the intervention compared to no intervention (Moser 2019), this study design likely generates larger outcome effect sizes, which facilitate the finding of a difference. However, despite these potential advantages, a smaller proportion of studies in our dataset utilised a negative controlled study design. Of the 83 studies included, 70 studies used positive or pseudo-negative controls, and only 13 trials enrolled participants in a placebo group. Our results reflect a wider tendency of many recent clinical trials toward the administration of some form of analgesia in all groups (Slingsby 2010, Moser 2019), due to the ethical implications of undertreatment of pain in placebo-treated participants.

The GCMPS-SF scores, depending on whether used as primary or secondary endpoints, showed a significant correlation with the finding of a difference in our analysis. A significant proportion of studies in which pain scores represented a secondary outcome did not find a statistically significant difference between groups (17 out of 21) in univariable tests. This finding, albeit potentially related to the fact that a difference did not exist, might also be related to the study design. *A priori* power calculations are usually implemented during the planning stage of a research project to reasonably ensures that an adequate number of subjects is enrolled in the study to detect the desired effect size in the primary outcome/s (Hofmeister *et al* 2007). However, if pain scores are a secondary outcome, the assay is not powered for this endpoint and might not consequently be sensitive enough to detect a difference. Despite a moderately wide error in prediction, results of the logistic regression model also supported this finding. GCMPS-SF scores as a primary outcome were associated with an almost four times higher probability to find a significant difference between groups compared to their inclusion in the study as secondary outcome measures. Interestingly, when we investigated the possible association between the finding of a difference and sample size calculation conducted specifically for the GCMPS-SF scores, no significant correlation was found in univariable analysis. In our dataset, the vast majority of studies powered the sample size calculation for an endpoint other than the GCMPS-SF scores (22 out of 37), leaving only 15 studies that included a pre-selected difference in pain scores between groups in the power analysis. No inferences can therefore be made on this aspect of the study design in our cohort, as a larger number of trials would be required to investigate the relationship

between the desired effect size in GCMPS-SF scores and the finding of a significant difference.

Modifications apported to the GCMPS-SF have the potential to influence results. As the scale was developed using a psychometric approach, any alteration would alter its measurement properties. All the studies included in our analysis that modified the scale in some way found a statistically significant difference between groups (6 out of 6), and they were associated with a 12 times higher probability of finding a difference compared to studies that used the scale as originally intended. These results might reflect a considerable impact of the appropriate use of the pain measurement instrument on drawing correct conclusions, as altering the scale seems to be associated with the possibility of rejecting a true null hypothesis (false positive results). The exclusion of some categories during pain assessment, for example those dedicated to the interaction with the patient (such as response to touch), might remove pain behaviours actually displayed by the patient under evaluation (but not captured) from the total score. It is possible that this practice introduces a bias in the pain scores derived in the study, thus leading to a higher probability of Type I statistical error. However, the number of studies that modified the scale in our dataset is very low. While this is reassuring, it also limits the possibility of making strong inferences about our results.

When considering the modifications implemented in the level of intervention for provision of rescue analgesia, approximately one third of the trials in our cohort modified the intervention score (27 out of 83). Increases in the threshold for administration of additional analgesia might result in finding bigger effect sizes between groups, thus favouring the finding of a spurious difference. However, results of our univariable tests showed no significant correlation between this practice and the finding of a statistically significant difference. Many of the alterations implemented in the intervention level for provision of rescue analgesia in the included studies were related to poor reporting (‘greater than’ instead of ‘greater than or equal to’), thus changing the intervention score by only one point. Our study might have failed to establish a significant correlation because this degree of variation either might require a larger sample size to be detected or does not actually influence the finding of a difference. In order to explore the

interplay between the appropriate use of the GCMPS-SF and study results, and to obtain more conclusive answers, a larger sample size would be ideally required.

The role of the presentation of a power calculation and the quality of reporting of its items on study outcomes have been investigated extensively in the veterinary literature. Under-representation of a sample size estimation has been found to be prevalent in the veterinary analgesia literature, and frequently associated with studies that were underpowered to detect an existing difference between groups, thus leading to a high prevalence of Type II statistical error (Hofmeister *et al* 2007). Lack of adherence to CONSORT guidelines (Moher *et al* 2010) has also been found prevalent in the veterinary literature. Deficiencies in reporting of power calculations have been demonstrated to affect the power of a study significantly and negatively, as highlighted in the canine and feline literature (Giuffrida 2014) and in a review across research subjects and species (Rufiange *et al* 2019). In accordance with the published evidence, less than half of the trials included in our study (37 out of 83) presented a sample size estimation, hence confirming that this essential requirement for high quality research still remains under-represented. Differently from previous studies, we found no significant disparity in the finding of a statistical difference in the 46 studies that did not conduct an *a priori* sample size estimation. Moreover, the median CONSORT score in studies that did not find a significant difference between groups (4.00) was significantly higher than the CONSORT score in studies that did find a difference (0.00), which seems to suggest that poor reporting might favour the finding of a spurious difference. However, when these factors were tested in multivariable analysis, no association emerged from our model between the finding of a difference and either the presentation of a power calculation or the quality of reporting (p = .60 and p = .33, respectively). These results might be explained by our small sample size, which might have been not sensitive enough to detect an association. In fact, the association between results and study design factors in previous published reports was investigated in 238 trials (Giuffrida 2014) and 120 trials (Rufiange *et al* 2019), while the present thesis included 83 trials. Our methodology was also different, in that we restricted the analysis to superiority, controlled trials specifically utilising the GCMPS-SF. It is therefore possible that other factors related to our inclusion criteria might have had a determinant influence on results. With respect to our study design, for example,

we divided the included studies in arbitrarily defined negative / positive / pseudo-negative controls, albeit definitions of our methods were given. This arbitrary subdivision might have potentially introduced a bias into our data, also impacting our results.

Surprisingly, no other factors pertinent to the study design were found to be associated with the ability of the assay to detect a significant difference.
The explanation for some of these factors might lie in the small sample size represented in our cohort. For example, only 12 studies were not blinded, only 8 studies were experimental, there were only 8 multicentre trials, and all studies reported to be randomised. Although the extent and quality of reporting of such factors as blinding and randomisation also represent a potential source of bias that might affect results (Di Girolamo *et al* 2017, Rufiange *et al* 2019), the risk of bias assessment via subsequent metanalysis was not carried out in the present work.
Other factors were characterised by a high degree of variability in our dataset, such as the type of surgery used as the pain-inducing procedure and the type of observer assigning the pain scores. Increasing levels of variation are negatively correlated with the power of a study (Rusticus *et al* 2014, Gilron *et al* 2019), and may have adversely impacted the ability of the statistical tests applied to our data to detect a significant association. In addition, more than half of the trials included in this chapter did not specify the type of observer (49 out of 83). Poor reporting might have reduced the power of our analysis even further (Rufiange *et al* 2019), leading to inconclusive results.
The mean number of dogs enrolled in studies that found a significant difference was higher than that of studies that did not find a difference (69.3 *versus* 35), thus possibly confirming that studies with a smaller sample size might not be powered enough to detect a difference (Hofmeister *et al* 2007). However, our data were not normally distributed, and therefore analysis was conducted on the medians (30 *versus* 32). No conclusive inferences can be made from our data, as no significant correlation was detected in univariate analysis between the finding of a significant difference and the median group size.

3.4.1. Limitations

The observational design of this study is one of the main limitations, with particular consideration to our inability to obtain conclusive results when investigating some factors of trial design that were poorly represented. However, being an exploratory hypothesis generating study rather than confirmatory, this study represents pilot work which may inform future research. Another possible limitation is the subjective coding for some variables, such as the type of control group. Nevertheless, detailed definitions were given in the published paper [chapter 2 of this thesis (Testa *et al* 2021)], and coding was applied consistently to each publication and independently reviewed by two authors. It is also possible that eligible studies were excluded from our cohort because not captured by our search terms. However, as we conducted a systematic and reproducible search of the literature, this eventuality seems unlikely.

3.4.2. Conclusions

Our results indicate that a considerable proportion of the variability in finding a difference was explained by the model, possibly suggesting that methodological factors did actually affect study outcomes in our dataset. The probability of finding a statistically significant difference was seven times higher in studies that used negative control groups, 3 times higher when the GCMPS-SF scores were used as a primary outcome, and 12 times higher if the pain scale was altered. Consistently with the published veterinary literature, poor reporting in multiple variables was also observed in our cohort. Our findings stress the importance of methodologically sound study design in order to obtain valid results, as these will influence evidence-based medicine, comparability of findings between studies, and will constitute the basis for the conduct of future research.

CHAPTER 4
MAXIMUM DIFFERENCE DETECTED IN GCMPS-SF SCORES BETWEEN GROUPS IN STUDIES THAT DECLARED STATISTICAL SIGNIFICANCE

4.1. Introduction

The difference in pain scores between groups that is deemed relevant by the patient or the clinician and that would mandate changes in patient management is referred to as the minimum clinically important difference (MCID). This concept was introduced to differentiate between statistical significance and clinical relevance (Jaeschke *et al* 1989), which may not necessarily overlap. In fact, while results in a study may be declared statistically significant, very small statistically significant effects may not have clinical relevance. In addition to its clinical utility, the MCID has two major roles in research.

First, the ability to measure the MCID can be used to test the responsiveness of a pain measurement instrument during the validation process (Reid *et al* 2018). A study conducted on the measurement of pain in infants (Barr 1998) suggested that responsiveness of a scale could be estimated by applying an intervention of known efficacy and measuring the magnitude of change. This process was adopted during the development of the GCMPS (Morton *et al* 2005), where the type of surgery was used as the intervention of known efficacy. The sensitivity of the scale was tested against its ability to differentiate between severities of pain induced by soft tissue and orthopaedic surgeries, testing the hypothesis that orthopaedic surgery would generate higher pain scores than soft tissue surgery. For a pain scale, however, the administration of an analgesic drug of known efficacy would best represent an appropriate intervention to assess the smallest meaningful amount of change the scale can detect (Morton *et al* 2005).

Second, the MCID can represent the desired effect size to be incorporated in the *a priori* sample size estimation. This will determine the minimum relevant effect

in the outcome of interest the assay should be sensitive to in order to estimate an adequate sample size (Olsen *et al* 2017).

The MCID has been studied extensively in the human literature in numerous studies, reviews and metanalyses (Farrar *et al* 2000, Olsen *et al* 2017, Bahreini *et al* 2020). A relatively recent study outlines a considerable variability in the MCID reported in the human literature irrespective of the pain measurement instrument used (Bahreini *et al* 2020). A meta-analysis of human trials that assessed the MCID in acute pain in the emergency department reported absolute values for the MCID ranging from 8 to 40 mm on a 100-mm scale (Olsen *et al* 2017).

Paucity of information can be retrieved on the MCID in pain scores in the veterinary literature. The only mention is a minimum clinically relevant difference in GCMPS-SF scores of 2.6 used as the desired effect size by McKune *et al* (2014). The authors stated that this reported value was based on the MCID detected in GCMPS-SF scores in a previous work (Morton *et al* 2005). However, in the work conducted by Morton and colleagues in 2005, the value of 2.6 represented the mean GCMPS score detected in dogs that had undergone a surgical procedure compared to a mean value of 1.2 in dogs that had not undergone surgery. Therefore, in the original publication this value of 2.6 was not indicative of a difference between groups, nor it was regarded as the MCID. Consequently, the only mention of the MCID in the veterinary literature was actually based on a mis-citation. To the best of this thesis' authors' knowledge, no formal assessment of the MCID exists in the veterinary literature.

The aim of the present work was to investigate whether statistically significant differences detected in the studies included in this thesis were of a magnitude that would be considered clinically relevant. Using our existing dataset from the scoping review, we first mapped desired and actual effect sizes in GCMPS-SF scores across different surgical procedures and interventions versus control, in studies that declared statistical significance. We then compared these values and investigated whether a consensus value for the MCID could be suggested from the available veterinary literature.

4.2. Methods

In order to compare the desired and actual effect sizes of GCMPS-SF scores between studies, we restricted our analysis to the 39 trials that reported a statistically significant difference between groups in pain scores only, or in both pain scores and the requirements for rescue analgesia (based on GCMPS-SF scores).

We recorded the desired effect size included in the *a priori* sample size estimation. For the actual effect size, we recorded the largest difference detected in GCMPS-SF scores between groups at a single time point as reported in the results section of the publications.
We then subdivided our dataset with respect to the type of intervention (drug, regional anaesthesia, or surgery) and the type of surgery.

As a great variety of surgical procedures was displayed in our dataset (Appendix 5), we first grouped them into broader classes as shown in Table 4.1. Some of these procedures were poorly or not represented (multiple orthopaedic surgeries, neurological procedures, and mixed surgeries). After exclusion of mixed and neurological surgeries, this recoding process led to a total of 27 studies subdivided by the type of procedure into two broad classes, namely soft tissue and orthopaedic surgeries.

The largest difference in pain scores between groups was investigated in this cohort overall and in the above-mentioned subgroups using descriptive statistics in Jamovi (The Jamovi project, Sydney, Australia). Data are presented as summary tables or box plots. To investigate whether the differences detected in each study were significantly different from each other based on type of surgery or intervention, Mann-Whitney U test or Kruskal-Wallis test were used as appropriate. In detail, a Mann-Whitney U test was used to compare the median GCMPS-SF scores between two groups (orthopaedic *versus* soft tissue, TPLO *versus* OVH), while a Kruskal-Wallis test was applied when the comparison of the median GCMPS-SF scores was made between more than two groups (analgesic drugs, regional anaesthesia, and surgery). The level of significance was set at $p < 0.05$.

Table 4.1 Summary of the type of surgeries grouped into broader categories. Mixed: multiple different surgical procedures; Neuro: neurological procedures; ortho mixed: multiple orthopaedic procedures; ortho single: single standardised orthopaedic procedure; soft tissue mixed: multiple soft tissue surgeries; soft tissue single: single standardised soft tissue surgery.

TYPE OF SURGERY – BROAD CATEGORY	NUMBER OF STUDIES THAT EMPLOYED IT
N Mixed	1
Neuro	2
Ortho mixed	0
Ortho single	6
Soft tissue mixed	7
Soft tissue single	14

4.3. Results

Of the 39 trials that declared a statistically significant difference between groups amongst the 114 included in the scoping review, 9 studies did not report the actual difference, hence leaving 30 studies to be included in our analysis.

Approximately 75% of the studies did not present a sample size calculation (29/39), whilst amongst the 10 studies that did only 2 were powered for a difference in GCMPS-SF scores.

When considering the 30 studies overall, the largest difference found was 2.00 (median 2.00, range 1.00-11.0) (Fig. 4.1).

We then considered separately the median differences in pain scores according to the type of procedure and the type of intervention.

Soft tissue surgeries were the type of procedure most largely represented in the analgesia studies in our cohort (soft tissue *versus* ortho 21 *versus* 6, respectively).

The largest difference in GCMPS-SF scores was higher in the orthopaedic than in the soft tissue group [3.00 (range 1.50-4.00) *versus* 2.00 (range 1.00-5.50), respectively] (Fig. 4.2), albeit it was not statistically significant, U ($n_{\text{soft tissue}} = 21$, $n_{\text{ortho}} = 6$) = 52.0, $p = .540$.

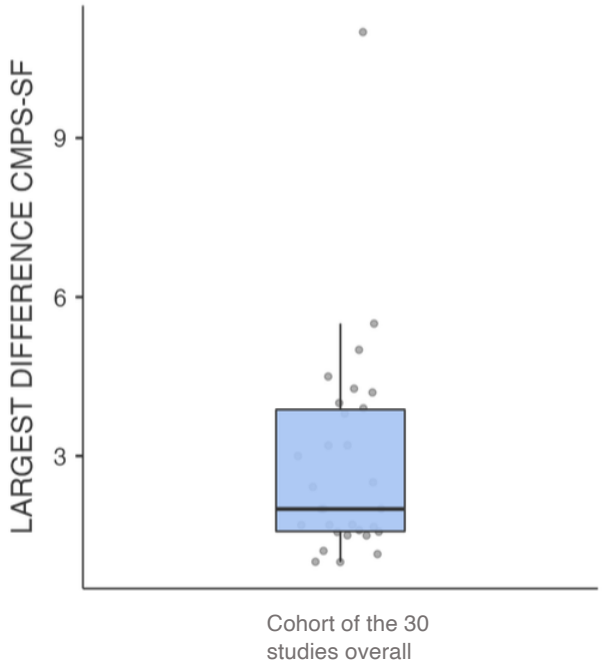

Figure 4.1 Box plot of the largest difference found in GCMPS-SF scores in the 30 studies that declared a statistically significant difference between groups. The box represents the 25th-75th quartile (interquartile range), the horizontal line within the box represents the median, the vertical lines (whiskers) represent the minimum and maximum value, and outliers are shown as dots beyond the whiskers.

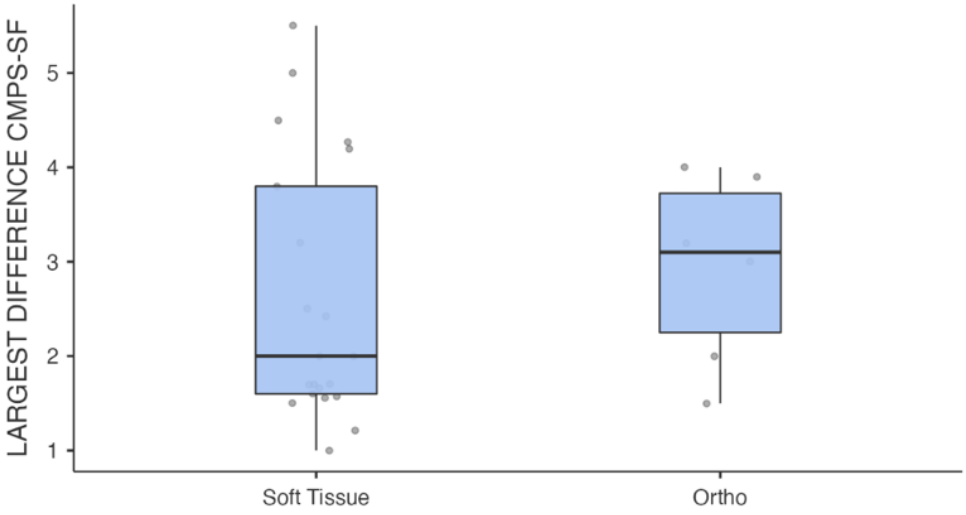

Figure 4.2 Box plot of the largest difference in GCMPS-sf scores detected in soft tissue and orthopaedic surgeries. The box represents the 25th-75th quartile (interquartile range), the horizontal line within the box represents the median, the vertical lines (whiskers) represent the minimum and maximum value, and outliers are shown as dots beyond the whiskers.

With the consideration that tibial plateau levelling osteotomy (TPLO) and ovariohysterectomy (OVH) were the most represented single orthopaedic and soft tissue procedures in our dataset, we further refined the type of surgery to investigate the differences in pain scores in studies that utilised these two single standardised surgeries. The largest difference in GCMPS-SF scores was higher in the TPLO than in the OVH group [2.20 (range 1.00-3.80) *versus* 1.80 (range 1.50-5.50), respectively] (Fig. 4.3). However, also in this analysis, the difference between groups was not statistically significant, U (n_{OVH} = 8, n_{TPLO} = 4) = 10.5, p = .392.

The type of interventions investigated in the 30 studies included in this chapter is summarised in Table 4.2. The use of analgesic drugs represented the most largely utilised intervention in this cohort of pain studies (24/30).

Largest differences in GCMPS-SF scores in studies that investigated the effects of analgesic drugs (Median 2.00, range 1.00-11.0), regional anaesthesia (Median 3.00, range 1.00-4.00), and surgery (Median 4.00, range 1.00-4.50) (Fig. 4.4) were not statistically different from each other [X^2 (2, n = 30) = 0.127, p = .938].

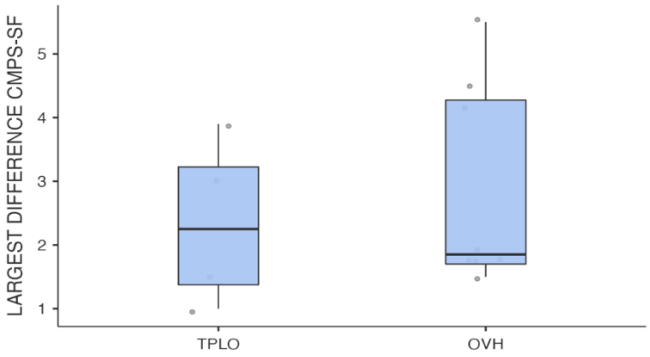

Figure 4.3 Box plot of the largest difference in GCMPS-sf scores detected in TPLO (tibial plateau levelling osteotomy) and OVH (ovariohysterectomy) surgeries. The box represents the 25th-75th quartile (interquartile range), the horizontal line within the box represents the median, the vertical lines (whiskers) represent the minimum and maximum value, and outliers are shown as dots beyond the whiskers.

Table 4.2 Summary of type of intervention utilised by the 30 included studies.

	TYPE OF INTERVENTION	NUMBER OF STUDIES THAT EMPLOYED IT
N	Drug	24
	Regional anaesthesia	3
	Surgery	3

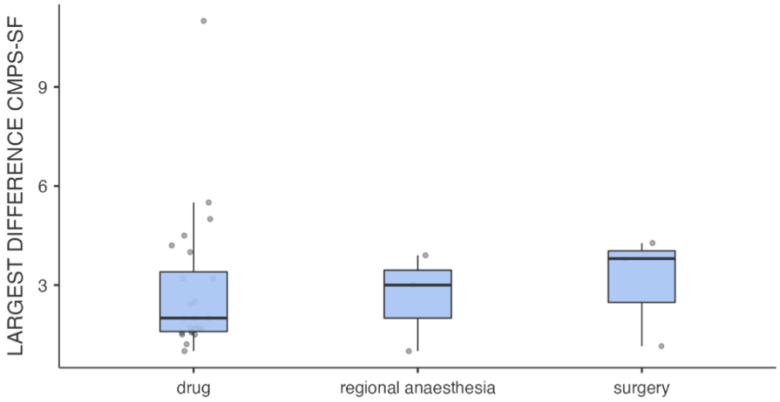

Figure 4.4. Box plot of the largest difference in GCMPS-sf scores based on the intervention category (drug, regional anaesthesia, and surgery). The box represents the 25th-75th quartile (interquartile range), the horizontal line within the box represents the median, the vertical lines (whiskers) represent the minimum and maximum value, and outliers are shown as dots beyond the whiskers.

4.4. Discussion

The investigation of the desired effect size in a cohort of 30 studies that declared a statistically significant difference between groups in pain scores revealed that approximately 75% of the trials did not conduct an *a priori* sample size estimation, and that only one quarter of the trials that did powered the study for a difference in GCMPS-SF scores. When we compared the largest differences in pain scores detected across a variety of interventions and surgical procedures, results of our analyses indicated a median GCMPS-SF score difference of 2.00, with no statistically significant differences between any of the subgroups tested.

4.4.1. Comparison of findings between surgical procedures

Orthopaedic procedures were associated with a difference in pain scores between groups higher than that detected in soft tissue surgeries (3.00 *versus* 2.00, respectively). In support of the hypothesis that orthopaedic surgery induces higher levels of pain than soft tissue (Morton *et al* 2005), this result seems to provide further evidence of the responsiveness of the GCMPS-SF in clinical studies. However, the difference in median scores between subgroups was only by 1 point, and it was not statistically significant.
A number of possible explanations can be postulated for this finding.

First, our sample size was small (27 studies) and was characterised by a marked inequality between groups (21 soft tissue *versus* 6 ortho).

In addition, the studies that investigated pain utilising soft tissue surgical procedures were characterised by a high degree of variation in the difference detected between groups, ranging from 1.00 to 11.0. These differences were possibly influenced by the great variation observed in multiple factors of the study design, such as the type of control group, the use of a modified scale, modifications to the intervention level for the provision of additional analgesia, the rescue analgesia plan, the methods employed to deal with data arising from

subjects that had received rescue, the presentation of a power calculation, and the outcome of interest the study was powered for. Considering the heterogenicity of study designs and the low number of trials that were included in our analysis, comparability of findings between studies was markedly affected, and it is, in the authors' opinion, one of the reasons imputable for the lack of significance of our results.

Differently from soft tissue, orthopaedic procedures showed much less variability in the differences detected between groups. This finding is perhaps not surprising, as orthopaedic procedures are considered more painful (Morton *et al* 2005), consequently leading researchers to include various combinations of multimodal analgesia in all study arms due to ethical implications, especially in clinical studies involving client-owned participants (Hansen 2003). This was reflected in our dataset, which encompassed a widespread use of positive and pseudo-negative control groups. This practice, resulting in lower levels of pain in all study groups, has been associated with the finding of smaller effect sizes (Moser 2019) and may have decreased the power of our study to detect a meaningful difference between groups (Page *et al* 2016).

When we investigated median differences in the restricted subset of data arising from the use of single surgical procedures (TPLO and OVH), our analysis produced results very similar to those observed in the broader subdivision of multiple orthopaedic and soft tissue surgeries. On one hand, the similar trends observed in these two subsets of data might support the validity of our findings. On the other hand, it would be expected that the decreased variability associated with a single standardised surgical procedure would generate bigger effect sizes (Rusticus *et al* 2014, Gilron *et al* 2019). However, the paucity of data included in this restricted analysis was a determinant factor in preventing such an evaluation.

4.4.2. Comparison of findings between interventions

The most utilised intervention was represented by the administration of analgesic drugs (24/30). Differences between groups with different interventions (drugs, regional anaesthesia, and surgery) were not statistically significant, with a largest

median difference detected of 2.00 points between the administration of analgesics and surgery groups, consistent with that detected when all the 30 studies were considered together.

The ability to measure the MCID is a fundamental aspect to assess the responsiveness of a measurement instrument for clinical use (Reid *et al* 2018). The responsiveness of a scale can be tested by applying an intervention of known efficacy and measuring the magnitude of change. This intervention may be represented by a surgical procedure (Hansen 2003, Morton *et al* 2005), in order to test the ability of the scale to differentiate between severities of pain induced by different standardised surgical procedures. However, as this might potentially generate big actual effect sizes, the smallest meaningful amount of change the scale can detect would be best assessed by the administration of an analgesic drug (Morton *et al* 2005).

In our study, the non-significant difference in pain scores between interventions was likely not imputable to lack of responsiveness of the GCMPS-SF. Instead, it might be the result of the low power of our study, arising from the small sample size (especially for the 'regional anaesthesia' and 'surgery' groups, encompassing only 3 studies each) and the magnitude of inequality in group sizes.

4.4.3. General considerations irrespective of the division in subgroups

Of the 114 studies included in the scoping review, only 39 found a statistically significant difference in pain scores between groups. In accordance with the published veterinary literature, this might reflect a common problem in small animal analgesia studies, which are frequently underpowered to detect an existing effect (Hofmeister *et al* 2007). Furthermore, approximately one quarter of the studies that declared statistical significance (9/39) did not report the actual difference found. Poor/selective reporting observed here is similarly mirrored more widely in the veterinary analgesia literature (Leung *et al* 2018, Gewandter *et al* 2019, Rufiange *et al* 2019).

We also document a great variability in the desired effect size the study was powered for and the statistically significant actual effect size detected.

4.4.3.1. Desired effect size

With respect to the desired effect size, some studies presented a sample size estimation powered for the difference between groups in isoflurane settings (McMillan *et al* 2012, Hunt *et al* 2013), duration of analgesia (Adami *et al* 2016), peak vertical force (Gurney *et al* 2012), incision site complications (Travis *et al* 2017), and fentanyl infusion rates (Palomba *et al* 2019). Of the only two studies powered specifically for the detection of a difference in GCMPS-SF scores, one set the desired effect size in pain scores at 0.37 points at the 24-hour assessment post-administration of the study intervention (Lascelles *et al* 2016). Many studies did not report a sample size estimation, examples of which are Bienhoff *et al* 2012, Kim *et al* 2012, Kongara *et al* 2012, Morgaz *et al* 2013, Perez *et al* 2013, Goudie-De Angelis *et al* 2016, and Friton *et al* 2017 (references of the publications cited in this paragraph can be found in Appendix 3 – supplementary references).

When investigating the clinical relevance of an outcome, the desired effect size plays a determinant role. If a study is powered for a different primary outcome measure than pain scores, or a sample size estimation is not presented, the study might not be sensitive enough to detect a significant difference in this endpoint (Hofmeister *et al* 2007). Hence, if a significant difference in pain scores between groups is detected, although potentially representing an actual existing difference, it might also be related to other aspects of the study design, such as modifications apported to the scale or the intervention for provision of additional analgesia, selective reporting, blinding and randomisation strategies. Despite multiple factors pertinent to the study design demonstrated an association with the finding of a difference (chapter 3), the risk of bias assessment via subsequent metanalysis was not part of this thesis. Therefore, a conclusive impact of potential sources of bias, such as factors of the study design that predispose to the finding of a difference, could not be established in this cohort.

Differently, if a very small, desired effect size in the difference in GCMPS-SF scores between groups is included in the sample size estimation (for example the 0.37 points mentioned previously - Lascelles *et al* 2016), the study will be sensitive to very small actual effect sizes, which might be statistically significant but not clinically relevant (Jaeschke *et al* 1989).
Given the high variability in the desired effect size studies were powered for, and the scarcity of studies powered specifically for a difference in pain scores, it is not possible to suggest an average desired effect size in GCMPS-SF scores utilised in the literature from our dataset.

4.4.3.2. Actual effect size

Considerable variation was also observed in the actual effect sizes reported as statistically significant. The median largest difference in GCMPS-SF scores between groups was 2.00 at a single time point of the post-operative assessment, implying that this difference was some degrees smaller at all the other time points. The range was considerably wide, from 1.00 to 11.0-point difference in GCMPS-SF scores between groups, with approximately 30% of the values (9/30) lying in the range of 1.15-1.60.

In order to investigate the clinical relevance of a statistically significant difference, it is worth considering how the MCID can be derived with the pain measurement instrument used. With subjective scales, like the VAS and the NRS, pain scores are anchored to the observer's perception of the patient's level of pain. Pain scores can then be integrated with the clinical assessment for requirements for additional analgesia to evaluate cut-off points. The minimum difference that would mandate changes in patient management can be considered clinically important and regarded as the MCID, if a consistent value is established across a variety of patients and conditions. These above explained criteria, used to define the minimum meaningful response to treatment with subjective pain scales, might differ according to the pain measurement instrument used. In fact, the clinical relevance of an outcome might change depending on whether we are using a VAS (more subjective) or the GCMPS-SF (that has an intervention level for provision of additional analgesia). Clinical relevance is perhaps more anchored to

the intervention level with the GCMPS-SF, rather than to a minimum difference in pain scores. Exemplified, a pain score difference of 3 might not be considered clinically relevant if the scores compared were below the intervention level (for example between 1 and 4 out of 24). In contrast, a 1 point-increase might be sufficient to mandate changes in patient management if it reached the threshold for administration of additional analgesia (for example from 5 to 6 out of 24). Hence, a 1- or a 3-point difference in pain scores between groups might or might not be clinically relevant with the GCMPS-SF, because this depends on the position of the score on the scale. With the VAS instead, the clinical relevance of a difference in values is more anchored to the value *per se*.
It is therefore questionable whether many of the differences detected, albeit statistically significant, are clinically relevant without accounting for their position on the scale.

In order to establish the clinical relevance of an outcome with the GCMPS-SF, it may be more meaningful to evaluate the proportion of dogs that required rescue analgesia *versus* those that did not. For this purpose, research would be required to determine a minimum consistent difference in the proportions of analgesic requirements that is deemed clinically relevant across multiple contexts and populations.

4.4.4. Limitations

One of the main limitations of this study is the observational design, influenced by factors such as low power, unequal group sizes, and the high degree of variability that characterised some of our subgroups. Another limitation is represented by the subjective recoding of surgical procedures. Grouping them into classes led to loss of some categories, with consequent loss of a small amount of data. However, the categories removed encompassed widely scattered and minimally represented data, which could have confounded our analysis without apporting any additional power to our study. Finally, as the risk of bias assessment was not conducted, the influence of study design factors that predispose to the finding of a difference on study results could not be evaluated.

4.4.5. Conclusions

Our results indicate that the largest median difference in pain scores between groups was 2.00 in this cohort of 30 trials that declared statistical significance. Given the wide range of statistically significant values reported and our small sample size, it was not possible to derive a consensus value from our data. Considerable variability was also identified in the desired effect sizes, with only two trials specifically powered for the detection of a difference in GCMPS-SF scores between groups. Consistently with the published veterinary literature, lack of presentation of a sample size estimation and poor reporting were also prevalent in our cohort. The numerous deficiencies and variability in study design potentially affected the clinical relevance of the results reported. In the authors' opinion, the clinical relevance of a set difference in pain scores without accounting for their position on the scale is questionable with a pain measurement instrument provided with an intervention level for administration of additional analgesia. When utilising the GCMPS-SF, it might be more meaningful to determine clinical relevance by comparing the proportions of dogs requiring / not requiring rescue analgesia.

CHAPTER 5
SUMMARY DISCUSSION

5.1. General discussion

The findings of our study highlight a high degree of methodological variability and multiple methodological deficiencies within prospective trials that utilised the GCMPS-SF to measure post-operative acute pain in dogs. A number of factors pertaining to the study design were identified that may predispose to Type I and Type II statistical errors. A median largest actual effect size in GCMPS-SF scores of 2.00 was detected, which appeared to be influenced by various aspects of the study design.

As one of few validated tools to measure post-operative acute pain in dogs, the GCMPS-SF was used widely in terms of geography and the contexts in which it was employed. We also document a considerable use of the scale in non-English speaking countries, despite the existence of validated context-related and linguistic translations (Murrell *et al* 2008, Della Rocca *et al* 2018). As previously reported (Tait *et al* 2011), and as shown from our data, the scale has been utilised to assess the efficacy of many different analgesic interventions on pain induced by a variety of surgical procedures, thus supporting its usefulness in a clinical setting and reinforcing the validity of its content.

We document a minor proportion of studies that apported non-validated modifications to the scale (7% of the papers included in the scoping review), which have the potential to affect results due to alteration of its measurement properties. All the studies included in the analysis presented in chapter 3 that had modified the scale in some way found a statistically significant difference between groups. These studies were associated with a 12 times higher probability of finding a difference compared to studies that used the scale as originally intended. These results might reflect the considerable impact of the appropriate use of the pain measurement instrument on drawing correct conclusions, as altering the scale

may be associated with an increased probability of incurring positive results which could represent Type I errors.

When considering the level of intervention for provision of rescue analgesia, this was not specified in a considerable proportion of studies (15%), and was modified in approximately one third of the trials in our cohort. While the intervention score was decreased in a negligible number of trials (3%), the majority of alterations related to increases in the threshold for administration of additional analgesia (15%). On many occasions, these increases were imputable to poor reporting ('greater than' instead of 'greater than or equal to'), thus changing the intervention score by only one point. However, some of these changes were arbitrary and substantial (some studies reported an intervention score between 11 and 15 out of 24). Reassuringly this practice was rare, although it raises ethical concerns as patients in severe pain would not receive any analgesia. Such changes might also result in finding exaggerated effect sizes between groups, which favour the finding of a difference (Moser 2019). Nevertheless, results of univariable tests conducted in chapter 3 showed no significant association between this practice and the finding of a statistically significant difference. As the majority of changes were by one point, our study might have failed to establish a significant association because this degree of variation either might require a larger sample size to be detected or does not actually influence the finding of a difference.

Aspects of the study design such as the type and number of evaluators assigning the pain scores have been shown to have the potential to influence study outcomes (Barletta *et al* 2016, Hofmeister *et al* 2018). Two major findings of our scoping review are a considerable variation between studies in both the type and number of observers detailed and a prevalence of poor reporting. With respect to the latter finding, the type and the number of evaluators were not specified in 16% and 62% of the studies included in the scoping review, respectively. When we investigated the association between observer-related factors and study outcomes, no significant association was detected in our analysis, possibly due to the high variability in our dataset and the prevalence of poor reporting.

Consistent with the previous veterinary literature (Di Girolamo *et al* 2017, Rufiange *et al* 2019), we also observed a high prevalence of selective reporting in

blinding and randomisation strategies. Although the extent and quality of reporting of these factors represent a potential source of bias that might affect results (Di Girolamo *et al* 2017, Rufiange *et al* 2019), a full risk of bias assessment was beyond the aims of this thesis. As all studies reported to be randomised, the association between randomisation and study outcomes could not be investigated in the present work.

Extreme variability was observed in the statistical approaches used to analyse absolute pain scores, in terms of the type of statistical tests utilised, the transformations applied to GCMPS-SF scores prior to statistical testing, and the approaches adopted to deal with data arising from animals after receiving rescue analgesia. All these factors have been demonstrated to have the potential to introduce a bias in either direction (Singla *et al* 2017), leading to higher rates of either Type I or Type II statistical errors.

The use of a negative controlled group was associated with a seven-time higher probability of finding a significant difference between groups in our multivariate analysis. This finding is perhaps not surprising, and in line with the reported veterinary literature. In fact, the decreased number of factors potentially influencing the response to treatment reduces sources of variability, conferring a higher power to the study (Lipsitch *et al* 2010). Furthermore, as the magnitude of attenuation of the pain response will be the result solely of the intervention compared to no intervention (Moser 2019), this study design likely generates larger outcome effect sizes, which facilitate the finding of a difference. However, despite these potential advantages, the use of negative controls was underrepresented amongst the controlled studies included in this thesis (13/104). Our results reflect a wider tendency of many recent clinical trials toward the administration of some form of analgesia in all study groups (Slingsby 2010, Moser 2019), due to the ethical implications of undertreatment of pain in placebo-treated participants.

Overall, the median number of patients per group was relatively low (15) in this scoping review, with a great variability observed (range 5-251). Relatively larger group sizes were noticed in studies with a non-inferiority *versus* superiority hypothesis (mean 50 ± 69 SD *versus* mean 21 ± 30 SD, respectively), possibly

reflecting the different statistical approach required to demonstrate non-inferiority. When we investigated the association between group size and study outcomes in the 83 controlled studies with a superiority hypothesis, the mean number of dogs enrolled in trials that found a significant difference was higher than that of trials that did not find a difference (69.3 *versus* 35, respectively), thus possibly confirming that studies with a smaller sample size might be underpowered to detect a difference, although this difference between groups was not statistically significant.

Under-reporting of a sample size estimation has been found prevalent in the veterinary analgesia literature, and frequently associated with studies that were underpowered to detect an existing difference between groups (Hofmeister *et al* 2007).

In accordance with the published evidence, less than half of the controlled trials included in our review (50/104) presented a sample size estimation. Similar results were obtained when we restricted the observation to the controlled trials with a superiority design (37/83), and to the 39 trials that declared a statistically significant difference between groups (10/39). Hence, our findings confirm that this essential requirement for high quality research still remains under-reported.

In contrast to previous studies (Hofmeister *et al* 2007, Giuffrida 2014), no significant association was found between presentation of a sample size estimation and the finding of a difference in the 46 superiority trials investigated in chapter 3. Furthermore, a sample size calculation was not presented in a great proportion of trials that found a statistically significant difference between groups (8/10), again possibly contrasting the previous evidence that lack of a sample size calculation predisposes to false negative results. However, the significance difference detected in these studies might also result from an actual difference between the cohorts, alongside methodological factors such as modifications apported to the scale, selective reporting, blinding and randomisation strategies.

We also identified other methodological deficits related to statistical power. The majority of studies that presented a sample size estimation in our review used a difference between groups in GCMPS-SF scores as a primary outcome (72%), yet only 67% of these were specifically powered to detect a difference in GCMPS-SF scores. More marked deficits were noticed among the superiority trials analysed

in chapter 3 and the trials that declared a significant difference analyzed in chapter 4, where the sample size calculation was powered for a primary endpoint other than the GCMPS-SF scores in 59% and in 75% of the trials, respectively.
Additionally, quality of reporting of items encompassed in the sample size calculation was found to be poor in the studies included in this review, as only 24% of the reported sample size calculations comprised sufficient information to judge their appropriateness.
The methodological deficits related to the presentation of a power calculation for a different primary endpoint and selective reporting have been demonstrated also in the wider veterinary literature (Giuffrida 2014, Rufiange *et al* 2019).
In the restricted analysis to the 83 superiority trials, the median CONSORT score in studies that did not find a significant difference between groups (4.00) was significantly higher than the CONSORT score in studies that did find a difference (0.00). In contrast with the published veterinary literature, where poor reporting has been associated with a significant decrease in the study power (Giuffrida 2014, Rufiange *et al* 2019), our findings seem to suggest that poor reporting might favour the finding of a spurious difference. However, our multivariate analysis showed no association between the finding of a difference and either the presentation of a power calculation or the quality of reporting (p = .60 and p = .33, respectively).

The 81% of superiority, controlled studies in which the GCMPS-SF scores represented a secondary outcome did not find a statistically significant difference between groups in these terms. This finding, although potentially related to the fact that a difference did not exist, might also be a consequence of the study design. In fact, if pain scores are a secondary outcome, the assay is not powered for this endpoint and might not consequently be sensitive enough to detect a difference.

In the separate analysis conducted on the 30 trials that declared statistical significance we detected a median largest difference in GCMPS-SF scores between groups of 2.00 (range 1.00 to 11.0), with approximately 30% of the values lying in the range of 1.15-1.60. As in many cases the differences were small, the clinical relevance of some of the statistically significant differences in the literature may be questionable. Additionally, considering that clinical relevance is perhaps more anchored to the intervention level with the GCMPS-SF, rather than to a minimum

difference in pain scores, we question whether many of the differences detected, albeit statistically significant, are clinically relevant without accounting for their position on the scale. In the authors' opinion, to establish the clinical relevance of an outcome with the GCMPS-SF, it may be more meaningful to evaluate the proportion of dogs that required rescue analgesia *versus* those that did not. For this purpose, research would be required to determine a minimum consistent difference in the proportions of analgesic requirements that is deemed clinically relevant across multiple contexts and populations. No formal studies have sought to establish a MCID for the GCMPS-SF, and it is not possible to suggest an average desired effect size in GCMPS-SF scores utilised in the literature from our dataset.

Furthermore, we did not reveal any meaningful difference in the magnitude of the actual effect size in GCMPS-SF pain scores between different types of surgery or interventions. Despite a seemingly higher difference in pain scores between groups in studies that utilised orthopaedic *versus* soft tissue surgeries (3.00 *versus* 2.00, respectively), this difference was not statistically significant.

5.1.1. Limitations

As an observational scoping review of the literature, the sample of studies included is fixed based on what has been conducted by other investigators. As such, in many of our formal analyses, unequal group sizes and a high degree of variability were present. Although we have attempted to statistically assess the effect of study design on outcomes and the finding of a statistically significant difference, these deficiencies did affect our ability to do so on many occasions. One major difficulty with our approach is that the actual ground-truth difference (the reality of the situation) in between groups which a study aimed to determine (whether the study drugs/interventions actually make a difference to pain scores), is not known. For example, most studies in the pain literature are presumably driven by the hypothesis that one drug is more effective than another and that pain scores in the superior drug group will be lower. Statistical approaches to proving this essentially imply sampling from that group in sufficient numbers to convincingly predict that that ground-truth/reality is that one drug is actually superior. However in some cases, for instance due to methodological deficiencies,

we fail to see the real difference in drug efficacy between groups, which is instead what we would see if all studies were devoid of Type I and Type II statistical errors. One would expect the magnitude of the actual effect to be a major predictor in the finding of a significant difference and we cannot account for that. We do however account for around a quarter of the variability in outcome by using study design factors as predictors and feel that this approach is informative and yields results that one would expect based on commonly reported deficiencies in the literature (Ioannidis 2005).

Another possible limitation is the subjective coding of some variables, such as the type of control group and superiority *versus* equivalence hypothesis, which was based on the information available in the publications. This reflects deficiencies in reporting evident in some of the included studies, as mirrored more widely in the analgesia literature (Leung *et al* 2018, Gerwandter *et al* 2019). Nevertheless, detailed definitions were given in our methods, and coding was applied consistently to each publication and independently reviewed by two authors. Furthermore, subjective grouping of surgical procedures into broader classes led to loss of some categories in one of our studies, with consequent loss of a small amount of data. However, the categories removed encompassed widely scattered and minimally represented data, which could have confounded our analysis without apporting any additional power to our analysis.

It is also possible that eligible studies were excluded from our cohort because not captured by our search terms. However, as we conducted a systematic and reproducible search of the literature, this eventuality seems unlikely.

5.1.2. Conclusions

This review demonstrates widespread use of the CMPS-SF across the canine acute pain literature.

For the most part, the scale has been adopted in a valid manner with only a few reported modifications to the scale and the intervention level. However, our results suggest that methodological factors did influence study outcomes in our dataset. The probability of finding a statistically significant difference was 7 times

higher in studies that used negative control groups, 3 times higher when the GCMPS-SF scores were used as a primary outcome, and 12 times higher if the pain scale was altered.

We detected a median largest actual effect size in pain scores of 2.00 in trials that declared statistical significance. In the authors' opinion, the clinical relevance of a set difference in pain scores without accounting for their position on the scale is questionable with a pain measurement instrument provided with an intervention level for administration of additional analgesia. When utilising the GCMPS-SF, it might be more meaningful to determine clinical relevance by comparing the proportions of dogs requiring / not requiring rescue analgesia.

Consistently with the published veterinary literature, we document several deficiencies in experimental reporting and design which may predispose to both Type I and Type II statistical errors. Based on our findings, we stress the importance of methodologically sound study design in order to obtain valid results, as these will influence evidence-based medicine, comparability of findings between studies, and will constitute the basis for the conduct of future research.

A number of solutions to this problem have been proposed, including submission of checklists and trial registries. Prior registration of clinical trials (trial registries) is also an essential requirement in human studies, and requires that primary outcome measures, hypotheses, sample size calculations and proposed statistical testing are declared before commencing the trial. Submission of checklists refers to checklists to be completed (in accordance with the CONSORT guidelines) by authors and reviewers to ensure that all the fundamental aspects of a randomised controlled trial mentioned above have been addressed and reported with transparency.
Trial registries and checklists are in their infancy in veterinary medicine, but, based on our findings, and as also already suggested in the veterinary literature (Murphey 2019), are to be recommended to those conducting companion animal pain research to overcome study weaknesses associated with the study design.

Appendix 1

Omission of the definitions of descriptors in the GCMPS-SF – authors' rationale

was modelled. However, during the development of the CMPS, items deemed to be redundant were excluded, and so the approach of Melzack (1987) was considered unlikely to be effective. Instead, the shortening of the CMPS consisted primarily of measures taken to reduce the time taken to complete the questionnaire, so increasing its usefulness. Although five items were removed, six were added, making a net increase of one item in the CMPS-SF compared with the CMPS. According to Landgraf and Abetz (1996), a useful clinical instrument must not only be valid, reliable and responsive, but also be 'practical and easy to administer, score and interpret'. Even if an instrument is valid and reliable, it may not be useful if it requires lengthy training, if it is time-consuming to administer, or if scoring is complex (Streiner 1993). Accordingly, it was decided to use a ranking system for the items in each category since this would simplify the scoring process and shorten the time taken to complete the questionnaire. Substitution of a rank number for the calculated weight converts the scale from interval to ordinal in nature, with a consequent decrease in level of precision. Interval level measurement provides more precise measurement, which is necessary for research purposes, hence its use in the CMPS. However, an ordinal scale was considered to have sufficient precision for the clinical purpose for which this instrument was being designed. The use of a ranking system can introduce some indirect weighting to the scale when there is an unequal number of items in each category. In the CMPS the category 'demeanour' contains seven items which would have ranked scores zero to six, assuming that 'happy and bouncy' would represent no pain, and the maximum score would be six. By comparison, 'comfort' contains only two items so the maximum score in this category would be one, yet demeanour is not known to be more important than comfort when measuring pain (Holton 2000). It was to minimise this bias that the number of items in each category was balanced as much as possible by combining those categories containing few items or by splitting combinations of word descriptors where appropriate, within each category. During the development of the CMPS the individual words in each combination (quiet/indifferent; licking/looking/rubbing) had been allocated the same weight, but the authors felt justified in splitting these and allocating ranked scores on the basis of clinical experience. These processes resulted in the CMPS-SF being better balanced in terms of number of items per category than the CMPS; CMPS-SF — six categories, four of which contain five items, one contains four items, and one contains six items; CMPS — seven categories, one category with seven items, two with five, one with four, two with three and one with two.

Videotaped data collected by Fox *et al* (2000) of canine behaviour following ovariohysterectomy demonstrated that pain modifies both spontaneous and interactive behaviour and thus accurate pain assessment must take account of both. Consequently, it was decided to retain the examination protocol devised for the CMPS. However, it was felt that the original mobility category was ambiguous in that 'assessment not carried out' did not make clear whether the animal elected not to move or if it was incapable of movement, or if movement was contraindicated for medical reasons. To resolve this confusion, 'refuses to move' was substituted for 'assessment not carried out' and the observer was instructed to omit the mobility assessment in those cases where moving the animal was contraindicated. Accordingly the total score for such animals is reduced by four, and although this would be likely to cause problems with statistical analysis in a group of dogs containing both mobile and immobile dogs, it was considered a satisfactory solution for the clinical purpose for which the CMPS-SF was designed.

Pre-testing and consideration of the layout of the categories in the CMPS indicated that its design was not optimal in terms of efficiency of use. Accordingly the order of items in each category was reversed and the categories were rearranged so that the CMPS-SF consisted of four distinct sections, A, B, C and D. 'Vocalisation' and 'attention to wound' are concerned with the animal's spontaneous behaviour and comprise section A, while in sections B and C, 'mobility' and 'response to touch' are interactive. It was considered that 'demeanour' and the combined 'posture' and 'activity' categories would best represent the observer's overall impression of the dog's well-being and so should be scored last (section D).

In non-verbal patients the difficulties of pain assessment are magnified, because the lack of effective communication means that assessment relies on the recognition and interpretation of behavioural signs by an independent observer; inter-observer variability has been shown to be unacceptable for the visual analogue scale when used to assess pain in the dog (Holton *et al* 1998). The problem of inter-observer variability has been addressed during the development of tools to monitor other functions such as the level of consciousness, notably in the widely recognised Glasgow Coma Scale (GCS) (Teasdale & Jennett 1974). This is a scale that focusses on three different aspects of behavioural response. The universality of the scale depends on identifying responses that can be clearly defined, and this was the approach adopted for the CMPS. Clear and specific definitions of each item used in the scale were provided for the user of the questionnaire. However, reference to the list of definitions added considerably to the time taken to complete the questionnaire; therefore, because all of the words had dictionary definitions and were in general use, it was decided to omit the definitions from the CMPS-SF. This and the other steps taken to streamline the questionnaire reduced the time taken to complete it from over 10 min for the CMPS to approximately 2 min for the CMPS-SF. However, removing the definitions may have affected the reliability with which different observers used the instrument. Additionally, two factors may have introduced bias to the intervention study: the fact that the same person generated the pain score and assessed whether or not the dog required analgesia; and the fact that some dogs were included in the study because the ward nurse believed them to be in pain. This may have affected the clinicians' judgement as to whether or not an animal required analgesia. However, this

Text extrapolated from Reid *et al* 2007. Authors' rationale for omission of definitions of descriptors in the GCMPS-SF is embordered in red.

Appendix 2

Search strategy

Search strategy	A systematic search of four bibliographic databases (PubMed, CAB abstracts, Web of Science and Google Scholar) was conducted for papers published between 2007 and 2019 (inclusive) using either Safari or Google Chrome as web browsers. Searches were carried out on each database using a combination of the following key words (and derivatives): dogs (dog *OR* dogs) *AND* the Glasgow Composite Measure Pain Scale - short form (GCMPS-SF *OR* GCMPS *OR* CMPS *OR* CMPS-SF *OR* Glasgow Composite Measure Pain Scale *OR* GCMPS short form *OR* CMPS short form *OR* GCPS) *AND* postoperative (post operative *OR* post-operative *OR* postoperative) *AND* pain. We first conducted restricted searches of titles and abstracts based on the terms 'dog' *AND* 'CMPS-SF' *AND* 'postoperative' AND 'pain', subsequently broadening our searches using the terms 'CMPS-SF' *AND* "dog' and their derivatives. However, neither of these search results contained several papers that the authors knew of that would have fully satisfied the inclusion criteria. Therefore, we adopted an additional broader search strategy using the terms 'postoperative' *AND* 'pain' *AND* 'dog' and their derivatives. As an example the detailed Pubmed search strategy is given below. Additional studies were identified by browsing the reference list of the included papers and by using the citing articles search feature in Google Scholar and Web of Science to identify any articles citing the original paper describing the development of the CMPS-SF (10).
Inclusion criteria	Each publication was initially assessed against the inclusion and exclusion criteria based on the title, abstract and further reading if necessary. Publications were included if they met the following criteria: (i) use of the Glasgow CMPS-SF to assess pain; (ii) investigating acute postoperative pain; (iii) prospective design; (iv) use of the English language; (v) published in a peer-reviewed journal; (vi) conducted in dogs, and (vii) available in full to the authors.

Pubmed search strategy. Keyword searches in the title and abstract of articles are marked with the syntax [tiab]. Results from searches #8, #9 and #10 were assessed further against inclusion criteria.

Search #	Search strategy
#1	"Dog" [tiab] OR "Dogs" [tiab]
#2	"Glasgow Composite Measure Pain Scale – Short Form" [tiab] OR "GCMPS-SF" [tiab] OR "GCMPS" [tiab] OR "GCMPS short form" [tiab] OR "CMPS-SF" [tiab] OR "CMPS" [tiab] OR "CMPS short form" [tiab] OR "Glasgow Composite Measure Pain Scale" [tiab] OR "GCPS" [tiab]
#3	"Postoperative" [tiab] OR "Post-operative" [tiab] OR "Post operative" [tiab]
#4	"Pain" [tiab]
#5	#1 AND #2 AND #3 AND #4
#6	#1 AND #2
#7	#1 AND #3 AND #4
#8	#5 Filters: English; 2007:2019
#9	#6 Filters: English; 2007:2019
#10	#7 Filters: English; 2007:2019

Appendix 3

Studies Using the GCMPS-SF included in the review. Supplementary table and supplementary references

Supplementary table 2 – Studies Using the GCMPS-SF included in the review. TPLO: tibial plateau leveling osteotomy; EHPSS: extra-hepatic portosystemic shunt; C-section: caesarean section.

Reference	Year	Clinical/ Experimental	Observational vs Comparative	Procedure	GCMPS-SF as Primary or Secondary Outcome
Adami et al. (1)	*2016*	Clinical	Controlled	TPLO	Primary
Adami et al. (2)	*2012*	Clinical	Controlled	TPLO	Primary
Aengwanich et al. (3)	*2019*	Clinical	Observational	Castration	Primary
Aghighi et al. (4)	*2012*	Clinical	Controlled	Hemilaminectomy	Primary
Amenegual et al. (5)	*2017*	Clinical	Controlled	Spinal decompressive surgery	Secondary
Andreoni et al (6)	*2009*	Clinical	Observational	Various elective surgeries	Secondary
Aprea et al (7)	*2012*	Clinical	Controlled	Dorsal hemilaminectomy	Primary
Barker et al (8)	*2013*	Clinical	Controlled	Hemilaminectomy	Primary
Barnes et al (9)	*2019*	Clinical	Controlled	TPLO	Primary
Bartel et al (10)	*2016*	Clinical	Controlled	Stifle arthroplasty	Primary
Bellei et al (11)	*2011*	Clinical	Observational	Spinal surgery	Primary
Bendinelli et al (12)	*2018*	Clinical	Controlled	Combined laparoscopic ovariectomy and laparoscopic-assisted gastropexy	Primary
Benitez et al (13)	*2015*	Clinical	Controlled	TPLO	Primary
Benitez et al (14)	*2015*	Clinical	Controlled	TPLO	Secondary
Bienhoff et al (15)	*2011*	Clinical	Controlled	Dental surgery (dental extraction)	Primary
Bienhoff et al (16)	*2012*	Clinical	Controlled	Soft tissue surgery	Primary
Bustamante et al (17)	*2018*	Clinical	Controlled	Ovariohysterectomy	Primary
Caniglia et al (18)	*2012*	Clinical	Controlled	TPLO	Primary
Cerasoli et al (19)	*2017*	Clinical	Controlled	TPLO	Primary

Chiavaccini et al (20)	*2017*	Experimental	Controlled	Thoracic skin incisions	Primary
Dancker et al (21)	*2019*	Clinical	Controlled	EHPSS attenuation	Primary
Davila et al (22)	*2013*	Clinical	Controlled	TPLO	Primary
Fitzpatrick et al (23)	*2010*	Clinical	Controlled	Ovariohysterectomy	Primary
Fizzano et al (24)	*2017*	Experimental	Controlled	Rhinoscopy and nasal biopsies	Secondary
Fransson et al (25)	*2015*	Clinical	Controlled	Laparoscopic Ovariohysterectomy	Primary
Friton et al (26)	*2017*	Clinical	Controlled	Soft tissue surgery	Primary
Friton et al (27)	*2017*	Clinical	Controlled	Soft tissue surgery	Primary
Giudice et al (28)	*2017*	Clinical	Controlled	Hemilaminectomy (acute vertebral disc extrusion)	Primary
Goudie-DeAngelis et al (29)	*2016*	Clinical	Controlled	Ovariohysterectomy	Primary
Groppetti et al (30)	*2019*	Clinical	Observational	C-section	Secondary
Gruet et al (31)	*2011*	Clinical	Controlled	Major orthopaedic surgery	Primary
Gruet et al (32)	*2013*	Clinical	Controlled	Major soft tissue surgery	Primary
Guerrero et al (33)	*2015*	Clinical	Controlled	Ovariohysterectomy	Primary
Guerrero et al (34)	*2016*	Clinical	Controlled	Ovariohysterectomy	Primary
Guillot et al (35)	*2011*	Experimental	Controlled	Bone Marrow Aspirate	Primary
Guimaraes Alves et al (36)	*2014*	Clinical	Controlled	Femoral, tibial, humeral or radial fracture repair	Primary
Gurney et al (37)	*2012*	Clinical	Controlled	Unilateral elbow arthroscopy	Primary
Gutierrez-Bautista et al (38)	*2018*	Clinical	Controlled	Orthopaedic surgery	Primary
Gutierrez-Blanco et al (39)	*2015*	Clinical	Controlled	Ovariohysterectomy	Primary
Hamilton et al (40)	*2014*	Clinical	Controlled	Orchidectomy	Secondary
Heffernan et al (41)	*2018*	Clinical	Controlled	TPLO	Primary
Hettlich et al (42)	*2017*	Clinical	Controlled	Hemilaminectomy	Primary
Hu et al (43)	*2017*	Clinical	Controlled	Ovariohysterectomy	Primary
Hunt et al (44)	*2013*	Clinical	Controlled	Orthopaedic surgery	Primary
Hunt et al (45)	*2013*	Clinical	Controlled	Various surgeries	Primary
Hunt et al (46)	*2014*	Clinical	Controlled	Mixed surgeries	Primary
Huuskonen et al (47)	*2013*	Clinical	Controlled	Castration	Secondary

Kaka et al (48)	*2018*	Clinical	Controlled	Ovariohysterectomy	Primary
Karrasch et al (49)	*2015*	Clinical	Controlled	Cutaneous tumour removal	Primary
Kibar et al (50)	*2019*	Clinical	Controlled	Ovariohysterectomy	Primary
Kim J.H. et al (51)	*2018*	Experimental	Controlled	Arthroscopic surgery (shoulder)	Primary
Kim Y.K. et al (52)	*2012*	Clinical	Controlled	Laparoscopic ovariohysterectomy	Primary
Kondo et al (53)	*2012*	Clinical	Controlled	Soft tissue surgery	Primary
Kongara et al (54)	*2012*	Clinical	Controlled	Ovariohysterectomy	Primary
Kongara et al (55)	*2013*	Clinical	Controlled	Castration	Primary
Kropf et al (56)	*2018*	Clinical	Controlled	Ovariohysterectomy	Secondary
Kropf et al (57)	*2019*	Clinical	Controlled	Ovariohysterectomy or castration	Secondary
Kushnir et al (58)	*2017*	Clinical	Controlled	Castration	Primary
Lambertini et al (59)	*2018*	Clinical	Controlled	Ovariohysterectomy	Primary
Lardone et al (60)	*2017*	Clinical	Controlled	Hip arthroplasty	Primary
Lascelles et al (61)	*2016*	Clinical	Controlled	Lateral retinacular suture procedure, including stifle arthrotomy	Primary
Lewis et al (62)	*2014*	Clinical	Controlled	TPLO	Primary
Li et al (63)	*2017*	Experimental	Controlled	Ovariohysterectomy	Secondary
Linton et al (64)	*2012*	Clinical	Controlled	Soft tissue or orthopaedic surgery	Primary
Little et al (65)	*2016*	Experimental	Observational	Surgical removal of cartilage from the head of the femur	Secondary
Luna et al (66)	*2015*	Clinical	Controlled	Ovariohysterectomy	Primary
Lush et al (67)	*2018*	Clinical	Observational	Castration	Primary
Martinez et al (68)	*2014*	Clinical	Controlled	Mixed surgeries	Primary
Martinez-Taboada et al (69)	*2017*	Clinical	Controlled	Elective surgery of the pelvic limbs or caudal abdomen	Secondary
McCally et al (70)	*2015*	Clinical	Controlled	TPLO	Primary
McKune et al (71)	*2014*	Clinical	Controlled	Ovariohysterectomy	Primary
McMillan et al (72)	*2012*	Clinical	Controlled	Castration	Secondary
Meakin et al (73)	*2016*	Clinical	Controlled	Abdominal surgeries, midline celiotomy	Secondary
Merema et al (74)	*2017*	Clinical	Controlled	Ovariohysterectomy	Primary
Meunier et al (75)	*2019*	Clinical	Controlled	Sterilisation	Primary
Morgaz et al (76)	*2013*	Clinical	Controlled	Ovariohysterectomy	Primary

Morgaz et al (77)	*2014*	Clinical	Controlled	Ovariohysterectomy	Primary
Nour et al (78)	*2013*	Experimental	Controlled	Not specified	Primary
Palomba et al (79)	*2019*	Clinical	Controlled	TPLO	Primary
Pascal et al (80)	*2013*	Clinical	Controlled	Genital surgery	Primary
Pascal et al (81)	*2019*	Clinical	Controlled	Hemilaminectomy	Primary
Peeters et al (82)	*2011*	Clinical	Controlled	Ovariohysterectomy or ovariectomy	Secondary
Perez et al (83)	*2013*	Clinical	Controlled	Castration	Primary
Perry et al (84)	*2015*	Clinical	Controlled	Distal limbs orthopaedic surgery	Secondary
Portela et al (85)	*2012*	Clinical	Observational	Orthopaedic surgery of the pelvic limb	Secondary
Re Bravo et al (86)	*2019*	Clinical	Controlled	Hemilaminectomy (single acute vertebral disc extrusion)	Primary
Read et al (87)	*2019*	Experimental	Controlled	Lateral thoracotomy	Primary
Reece et al (88)	*2012*	Clinical	Observational	Ovariohysterectomy	Secondary
Rioja et al (89)	*2012*	Clinical	Controlled	Ovariohysterectomy	Secondary
Romano et al (90)	*2016*	Clinical	Controlled	TPLO	Secondary
Sarotti et al (91)	*2015*	Clinical	Controlled	Pelvic limb orthopaedic surgery above the knee	Secondary
Sarotti et al (92)	*2019*	Clinical	Controlled	Hindlimb surgery	Secondary
Scott et al (93)	*2018*	Experimental	Controlled	Laparoscopy	Secondary
Shah et al (94)	*2018*	Clinical	Controlled	Ovariohysterectomy	Primary
Shih et al (95)	*2008*	Clinical	Controlled	Ovariohysterectomy	Primary
Shilo-Benjamini et al (96)	*2019*	Clinical	Controlled	Enucleation	Primary
Shivley et al (97)	*2018*	Clinical	Controlled	Ovariohysterectomy (suspensory ligament: sharp transection or digital strumming)	Secondary
Skelding et al (98)	*2019*	Clinical	Controlled	Various surgical procedures of the thoracic limb	Primary
Srithunyarat et al (99)	*2016*	Clinical	Observational	Ovariohysterectomy	Primary
Swallow et al (100)	*2017*	Clinical	Controlled	Ovariohysterectomy	Secondary
Tallant et al (101)	*2016*	Clinical	Controlled	Ovariohysterectomy/O variectomy	Primary
Tayari et al (102)	*2017*	Clinical	Controlled	TPLO	Primary
Tayari et al (103)	*2019*	Clinical	Observational	Thoracic limb orthopaedic surgery (distal to the mid-humerus)	Secondary

Travis et al (104)	*2017*	Clinical	Controlled	Midline celiotomy	Secondary
Valtolina et al (105)	*2009*	Clinical	Controlled	Exploratory laparotomy, thoracotomy, orthopaedic surgery	Primary
Vettorato et al (106)	*2010*	Clinical	Controlled	TPLO	Primary
Wagner et al (107)	*2008*	Clinical	Controlled	Castration or Ovariohysterectomy	Primary
Wagner et al (108)	*2010*	Clinical	Controlled	Forelimb amputation	Primary
Wang-Leandro et al (109)	*2019*	Experimental	Controlled	Tranresctal intraprostatic steam application	Secondary
Watanabe et al (110)	*2018*	Clinical	Controlled	Ovariohysterectomy	Primary
Weil et al (111)	*2016*	Clinical	Controlled	Soft tissue or orthopaedic surgery	Secondary
Zhang et al (112)	*2017*	Clinical	Controlled	Ovariohysterectomy	Primary
Zidan et al (113)	*2018*	Clinical	Controlled	Hemilaminectomy	Secondary
Zidan et al (114)	*2018*	Clinical	Controlled	Hemilaminectomy	Secondary

Supplementary References

1. Adami C, Casoni D, Noussitou F, Rytz U, Spadavecchia C. Addition of magnesium sulphate to ropivacaine for spinal analgesia in dogs undergoing tibial plateau levelling osteotomy. *Vet J* (2016) **209**:163–168. doi:10.1016/j.tvjl.2015.11.017

2. Adami C, Veres-Nyéki K, Spadavecchia C, Rytz U, Bergadano A. Evaluation of peri-operative epidural analgesia with ropivacaine, ropivacaine and sufentanil, and ropivacaine, sufentanil and epinephrine in isoflurane anesthetized dogs undergoing tibial plateau levelling osteotomy. *Vet J* (2012) **194**:229–234. doi:10.1016/j.tvjl.2012.04.019

3. Aengwanich W, Sakundech K, Chompoosan C, Tuchpramuk P, Boonsorn T. Physiological changes, pain stress, oxidative stress, and total antioxidant capacity before, during, and after castration in male dogs. *Journal of Veterinary Behavior* (2019) **32**:76–79. doi:10.1016/j.jveb.2019.04.004

4. Aghighi SA, Tipold A, Piechotta M, Lewczuk P, Kästner SBR. Assessment of the effects of adjunctive gabapentin on postoperative pain after intervertebral disc surgery in dogs. *Vet Anaesth Analg* (2012) **39**:636–646. doi:10.1111/j.1467-2995.2012.00769.x

5. Amengual M, Leigh H, Rioja E. Postoperative respiratory effects of intravenous fentanyl compared to intravenous methadone in dogs following spinal surgery. *Vet Anaesth Analg* (2017) **44**:1042–1048. doi:10.1016/j.vaa.2016.11.010

6. Andreoni V, Hughes JML. Propofol and fentanyl infusions in dogs of various breeds undergoing surgery. *Veterinary Anaesthesia and Analgesia* (2009) **36**:523–531. doi:10.1111/j.1467-2995.2009.00490.x

7. Aprea F, Cherubini GB, Palus V, Vettorato E, Corletto F. Effect of extradurally administered morphine on postoperative analgesia in dogs undergoing surgery for thoracolumbar intervertebral disk extrusion. *Scientific Reports* (2012) **241**:6.

8. Barker J, Clark-Price S. Evaluation of Topical Epidural Analgesia Delivered in Gelfoam for Postoperative Hemilaminectomy Pain Control. *Veterinary Surgery* (2013)7.

9. Barnes K, Faludi A, Takawira C, Aulakh K. Extracorporeal Shock Wave Therapy Improves Short Term Limb use after Canine Tibial Plateau Leveling Osteotomy. (2019)10.

10. Bartel AK, Campoy L, Martin-Flores M, Gleed RD, Walker KJ, Scanapico CE, Reichard AB. Comparison of bupivacaine and dexmedetomidine femoral and sciatic nerve blocks with bupivacaine and buprenorphine epidural injection for stifle arthroplasty in dogs. *Veterinary Anaesthesia and Analgesia* (2016) **43**:435–443. doi:10.1111/vaa.12318

11. Bellei E, Roncada P, Pisoni L, Joechler M, Zaghini A. The use of fentanylpatch in dogs undergoing spinal surgery: plasma concentration and analgesic efficacy. (2011)5.

12. Bendinelli C, Properzi R, Boschi P, Bresciani C, Rocca E, Sabbioni A, Leonardi F. Meloxicam vs robenacoxib for postoperative pain management in dogs undergoing combined laparoscopic ovariectomy and laparoscopic-assisted gastropexy. (2018)6.

13. Benitez ME, Roush JK, McMurphy R, KuKanich B, Legallet C. Clinical efficacy of hydrocodone-acetaminophen and tramadol for control of postoperative pain in dogs following tibial plateau leveling osteotomy. *American Journal of Veterinary Research* (2015) **76**:755–762. doi:10.2460/ajvr.76.9.755

14. Benitez ME, Roush JK, KuKanich B, McMurphy R. Pharmacokinetics of hydrocodone and tramadol administered for control of postoperative pain in dogs following tibial plateau leveling osteotomy. (2015) **76**:8.

15. Bienhoff SE, Smith ES, Roycroft LM, Roberts ES, Baker LD. Efficacy and Safety of Deracoxib for the Control of Postoperative Pain and Inflammation Associated with Dental Surgery in Dogs. *ISRN Veterinary Science* (2011)8.

16. Bienhoff SE, Smith ES, Roycroft LM, Roberts ES. Efficacy and Safety of Deracoxib for Control of Postoperative Pain and Inflammation Associated with Soft Tissue Surgery in Dogs. *Veterinary Surgery* (2012)9.

17. Bustamante R, Daza M, Canfran S, Garcia P, Suarez M, Trobo I, Gomez de Segura I. Comparison of the postoperative analgesic effects of cimicoxib, buprenorphine and their combination in healthy dogs undergoing ovariohysterectomy. (2018)12.

18. Caniglia AM, Driessen B, Puerto DA, Bretz B, Boston RC, Larenza MP. Intraoperative antinociception and postoperative analgesia following epidural anesthesia versus femoral and sciatic nerve blockade in dogs undergoing stifle joint surgery. *Journal of the American Veterinary Medical Association* (2012) **241**:1605–1612. doi:10.2460/javma.241.12.1605

19. Cerasoli I, Tutunaru A, Cenani A, Ramirez J, Detilleux J, Belligand M, Sendersen C. Comparison of clinical effects of epidural levobupivacaine morphine versus bupivacaine morphine in dogs undergoing elective pelvic limb surgery. (2017)9.

20. Chiavaccini L, Claude AK, Meyer RE. Comparison of Morphine, Morphine-Lidocaine, and Morphine-Lidocaine-Ketamine Infusions in Dogs Using an Incision-Induced Pain Model. *Journal of the American Animal Hospital Association* (2017) **53**:65–72. doi:10.5326/JAAHA-MS-6442

21. Dancker C, MacFarlane PD, Love EJ. The effect of neuraxial morphine on postoperative pain in dogs after extrahepatic portosystemic shunt attenuation. *Veterinary Anaesthesia and Analgesia* (2020) **47**:111–118. doi:10.1016/j.vaa.2019.06.011

22. Davila D, Keeshen TP, Evans RB, Conzemius MG. Comparison of the analgesic efficacy of perioperative firocoxib and tramadol administration in dogs undergoing tibial plateau leveling osteotomy. *Journal of the American Veterinary Medical Association* (2013) **243**:225–231. doi:10.2460/javma.243.2.225

23. Fitzpatrick CL, Weir HL, Monnet E. Effects of infiltration of the incision site with bupivacaine on postoperative pain and incisional healing in dogs undergoing

ovariohysterectomy. *Journal of the American Veterinary Medical Association* (2010) **237**:395–401. doi:10.2460/javma.237.4.395

24. Fizzano KM, Claude AK, Kuo L-H, Eells JB, Hinz SB, Thames BE, Ross MK, Linford RL, Wills RW, Olivier AK, et al. Evaluation of a modified infraorbital approach for a maxillary nerve block for rhinoscopy with nasal biopsy of dogs. (2017) **78**:11.

25. Fransson BA, Perez TE, Flores K, Gay JM, Acvpm D. Cardiorespiratory Changes and Pain Response of Lift Laparoscopy Compared to Capnoperitoneum Laparoscopy in Dogs. *Veterinary Surgery* (2015)9.

26. Friton G, Thompson C, Karadzovska D, King S, King JN. Efficacy and safety of oral robenacoxib (tablet) for the treatment of pain associated with soft tissue surgery in client-owned dogs. (2017)12.

27. Friton G, Thompson C, Karadzovska D, King S, King JN. Efficacy and Safety of Injectable Robenacoxib for the Treatment of Pain Associated With Soft Tissue Surgery in Dogs. (2017)10.

28. Giudice E, Barillaro G, Crinò C, Alaimo A, Macrì F, Di Pietro S. Postoperative pain in dogs undergoing hemilaminectomy: Comparison of the analgesic activity of buprenorphine and tramadol. *Journal of Veterinary Behavior* (2017) **19**:45–49. doi:10.1016/j.jveb.2017.02.003

29. Goudie-DeAngelis EM, Woodhouse KJ. Evaluation of Analgesic Efficacy and Associated Plasma Concentration of Tramadol and O-desmethyltramadol Following Oral Administration Post Ovariohysterectomy. (2016) **14**:9.

30. Groppetti D, Di Cesare F, Pecile A, Cagnardi P, Merlanti R, D'Urso E, Gioeni D, Boracchi P, Ravasio G. Maternal and neonatal wellbeing during elective C-section induced with a combination of propofol and dexmedetomidine: How effective is the placental barrier in dogs? (2019)9.

31. Gruet P, Seewald W, King JN. Evaluation of subcutaneous and oral administration of robenacoxib and meloxicam for the treatment of acute pain and inflammation associated with orthopedic surgery in dogs. (2011) **72**:10.

32. Gruet P, Seewald W, King JN. Robenacoxib versus meloxicam for the management of pain and inflammation associated with soft tissue surgery in dogs: a randomized, non-inferiority clinical trial. (2013)12.

33. Guerrero KSK, Schwarz A, Wuhrmann R, Feldmann S, Hartnack S, Bettschart-Wolfensberger R. Comparison of a new metamizole formulation and carprofen for extended post-operative analgesia in dogs undergoing ovariohysterectomy. *The Veterinary Journal* (2015)6.

34. Guerrero KSK, Campagna I, Bruhl-Day R, Hegamin-Younger C, Guerrero T. Intraperitoneal bupivacaine with or without incisional bupivacaine for postoperative analgesia in dogs undergoing ovariohysterectomy. (2016)8.

35. Guillot M, Rialland P, Nadeau M-È, del Castillo JRE, Gauvin D, Troncy E. Pain Induced by a Minor Medical Procedure (Bone Marrow Aspiration) in Dogs: Comparison of Pain Scales in a Pilot Study. *Journal of Veterinary Internal Medicine* (2011) **25**:1050–1056. doi:10.1111/j.1939-1676.2011.00786.x

36. Alves IPG, Nicácio GM, Diniz MS, Rocha TLA, Prada G, Cassu RN. Analgesic comparison of systemic lidocaine, morphine or lidocaine plus morphine infusion in dogs undergoing fracture repair. (2014)7.

37. Gurney MA, Rysnik M, Comerford EJ, Cripps PJ, Iff I. Intra-articular morphine, bupivacaine or no treatment for postoperative analgesia following unilateral elbow joint arthroscopy. *Journal of Small Animal Practice* (2012) **53**:6.

38. Gutiérrez-Bautista ÁJ, Morgaz J, Granados M del M, Gómez-Villamandos RJ, Dominguez JM, Fernandez-Sarmiento JA, Aguilar-García D, Navarrete-Calvo R. Evaluation and comparison of postoperative analgesic effects of dexketoprofen and methadone in dogs. *Veterinary Anaesthesia and Analgesia* (2018) **45**:820–830. doi:10.1016/j.vaa.2018.06.016

39. Gutierrez-Blanco E, Victoria-Mora JM, Ibancovichi-Camarillo JA, Sauri-Arceo CH, Bolio-González ME, Acevedo-Arcique CM, Marin-Cano G, Steagall PV. Postoperative analgesic effects of either a constant rate infusion of fentanyl, lidocaine, ketamine, dexmedetomidine, or the combination lidocaine-ketamine-

dexmedetomidine after ovariohysterectomy in dogs. *Veterinary Anaesthesia and Analgesia* (2015) **42**:309–318. doi:10.1111/vaa.12215

40. Hamilton KH, Henderson ER, Toscano M, Chanoit GP. Comparison of postoperative complications in healthy dogs undergoing open and closed orchidectomy. *J Small Anim Pract* (2014) **55**:521–526. doi:10.1111/jsap.12266

41. Heffernan AE, Katz EM, Sun Y, Rendahl AK, Conzemius MG. Once daily oral extended-release hydrocodone as analgesia following tibial plateau leveling osteotomy in dogs. *Veterinary Surgery* (2018) **47**:516–523. doi:10.1111/vsu.12792

42. Hettlich BF, Cook L, London C, Fosgate GT. Comparison of harmonic blade versus traditional approach in canine patients undergoing spinal decompressive surgery for naturally occurring thoracolumbar disk extrusion. *PLoS ONE* (2017) **12**:e0172822. doi:10.1371/journal.pone.0172822

43. Hu XY, Luan L, Guan W, Shi J, Zhao YB, Fan HG. Tolfenamic acid and meloxicam both provide an adequate degree of postoperative analgesia in dogs undergoing ovariohysterectomy. *Veterinarni Medicina* (2017)9.

44. Hunt JR, Attenburrow PM, Slingsby LS, Murrell JC. Comparison of premedication with buprenorphine or methadone with meloxicam for postoperative analgesia in dogs undergoing orthopaedic surgery. *J Small Anim Pract* (2013) **54**:418–424. doi:10.1111/jsap.12103

45. Hunt JR, Grint NJ, Taylor PM, Murrell JC. Sedative and analgesic effects of buprenorphine, combined with either acepromazine or dexmedetomidine, for premedication prior to elective surgery in cats and dogs. *Veterinary Anaesthesia and Analgesia* (2013) **40**:297–307. doi:10.1111/vaa.12003

46. Hunt JR, Slingsby LS, Murrell JC. The effects of an intravenous bolus of dexmedetomidine following extubation in a mixed population of dogs undergoing general anaesthesia and surgery. *The Veterinary Journal* (2014) **200**:133–139. doi:10.1016/j.tvjl.2014.01.015

47. Huuskonen V, Hughes JL, Estaca Bañon E, West E. Intratesticular lidocaine reduces the response to surgical castration in dogs. *Veterinary Anaesthesia and Analgesia* (2013) **40**:74–82. doi:10.1111/j.1467-2995.2012.00775.x

48. Kaka U, Rahman N-A, Abubakar AA, Goh YM, Fakurazi S, Omar MA, Chen HC. Pre-emptive multimodal analgesia with tramadol and ketamine-lidocaine infusion for suppression of central sensitization in a dog model of ovariohysterectomy. *JPR* (2018) **Volume 11**:743–752. doi:10.2147/JPR.S152475

49. Karrasch NM, Lerche P, Aarnes TK, Gardner HL, London CA. The effects of preoperative oral administration of carprofen or tramadol on postoperative analgesia in dogs undergoing cutaneous tumor removal. (2015) **56**:6.

50. Kibar M, Tuna B, Kisadere I, Güzelbektes H. Comparison of Instilled Lidocaine and Procaine Effects on Pain Relief in Dogs Undergoing Elective Ovariohysterectomy. *Israel Journal of Veterinary Medicine* (2019) **74**: Available at: https://www.ivis.org/library/israel-journal-of-veterinary-medicine/israel-journal-of-veterinary-medicine-vol-743-sep-7

51. Kim JH, Seok SH, Park TY, Kim HJ, Lee SW, Lee HC, Yeon SC. Analgesic effect of intra-articular ropivacaine injection after arthroscopic surgery on the shoulder joint in dogs. *Veterinarni Medicina* (2018) **63**:513–521. doi:10.17221/37/2017-VETMED

52. Kim YK, Lee S, Suh E, Lee L, Lee H, Lee H, Yeon SC. Sprayed intraperitoneal bupivacaine reduces early postoperative pain behavior and biochemical stress response after laparoscopic ovariohysterectomy in dogs. *The Veterinary Journal* (2012)5.

53. Kondo Y, Takashima K, Matsumoto S, Shiba M, Otsuki T, Kinoshita G, Rosentel J, Gross SJ, Fleishman C, Yamane Y. Efficacy and Safety of Firocoxib for the Treatment of Pain Associated with Soft Tissue Surgery in Dogs under Field Conditions in Japan. (2012)7.

54. Kongara K, Chambers J, Johnson C. Effects of tramadol, morphine or their combination in dogs undergoing ovariohysterectomy on peri-operative electroencephalographic responses and post-operative pain. *New Zealand Veterinary Journal* (2012) **60**:129–135. doi:10.1080/00480169.2011.641156

55. Kongara K, Chambers J, Johnson C, Dukkipati V. Effects of tramadol or morphine in dogs undergoing castration on intra-operative electroencephalogram

responses and post-operative pain. *New Zealand Veterinary Journal* (2013) **61**:349–353. doi:10.1080/00480169.2013.780280

56. Kropf J, Hughes JML. Effects of midazolam on cardiovascular responses and isoflurane requirement during elective ovariohysterectomy in dogs. *Ir Vet J* (2018) **71**:26. doi:10.1186/s13620-018-0136-y

57. Kropf J, Hughes JML. Effect of midazolam on the quality and duration of anaesthetic recovery in healthy dogs undergoing elective ovariohysterectomy or castration. (2019)10.

58. Kushnir Y, Toledano N, Cohen L, Bdolah-Abram T, Shilo-Benjamini Y. Intratesticular and incisional line infiltration with ropivacaine for castration in medetomidine-butorphanol-midazolam sedated dogs. (2017)10.

59. Lambertini C, Kluge K, Lanza-Perea M, Bruhl-Day R, Guerrero KSK. Comparison of intraperitoneal ropivacaine and bupivacaine for postoperative analgesia in dogs undergoing ovariohysterectomy. (2018)6.

60. Lardone E, Peirone B, Adami C. Combination of magnesium sulphate and ropivacaine epidural analgesia for hip arthroplasty in dogs. (2017)9.

61. Lascelles BDX, Rausch-Derra LC, Wofford JA, Huebner M. Pilot, randomized, placebo-controlled clinical field study to evaluate the effectiveness of bupivacaine liposome injectable suspension for the provision of post-surgical analgesia in dogs undergoing stifle surgery. *BMC Vet Res* (2016) **12**:168. doi:10.1186/s12917-016-0798-1

62. Lewis KA, Bednarski RM, Aarnes TK, Dyce J, Hubbell JAE. Postoperative comparison of four perioperative analgesia protocols in dogs undergoing stifle joint surgery. *Journal of the American Veterinary Medical Association* (2014) **244**:1041–1046. doi:10.2460/javma.244.9.1041

63. Li L, Dong J, Fen X, Li B, Chen Y, Sha J, Fan H. Effects of dexmedetomidine on plasma glucose, cortisol and adrenocorticotropic hormone concentrations of canine undergoing ovariohysterectomy. *Thai J Vet Med* (2017)6.

64. Linton DD, Wilson MG, Newbound GC, Freise KJ, Clark TP. The effectiveness of a long-acting transdermal fentanyl solution compared to buprenorphine for the control of postoperative pain in dogs in a randomized, multicentered clinical study. *Journal of Veterinary Pharmacology and Therapeutics* (2012) **35**:53–64. doi:10.1111/j.1365-2885.2012.01408.x

65. Little D, Johnson S, Hash J, Olson SA, Estes BT, Moutos FT, Lascelles BDX, Guilak F. Functional outcome measures in a surgical model of hip osteoarthritis in dogs. *J EXP ORTOP* (2016) **3**:17. doi:10.1186/s40634-016-0053-5

66. Luna SPL, Martino I, Lorena S, Capua M, Lima A, Santos B, Brondani J, Vesce G. Acupuncture and pharmacopuncture are as effective as morphine or carprofen for postoperative analgesia in bitches undergoing ovariohysterectomy. (2015)7.

67. Lush J, Ijichi C. A preliminary investigation into personality and pain in dogs. (2018)7.

68. Martinez SA, Wilson MG, Linton DD, Newbound GC, Freise KJ, Lin T -L., Clark TP. The safety and effectiveness of a long-acting transdermal fentanyl solution compared with oxymorphone for the control of postoperative pain in dogs: a randomized, multicentered clinical study. *J vet Pharmacol Therap* (2014) **37**:394–405. doi:10.1111/jvp.12096

69. Martinez-Taboada F, Redondo JI. Comparison of the hanging-drop technique and running-drip method for identifying the epidural space in dogs. *Veterinary Anaesthesia and Analgesia* (2017) **44**:329–336. doi:10.1016/j.vaa.2016.03.002

70. McCally RE, Bukoski A, Branson KR, Fox DB, Cook JL. Comparison of Short-Term Postoperative Analgesia by Epidural, Femoral Nerve Block, or Combination Femoral and Sciatic Nerve Block in Dogs Undergoing Tibial Plateau Leveling Osteotomy: Comparison of Epidural or Peripheral Nerve Blocks After TPLO. *Veterinary Surgery* (2015) **44**:983–987. doi:10.1111/vsu.12406

71. McKune CM, Pascoe PJ, Lascelles BDX, Kass PH. The challenge of evaluating pain and a pre-incisional local anesthetic block. *PeerJ* (2014) **2**:e341. doi:10.7717/peerj.341

72. McMillan MW, Seymour CJ, Brearley JC. Effect of intratesticular lidocaine on isoflurane requirements in dogs undergoing routine castration. *Journal of Small Animal Practice* (2012) **53**:5.

73. Meakin LB, Murrell JC, Doran ICP, Knowles TG, Tivers MS, Chanoit GPA. Electrosurgery reduces blood loss and immediate postoperative inflammation compared to cold instruments for midline celiotomy in dogs: A randomized controlled trial. (2016)5.

74. Merema DK, Schoenrock EK, Boedec KL, McMichael MA. Effects of a transdermal lidocaine patch on indicators of postoperative pain in dogs undergoing midline ovariohysterectomy. (2017) **250**:8.

75. Meunier NV, Panti A, Mazeri S, Fernandes KA, Handel IG. Randomised trial of perioperative tramadol for canine sterilisation pain management. (2019)8.

76. Morgaz J, Navarrete R, Muñoz-Rascón P, Domínguez JM, Fernández-Sarmiento JA, Gómez-Villamandos RJ, Granados MM. Postoperative analgesic effects of dexketoprofen, buprenorphine and tramadol in dogs undergoing ovariohysterectomy. *Research in Veterinary Science* (2013) **95**:278–282. doi:10.1016/j.rvsc.2013.03.003

77. Morgaz J, Muñoz-Rascón P, Serrano-Rodriguez J, Navarrete R, Dominguez JM, Fernandez-Sarmiento JA, Gomez-Villamandos R, Serrano J, Granados MM. Effectiveness of pre-peritoneal continuous wound infusion with lidocaine for pain control following ovariohysterectomy in dogs. *The Veterinary Journal* (2014)5.

78. Nour E, Othman M, Karrouf G, Zaghloul A. Glasgow Composite Measure Pain Scale score and comaprison between several adjuvants in association with bupivacaine. *Life Science Journal* (2013) **10**: Available at: https://www.researchgate.net/publication/256399795_Glasgow_Composite_Measure_Pain_Scale_score_and_comaprison_between_several_adjuvants_in_association_with_bupivacaine [Accessed January 23, 2021]

79. Palomba N, Vettorato E, De Gennaro C, Corletto F. Peripheral nerve block versus systemic analgesia in dogs undergoing tibial plateau levelling osteotomy: Analgesic efficacy and pharmacoeconomics comparison. (2019)10.

80. Pascal M, Burac M, Diaconescu A, Togoe D, Vitalaru A, Bîrtoiu A. Comparison of tramadol and robenacoxib postoperative analgesic efficacy in dogs. *Scientific Works Series C Veterinary Medicine* (2013) **LIX**:72–75.

81. Pascal M, Allison A, Kaartinen J. Opioid-sparing effect of a medetomidine constant rate infusion during thoraco-lumbar hemilaminectomy in dogs administered a ketamine infusion. *Veterinary Anaesthesia and Analgesia* (2020) **47**:61–69. doi:10.1016/j.vaa.2019.06.012

82. Peeters ME, Kirpensteijn J. Comparison of surgical variables and short-term postoperative complications in healthy dogs undergoing ovariohysterectomy or ovariectomy. *Journal of the American Veterinary Medical Association* (2011) **238**:189–194. doi:10.2460/javma.238.2.189

83. Perez TE, Grubb TL, Greene SA, Meyer S, Valdez N, Bingman J, Farnsworth R. Effects of intratesticular injection of bupivacaine and epidural administration of morphine in dogs undergoing castration. *Scientific Reports* (2013) **242**:12.

84. Perry K, Rutherford L, Sajik D, Bruce M. A preliminary study of the effect of closed incision management with negative pressure wound therapy over high-risk incisions. (2015)12.

85. Portela D, Otero P, Briganti A, Romano M, Corletto F, Breghi G. Femoral nerve block: a novel psoas compartment lateral pre-iliac approach in dogs. (2012)11.

86. Re Bravo V, Aprea F, Bhalla RJ, De Gennaro C, Cherubini GB, Corletto F, Vettorato E. Effect of 5% transdermal lidocaine patches on postoperative analgesia in dogs undergoing hemilaminectomy. *J Small Anim Pract* (2019) **60**:161–166. doi:10.1111/jsap.12925

87. Read K, Khatun M, Murphy H. Comparison of transdermal fentanyl and oral tramadol for lateral thoracotomy in dogs: cardiovascular and behavioural data. *Veterinary Anaesthesia and Analgesia* (2019) **46**:116–125. doi:10.1016/j.vaa.2018.09.046

88. Reece JF, Nimesh MK, Wyllie RE, Jones AK, Dennison AW. Description and evaluation of a right flank, mini-laparotomy approach to canine

ovariohysterectomy. *Veterinary Record* (2012) **171**:248–248. doi:10.1136/vr.100907

89. Rioja E, Dzikiti B, Fosgate G, Goddaard A, Stegmann F, Schoeman J. Effects of a constant rate infusion of magnesium sulphate in healthy dogs anaesthetized with isoflurane and undergoing ovariohysterectomy. (2012)12.

90. Romano M, Portela DA, Breghi G, Otero PE. Stress-related biomarkers in dogs administered regional anaesthesia or fentanyl for analgesia during stifle surgery. *Veterinary Anaesthesia and Analgesia* (2016) **43**:44–54. doi:10.1111/vaa.12275

91. Sarotti D, Rabozzi R, Franci P. Comparison of epidural versus intrathecal anaesthesia in dogs undergoing pelvic limb orthopaedic surgery. *Veterinary Anaesthesia and Analgesia* (2015) **42**:405–413. doi:10.1111/vaa.12229

92. Sarotti D, Rabozzi R, Franci P. Effects of intravenous dexmedetomidine infusion on local anaesthetic block: A spinal anaesthesia clinical model in dogs undergoing hind limb surgery. *Research in Veterinary Science* (2019) **124**:93–98. doi:10.1016/j.rvsc.2019.03.001

93. Scott JE, Singh A, Valverde A, Blois SL, Foster RA, Kilkenny JJ, Linden A zur. Effect of pneumoperitoneum with warmed humidified or standard-temperature carbon dioxide during laparoscopy on core body temperature, cardiorespiratory and thromboelastography variables, systemic inflammation, peritoneal response, and signs of postoperative pain in healthy mature dogs. *American Journal of Veterinary Research* (2018) **79**:1321–1334. doi:10.2460/ajvr.79.12.1321

94. Shah MD, Yates D, Hunt J, Murrell JC. A comparison between methadone and buprenorphine for perioperative analgesia in dogs undergoing ovariohysterectomy: A comparison of methadone and buprenorphine. *J Small Anim Pract* (2018) **59**:539–546. doi:10.1111/jsap.12859

95. Shih AC, Robertson S, Isaza N, Pablo L, Davies W. Comparison between analgesic effects of buprenorphine, carprofen, and buprenorphine with carprofen

for canine ovariohysterectomy. *Veterinary Anaesthesia and Analgesia* (2008) **35**:69–79. doi:10.1111/j.1467-2995.2007.00352.x

96. Shilo-Benjamini Y, Slav SA, Kahane N, Kushnir Y, Sarfaty H, Ofri R. Analgesic effects of intraorbital insertion of an absorbable gelatin hemostatic sponge soaked with 1% ropivacaine solution following enucleation in dogs. *Journal of the American Veterinary Medical Association* (2019) **255**:1255–1262. doi:10.2460/javma.255.11.1255

97. Shivley J, Richardson JM, Woodruff KA, Brookshire W, Meyer R, Smith D. Sharp Transection of the Suspensory Ligament as an Alternative to Digital Strumming during Canine Ovariohysterectomy. (2018)6.

98. Skelding A, Valverde A, Aguilera R, Moens NM, Sinclair M, Thomason JJ. Comparison of 3 blind brachial plexus block techniques during maintenance of anesthesia and postoperative pain scores in dogs undergoing surgical procedures of the thoracic limb. (2019)9.

99. Srithunyarat T, Höglund OV, Hagman R, Olsson U, Stridsberg M, Lagerstedt A-S, Pettersson A. Catestatin, vasostatin, cortisol, temperature, heart rate, respiratory rate, scores of the short form of the Glasgow composite measure pain scale and visual analog scale for stress and pain behavior in dogs before and after ovariohysterectomy. *BMC Res Notes* (2016) **9**:381. doi:10.1186/s13104-016-2193-1

100. Swallow A, Rioja E, Elmer T, Dugdale A. The effect of maropitant on intraoperative isoflurane requirements and postoperative nausea and vomiting in dogs: a randomized clinical trial. *Veterinary Anaesthesia and Analgesia* (2017) **44**:785–793. doi:10.1016/j.vaa.2016.10.006

101. Tallant A, Ambros B, Freire C, Sakals S. Comparison of intraoperative and postoperative pain during canine ovariohysterectomy and ovariectomy. (2016) **57**:6.

102. Tayari H, Tazioli G, Breghi G, Briganti A. Ultrasound-guided femoral and obturator nerves block in the psoas compartment in dogs: anatomical and randomized clinical study. (2017)11.

103. Tayari H, Otero P, Rossetti A, Breghi G, Briganti A. Proximal RUMM block in dogs: preliminary results of cadaveric and clinical studies. *Veterinary Anaesthesia and Analgesia* (2019) **46**:384–394. doi:10.1016/j.vaa.2018.11.009

104. Travis BM, Hayes GM, Vissio K, Harvey HJ, Flanders JA, Sumner JP. A quilting subcutaneous suture pattern to reduce seroma formation and pain 24 hours after midline celiotomy in dogs: A randomized controlled trial. *Veterinary Surgery* (2018) **47**:204–211. doi:10.1111/vsu.12754

105. Valtolina C, Robben JH, Uilenreef J, Murrell JC, Aspegrén J, McKusick BC, Hellebrekers LJ. Clinical evaluation of the efficacy and safety of a constant rate infusion of dexmedetomidine for postoperative pain management in dogs. *Veterinary Anaesthesia and Analgesia* (2009) **36**:369–383. doi:10.1111/j.1467-2995.2009.00461.x

106. Vettorato E, Zonca A, Isola M, Villa R, Gallo M, Ravasio G, Beccaglia M, Montesissa C, Cagnardi P. Pharmacokinetics and efficacy of intravenous and extradural tramadol in dogs. *The Veterinary Journal* (2010) **183**:310–315. doi:10.1016/j.tvjl.2008.11.002

107. Wagner AE, Worland GA, Glawe JC, Hellyer PW. Multicenter, randomized controlled trial of pain-related behaviors following routine neutering in dogs. *Journal of the American Veterinary Medical Association* (2008) **233**:109–115. doi:10.2460/javma.233.1.109

108. Wagner AE, Mich PM, Uhrig SR, Hellyer PW. Clinical evaluation of perioperative administration of gabapentin as an adjunct for postoperative analgesia in dogs undergoing amputation of a forelimb. *Journal of the American Veterinary Medical Association* (2010) **236**:751–756. doi:10.2460/javma.236.7.751

109. Wang-Leandro A, Willmitzer F, Karol A, Porcellini B, Kronen P, Hiltbrand EM, Rüfenacht D, Kircher PR, Richter H. MRI-guided, transrectal, intraprostatic steam application as potential focal therapeutic modality for prostatic diseases in a large animal translational model: A feasibility follow-up study. *PLoS ONE* (2019) **14**:e0226764. doi:10.1371/journal.pone.0226764

110. Watanabe R, Monteiro BP, Evangelista MC, Castonguay A, Edge D, Steagall PV. The analgesic effects of buprenorphine (Vetergesic or Simbadol) in combination with carprofen in dogs undergoing ovariohysterectomy: a randomized, blinded, clinical trial. *BMC Vet Res* (2018) **14**:304. doi:10.1186/s12917-018-1628-4

111. Weil C, Tünsmeyer J, Tipold A, Hoppe S, Beyerbach M, Pankow W-R, Kästner SB. Effects of concurrent perioperative use of marbofloxacin and cimicoxib or carprofen in dogs: Marbofloxacin and cimicoxib or carprofen in dogs. *J Small Anim Pract* (2016) **57**:311–317. doi:10.1111/jsap.12464

112. Zhang S, Li J, Luan L, Guan W, Hu X, Fan H. Comparison of the effects of nefopam and tramadol on postoperative analgesia in dogs undergoing ovariohysterectomy. *Veterinarni Medicina* (2017) **62**:131–137. doi:10.17221/53/2016-VETMED

113. Zidan N, Fenn J, Griffith E, Early PJ, Mariani CL, Muñana KR, Guevar J, Olby NJ. The Effect of Electromagnetic Fields on Post-Operative Pain and Locomotor Recovery in Dogs with Acute, Severe Thoracolumbar Intervertebral Disc Extrusion: A Randomized Placebo-Controlled, Prospective Clinical Trial. *Journal of Neurotrauma* (2018) **35**:1726–1736. doi:10.1089/neu.2017.5485

114. Zidan N, Sims C, Fenn J, Williams K, Griffith E, Early PJ, Mariani CL, Munana KR, Guevar J, Olby NJ. A randomized, blinded, prospective clinical trial of postoperative rehabilitation in dogs after surgical decompression of acute thoracolumbar intervertebral disc herniation. *J Vet Intern Med* (2018) **32**:1133–1144. doi:10.1111/jvim.15086

Appendix 4

Univariable tests investigating the association between study design factors and the finding of a statistically significant difference

Type of control

	STATISTICAL DIFFERENCE FOUND		
CONTROL (POSITIVE / NEGATIVE / PSEUDO-NEGATIVE)	No	Yes	Total
Negative	2	11	13
Positive	23	17	40
Pseudo-negative	20	10	30
Total	45	38	83

χ^2 Tests

	Value	df	p
χ^2	9.94	2	0.007
N	83		

GCMPS-SF primary/secondary outcome

	STATISTICAL DIFFERENCE FOUND		
PAIN SCORES (BOTH GUIDING RESCUE OR ABSOLUTE SCORES): PRIMARY/SECONDARY OUTCOME?	No	Yes	Total
Primary	28	34	62
Secondary	17	4	21
Total	45	38	83

χ^2 Tests

	Value	df	p
χ^2	8.10	1	0.004
N	83		

Modifications of the scale

MODIFICATIONS OF THE SCALE	STATISTICAL DIFFERENCE FOUND		
	No	Yes	Total
No	45	32	77
Yes	0	6	6
Total	45	38	83

χ^2 Tests

	Value	df	p
χ^2	7.66	1	0.006
N	83		

Presentation of a power calculation

POWER CALCULATION Y/N	STATISTICAL DIFFERENCE FOUND		
	No	Yes	Total
Yes	24	13	37
No	21	25	46
Total	45	38	83

χ^2 Tests

	Value	df	p
χ^2	3.05	1	0.081
N	83		

CONSORT score

		Statistic	p
CONSORT SCORE	Mann-Whitney U	636	0.031

Group Descriptives

	Group	N	Mean	Median	SD	SE
CONSORT score	No	45	3.71	4.00	3.57	0.531
	Yes	38	2.11	0.00	3.35	0.544

Modifications of the intervention level

	STATISTICAL DIFFERENCE FOUND		
INTERVENTION LEVEL	**No**	**Yes**	**Total**
Altered	17	10	27
No	28	28	56
Total	45	38	83

χ^2 Tests

	Value	df	p
χ^2	1.23	1	0.267
N	83		

Group size

		Statistic	p
GROUP SIZE: TOTAL NUMBER OF DOGS	Mann-Whitney U	743	0.307

Group Descriptives

	Group	N	Mean	Median	SD	SE
Group size: total number of dogs	No	45	35.0	30.0	21.1	3.14
	Yes	38	69.3	32.0	103	16.7

Blinding

	STATISTICAL DIFFERENCE FOUND		
STUDY DESIGN: BLINDED?	**No**	**Yes**	**Total**
No	8	4	12
Yes	37	34	71
Total	45	38	83

χ^2 Tests

	Value	df	p
χ^2	0.876	1	0.349
N	83		

Clinical *versus* experimental study design

CLINICAL / EXPERIMENTAL	STATISTICAL DIFFERENCE FOUND		Total
	No	Yes	
Clinical	41	34	75
Experimental	4	4	8
Total	45	38	83

χ^2 Tests

	Value	df	p
χ^2	0.0634	1	0.801
N	83		

Was the study specifically powered for the GCMPS-SF?

POWERED FOR GCMPS-SF?	STATISTICAL DIFFERENCE FOUND		Total
	No	Yes	
No	35	33	68
Yes	10	5	15
Total	45	38	83

χ^2 Tests

	Value	df	p
χ^2	1.14	1	0.285
N	83		

Single *versus* multicentre study design

SINGLE / MULTICENTRE	STATISTICAL DIFFERENCE FOUND		Total
	No	Yes	
Multicentre	2	6	8
Single centre	43	32	75
Total	45	38	83

χ^2 Tests

	Value	df	p
χ^2	3.04	1	0.081
N	83		

Type of surgery

TYPE OF SURGERY	STATISTICAL DIFFERENCE FOUND		Total
	No	Yes	
Soft tissue single	21	15	36
Neuro	6	4	10
Mixed	4	3	7
Ortho single	8	8	16
Ortho mixed	3	1	4
Soft tissue mixed	3	7	10
Total	45	38	83

χ^2 Tests

	Value	df	p
χ^2	3.58	5	0.612
N	83		

Type of observer

	STATISTICAL DIFFERENCE FOUND		
OBSERVER TYPE	No	Yes	Total
Vet	12	11	23
Student	3	0	3
Nurse	5	3	8
Not specified	25	24	49
Total	45	38	83

χ^2 Tests

	Value	df	p
χ^2	2.99	3	0.392
N	83		

Statistical techniques used

	STATISTICAL DIFFERENCE FOUND		
TYPE OF STATISTICAL TEST	No	Yes	Total
Parametric	17	19	36
Other	28	19	47
Total	45	38	83

χ^2 Tests

	Value	df	p
χ^2	1.25	1	0.263
N	83		

Appendix 5

Summary of all the procedures used in the 30 clinical trials that reported the largest difference in GCMPS-sf scores detected.

TYPE OF SURGERY	NUMBER OF STUDIES THAT EMPLOYED IT
Ovariohysterectomy	8
Hemilaminectomy	1
Laparoscopic ovariohysterectomy	1
Castration	2
TPLO	4
Lateral retinacular suture procedure, including stifle arthrotomy	1
Orthopaedic surgery	0
Soft tissue surgery	4
Orthopaedic surgery of the pelvic limb	0
Hemilaminectomy (acute vertebral disc extrusion)	1
Arthroscopic surgery (shoulder)	1
Multiple surgeries	0
Castration - OHE	1
BMA	1
Unilateral elbow arthroscopy	1
Not specified	1
Midline celiotomy	0
Combined laparoscopic ovariectomy and laparoscopic-assisted gastropexy	1
Lateral thoracotomy	1
Enucleation	1
EHPSS attenuation	0

REFERENCES

Abdallah CG, Geha P (2017) Chronic pain and chronic stress: two sides of the same coin? *Chronic stress* 1: 1-10

Ahmad AH, Zakaria R (2015) Pain in times of stress. *Malays J Med Sci* Special issue: 52-61

Alford P, Geller S, Richardson B, Slater M, Honnas C, Foreman J, Robinson J, Messer M et al (2001) A multicenter, matched case-control study of risk factors for equine laminitis. *Prev Vet Med* 49: 209-222

Arksey H, O'Malley L (2005) Scoping studies: towards a methodological framework. *Int J Social Research Methodology* 8 (1): 19-32

Bartlett PC, Van Buren JW, Bartlett AD, Zhou C (2009) Case-control study of risk factors associated with feline and canine chronic kidney disease. *Vet Med Int* ID 957570

Bafeta A, Dechartres A, Trinquart L, Yavchitz A, Boutron I, Ravaud P (2012) Impact of single centre status on estimates of intervention effects in trials with continuous outcomes: meta-epidemiological study. *BMJ* 344: e813

Bahreini M, Safaie A, Mirfazaelian H, Jalili M (2020) How much change in pain score does rally matter to patients? *Am J Em Med* 38: 1641-1646

Barletta M, Young CN, Quandt JE, Hofmeister EH (2016) Agreement between veterinary students and anesthesiologists regarding postoperative pain assessment in dogs. *VAA* 43: 91-98

Barr RG (1998) Reflections on measuring pain in infants: dissociation in responsive systems and "honest signaling. *Arch Dis Child* 79:152-156

Bell A (2018) The neurobiology of acute pain. *The Vet Journal* 237: 55-62

Bertout JA, Baneux JR, Robertson-Plouch CK (2021) Recommendations for ethical review of veterinary clinical trials. *Front Vet Sci* 8: 715926

Bewick V, Cheek L, Ball J (2004) Statistics review 12: Survival analysis. *Critical Care* 8: 389-394

Bhaskaran K, Smeeth L (2014) What is the difference between missing completely at random and missing at random? *Int J Epidemiol* 43 (4): 1336-1339

Bilgic Temel A, Murrell DF (2019) Pharmacological advances in pemphigus. *Curr Opin Pharmacol*. 46: 44–49

Brondani JT, Luna SPL, Padovani CR (2011) Refinement and initial validation of a multidimensional composite scale for use in assessing acute postoperative pain in cats. *Am J of Vet Res* 72: 174–183

Brondani JT, Mama KR, Luna SP, Wright BD, Niyom S, Ambrosio J, Vogel PR, Padovani CR (2013) Validation of the English version of the UNESP-Botucatu multidimensional composite pain scale for assessing postoperative pain in cats. *BMC Vet Res* 9: 143

Brotman DJ, Golden SH, Wittstein IS (2007) The cardiovascular toll of stress. *The Lancet* 22: 370 (9592)

Buisman M, Wagner MC, Hasiuk MMM, Prebble M, Law L, Pang DSJ (2016) Effects of ketamine and alfaxalone on application of a feline pain assessment scale. *J Fel Med Surg* 18 (8): 643-651

Buisman M, Hasiuk MMM, Gunn M, Pang DSJ (2017) The influence of demeanor on scores from two validated feline pain assessment scales during the perioperative period. *VAA* 44 (3): 646-655

Bussières G, Jacques C, Lainay O, Beauchamp G, Leblond A, Cadoré JL, Desmaizières LM et al (2008) Development of a composite orthopaedic pain scale in horses. *Res Vet Sci* 85: 294-306

Burton AR, Fazalbhoy A, Macefield VG (2016) Sympathetic responses to noxious stimulation of muscle and skin. *Front Neurology* 7: 109

Calvo G, Holden E, Reid J, Scott EM, Firth A, Bell A, Robertson S, Nolan AM (2014) Development of a behaviour-based measurement tool with defined intervention level for assessing acute pain in cats. *J Small Anim Pract* 55: 622-629

Carsten RE, Hellyer PW, Bachand AM, LaRue SM (2008) Correlations between acute radiation scores and pain scores in canine radiation patients with cancer of the forelimb. *VAA* 35: 355-362

Chapman RC (1976) Measurement of pain: problems and issues. *Advances in Pain Research and Therapy* 1: 345-353

Clark TG, Bradburn MJ, Love SB, Altman DG (2003) Survival analysis Part I: basic concepts and first analyses. *Br J Cancer* 89 (2): 232-238

Colquhoun HL, Levac D, O'Brien KK, Straus S, Tricco AC, Perrier L, Kastner M, Moher D (2014) Scoping reviews: time for clarity in definition, methods and reporting. *J Clin Epidemiol* 67 (12): 1291-1294

Conzemius MG, Hill CM, Sammarco JL, Perkowski SZ (1997) Correlation between subjective and objective measures used to determine severity of postoperative pain in dogs. *JAVMA* 210: 1619-1622

Dalla Costa E, Minero M, Lebelt D, Stucke D, Canali E, Leach MC (2014) Development of the Horse Grimace Scale (HGS) as a pain assessment tool in horses undergoing routine castration. *PLoS ONE* 9 (3): e92281

Day JD, Altman DG (2000) Blinding in clinical trials and other studies. *BMJ* 321: 504

Davis RB, Mukamal KJ (2006) Hypothesis testing means. *Circulation* 114: 1078-1082

Della Rocca G, Colpo R, Reid J, Di Salvo A, Scott M (2018) Creation and validation of the Italian version of the Glasgow Composite Measure Pain Scale - Short Form (ICMPS-SF). *Veterinaria Italiana* 54 (3): 251-260

Della Rocca G, Di Salvo A, Marenzoni ML, Bellezza E, Pastorino G, Monteiro B, et al (2019) Development, preliminary validation, and refinement of the Composite Oral and Maxillofacial Pain Scale-Canine/Feline (COPS-C/F) *Front Vet Sci* 6: 274

Des Roches AB, Faure M, Lussert A, Harry V, Rainard P, Durand D, Foucras G (2017) Behavioral and patho-physiological response as possible signs of pain in dairy cows during *Escherichia Coli* mastitis: a pilot study. *J Dairy Sci* 100 (10): 8385-8397

Di Giminiani P, Brierley VLMH, Scollo A, Gottardo F, Malcom EM, Edwards SA, Leach MC (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

Di Girolamo N, Giuffrida MA, Winter AL, Meursinge Reynders R (2017) Reporting and communication of randomisation procedures is suboptimal in veterinary trials. *Vet Rec* 181 (8): 195-195

Downie WW, Leatham PA, Rhind VM, Wright V, Branco JA, Anderson JA (1978) Studies with pain rating scales. *Annals of rheumatic diseases* 37: 378-381

Eckersall PD, Young FJ, McComb C, Hogarth CJ, Safi S, Weber A et al (2001) Acute phase proteins in serum and milk from dairy cows with clinical mastitis. *Vet Rec* 148 (2): 35-41

Eckersall PD, Bell R (2010) Acute phase proteins: biomarkers of infection and inflammation in veterinary medicine. *The Vet J* 185 (1): 23-27

Ekman P, Friesen W (1978) Facial Action Coding System. *Consulting Psychologists Press*

Elwood B, Murison PJ (2022) Investigating the effect of anxiety on pain scores in dogs. *VAA* 49: 135-142

English T, Keeley JW (2015) Internal consistency approach to test construction. *The Encyclopedia of Clinical Psychology*, First Edition, John Wiley & Sons, Inc. (eds R.L. Cautin and S.O. Lilienfeld)

Epstein M, Rodan I, Griffenhagen G, Kadrlik J, Petty M, Robertson S, et al (2015) AAHA/AAFP pain management guidelines for dogs and cats. *J Am Anim Hosp Assoc* 51: 67–84

Evangelista MC, Watanabe R, Leung VSY, Monteiro BP, O'Toole E, Pang DSJ, Stegall PV (2019) Facial expressions of pain in cats: the development and validation of a Feline Grimace Scale. *Sci Rep* 9: 19128

Farrar JT, Portenoy RK, Berlin JA, Kinman JL, Strom BL (2000) Defining the clinically important difference in pain outcome measures. *Pain* 88: 287–294

Firth AM, Haldane SL (1999) Development of a scale to evaluate postoperative pain in dogs. *J Am Vet Med Assoc* 214 (5): 651-659

Fiske DW, Fiske ST (2005) Laboratory studies. *Encyclopedia of social measurement*, Elsevier 435-439

Fitzcharles MA, Cohen SP, Clauw DJ, Littlejohn G, Usui C, Häuser W (2021) Nociplastic pain: towards an understanding of prevalent pain conditions. *The Lancet* 397 (10289): 2098-2110

Flecknell PA (2010) Do mice have a pain face? *Nature Methods* 7 (6): 437-439

Fox SM, Mellor DJ, Firth EC, Hodge H, Lawoko CRO (1994) Changes in plasma cortisol concentrations before, during and after analgesia, an- aesthesia and anaesthesia plus ovariohysterectomy in bitches. *Res Vet Sci* 57:110-118.

Fox SM, Mellor DJ, Stafford KJ, Lowoko CR, Hodge H (2000) The effects of ovariohysterectomy plus different combinations of halothane anaesthesia and butorphanol analgesia on behavior in the bitch. *Res Vet Sci* 68 (3): 265–274

Frayers PM, Hand DJ (2002) Casual variables, indicator variables and measurement scales: an example from quality of life. *J R Statist Soc* 165 (2): 233-261

Furr M (2012) Multicentre clinical research and the veterinary clinician. *Eq Vet J* 44 (suppl 41): 3-4

Gewandter JS, Smith SM, McKeown A, Burke LB, Hertz SH, Hunsinger M, et al (2014) Reporting of primary analyses and multiplicity adjustment in recent analgesic clinical trials: ACTTION systematic review and recommendations. *Pain* 155: 461–466

Gewandter JS, Eisenach JC, Gross RA, Jensen MP, Keefe FJ, Lee DA, et al (2019) Checklist for the preparation and review of pain clinical trial publications: a pain-specific supplement to CONSORT. *PAIN Rep* 4: e621

Gilron I, Carr DB, Desjardins PJ, Kehlet H (2019) Current methods and challenges for acute pain clinical trials. *Pain* 4: e647

Giuffrida MA (2014) Type II error and statistical power in reports of small animal clinical trials. *JAVMA* 244 (9): 1075-1080

Gleerup KB, Forkman B, lindegaard C, Andersen PH (2015) An equine pain face. *VAA* 42: 103-114

Gleerup KB, Andersen PH, Munksgaard L, Forkman B (2015) Pain evaluation in dairy cattle. *Applied Animal Behaviour Science* 171: 25-32

Goldman L, Schafer AI (2020) Goldman-Cecil Medicine 26th Edition. Chapter 27

Golubczyk D, Malysz-Cymborska I, Kalkowski L, Janowski M, Coates JR, Wojtkiewicz J, et al. (2019) The role of glia in canine degenerative myelopathy: relevance to human amyotrophic lateral sclerosis. *Mol Neurobiol* 56: 5740–5748

Gough D, Davies P, Jamtvedt G, Langlois E, Little J, Lotfi T, Masset E, Merlin T et al. (2020) Evidence Synthesis International (ESI): Position statement. *BMC* 9: 155

Green SB, Yang Y, Alt M, Brinkley S, Gray S, Hogan T, Cowan N (2016) Use of internal consistency for estimating reliability of experimental tasks scores. *Psychon Bull Rev* 23 (3): 750-763

Grichnik KP, Ferrante FM. (1991) The difference between acute and chronic pain. *Mt Sinai J Med*. 58 (3):217-20

Gruen ME, White P, Hare B (2020) Do dog breeds differ in pain sensitivity? Veterinarians and the public believe they do. *PLoS ONE* 15(3): e0230315

Guesgen MJ, Beausoleil NJ, Leach M, Minot EO, Stewart M, Stafford KJ (2016) Coding and quantification of a facial expression for pain in lambs. *Behavioural Processes* 132: 49-56

Guillot M, Rialland P, Nadeau ME et al. (2011) Pain induced by a minor medical procedure (bone marrow aspiration) in dogs: comparison of pain scales in a pilot study. *J Vet Intern Med* 25: 1050–1056.

Guyatt GH, Kirshner B, Jaeschke R (1992) Measuring health status: what are the necessary measurement properties? *Clin Epidemiol* 45: 1341-1345

Ha ̈ger C, Biernot S, Buettner M, Glage S, Keubler LM, Held N, Belich EM et al (2017) The Sheep Grimace Scale as an indicator of post-operative distress and pain in laboratory sheep. *PLoS ONE* 12 (4): e0175839

Han S, Olonisakin TF, Pribis JP, Zupetic J, Yoon JH, Holleran KM, et al (2017) A checklist is associated with increased quality of reporting preclinical biomedical research: a systematic review. *PLoS ONE* 12: e0183591

Hansen BD (1997) Through a glass darkly: Using behavior to assess pain. *Semin Vet Med Surg (Small Anim)* 12:61-74

Hansen BD (2003) Assessment of pain in dogs: veterinary clinical studies. *ILAR J* 44 (3): 197-205

Hariton E, Locascio JJ (2018) Randomised controlled trials – the gold standard for effectiveness research. *BJOG* 125 (13): 1716

Harkins SW, Price DD, Braith J (1989) Effects of extraversion and neuroticism on experimental pain, clinical pain, and illness behavior. *Pain* 36: 209-218

Hellyer PW, Gaynor JS (1998) Acute post-surgical pain in dogs and cats. *The Compendium on continuing education for the practicing veterinarian (USA)* 20: 140-153

Hernandez-Avalos I, Mota-Rojas D, Mora-Medina P, Martinez-Burnes J, Casas Alvarado A, Verduzco-Mendoza A et al (2019) Review of different methods used for clinical recognition and assessment of pain in dogs and cats. *Int J Vet Sci and Med* 7(1): 43-54

Herzberg DE, Bustamante HA (2021) Animal models of chronic pain. Are naturally occurring diseases a potential model for translational research? *Austral J Vet Sci* 53: 47-54

Higgins J, Green S (2011) Cochrane handbook for systematic reviews of interventions. Version 5.1.0 ed *The Cochrane Collaboration*

Hoffman JM, Creevy KE, Franks A, O'Neill DG, Promislow DEL (2018) The companion dog as a model for human aging and mortality. *Aging Cell* 17: e12737

Hofmeister EH, King J, Read MR, Budsberg SC (2007) Sample size and statistical power in the small-animal analgesia literature. *J Small Anim Pract*. 48:76–79

Hofmeister EH, Barletta M, Shepard M, Brainard BM, Trim CM, Quandt J (2018) Agreement amongst anesthesiologists regarding postoperative pain assessment in dogs. *VAA* 45 (5): 695-702

Holden E, Calvo G, Collins M, Bell A, Reid J, Scott EM, Nolan AM (2014) Evaluation of facial expression in acute pain in cats. *JSAP* 55: 615-621

Holopherne-Doran D, Laboissière B, Gogny M (2010). Validation of the 4Avet postoperative pain scale in dogs and cats. *VAA* 37: 1-17
37, 1–17.

Holton LL, Scott EM, Nolan AM, Reid J, Welsh E, Flaherty D (1998a) Comparison of three methods used for assessment of pain in dogs. *J Am Vet Med Assoc* 212:61–66.

Holton LL, Scott EM, Nolan AM, Reid J, Welsh E (1998b) Relationship between physiological factors and clinical pain in dogs scored using a numerical rating scale. *J Small Anim Pract* 39: 469-474

Holton LL (2000) The measurement of pain in dogs. PhD thesis, University of Glasgow, UK

Holton LL, Pawson P, Nolan A, Reid J, Scott EM. (2001) Development of a behaviour- based scale to measure acute pain in dogs. *Vet Rec* 148:525–31

Ijichi C, Collins LM, Elwood RW (2014) Pain expression is linked to personality in horses. *Appl Anim Behav* Sci 152: 38e43.

Ioannidis JPA (2005) Why most published research findings are false. *PLoS Med* 2 (8): e124

Jaeschke R, Singer J, Guyatt G H (1989) Measurement of health status. Ascertaining the minimal clinically important difference. *Controlled Clinical Trials* 10: 407–415

Johnston DF and Turbitt LR (2021) Defining success in regional anaesthesia. *Anaesthesia* 76 (1): 40-52

Jones DS, Podolsky SH (2015) The history and fate of the gold standard. *The Lancet* 385: 1502-1503

Kazis LE, Anderson JJ, Meenan RF (1989) Effect sizes for interpreting changes in health status. *Med Care* 27 (3): S178-189

Keating SCJ, Thomas AA, Flecknell PA, leach MC (2012) Evaluation of EMLA cream for preventing pain during tattoing of rabbits: changes in physiological, behavioural and facial expression responses. *PLoS ONE* 7 (9): e44437

Keppel G, Wickens TD (2004) Design and analysis: a researcher's handobook. *Pearson*, fourth edition

Klinck MP, Mogil JS, Moreau M, Lascelles BDX, Flecknell PA, Poitte T, et al. (2017) Translational pain assessment: could natural animal models be the missing link? *Pain* 158:1633–1646

Kol A, Arzi B, Athanasiou KA, Farmer DL, Nolta JA, Rebhun RB, et al. (2015) Companion animals: translational scientist's new best friends. *Sci Transl Med*. 7: 308ps21

Koolhaas JM, Korte SM, De Boer SF, Van Der Vegt BJ, Van Reenen CG, Hopster H, De Jong IC, Ruis MAW, Blokhuis HJ (1999) Coping styles in animals: current status in behavior and stress-physiology. *Neurosci Biobehav Rev* 23: 925-935.

Kyles AE, Hardie EM, Hansen BD, Papich MG (1998) Comparison of transdermal fentanyl and intramuscular oxymorphone on post-operative behaviour after ovariohysterectomy in dogs. *Res Vet Sci* 65:245-251

Lachin JM (2016) Fallacies of Last Observation Carried Forward. *Clin Trials* 13 (2): 161-168

Langford DJ, Bailey AL, Chanda ML, Clarke SE, Drummond TE, Echol S, Glick S et al (2010) Coding of facial expressions of pain in the laboratory mouse. *Nature Methods* 7(6): 447-449

Lesaffre E (2008) Superiority, equivalence, and non-inferiority trials. *Bull NYU Hosp Jt Dis* 66 (2): 150-154

Leung V, Rousseau-Blass F, Beauchamp G, Pang DSJ (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: e0197882

Levine M, Ensom MHH (2001) Post hoc power analysis: an idea whose time has passed? *Pharmacotherapy* 21: 405-409

Lipsitch M, Tchetgen ET, Gohen T (2010) Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiol* 21 (3): 383-388

Lush L, Ijichi C (2018) A preliminary investigation into personality and pain in dogs. *J Vet Behav* 24: 62-68

Machado M, Oliveira da Silva IJ (2020) Body expressions of emotions: does animals have it? *J Anim Behav Biometeorol* 8: 1-10

MacRae AM, Makowska IJ, 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

Mantha S, Thisted R, Foss J, Ellis J, Roizen M (1993) A proposal to use confidence intervals for visual analogue scale data for pain measurement to determine clinical significance. *Anesth Analg* 77 (5): 1041-1047

Mathews K, Kronen PW, Lascelles D (2014) Guidelines for recognition, assessment and treatment of pain. *J Small Anim Pract* 55 (6): E10-E68

McCluskey A, Lalkhen AG (2007) Statistics II: central tendency and spread of data. *Continuing Education in Anaesthesia, Critical Care and Pain* 7 (4): 127-130

McKune CM, Pascoe PJ, Lascelles BDX, Kass PH (2014) The challenge of evaluating pain and a pre-incisional local anesthetic block. *PeerJ* 2:3341

McKune CM, Murrell JC, Nolan AM, White KL, Wright BD (2015) Nociception and pain. *Veterinary Anaesthesia and Analgesia: the fifth edition of Lumb and Jones* Chapter 29: 584-623

McLennan KM, Rebelo CJB, Corke MJ, Holmes MA, Leach MC, 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

Meeson RL, Todhunter RJ, Blunn G, Nuki G, Pitsillides AA (2019) Spontaneous dog osteoarthritis - a One Medicine vision. *Nat Rev Rheumatol* 15: 273–287

Melzack R, Torgerson WS (1971) On the language of pain. *Anesthesiology* 34: 50-59

Melzack R (1975) the McGill pain questionnaire: major properties and scoring methods. *Pain* 1: 277-299

Messick, S (1986), The once and future issues of validity: assessing the meaning and consequences of measurement. *ETS Research Report Series* i-24

Mich PM, Hellyer PW, Kogan L, Schoenfeld-Tacher R (2010) Effects of a pilot training program on veterinary students' pain knowledge, attitude, and assessment skills. *JVME* 37 (4): 358-368

Mischkowski D, Palacios-Barrios EE, Banker L, Dildine TC, Atlas LY (2018) Pain or nociception? Subjective experience mediates the effects of acute noxious heat on autonomic responses. *Pain* 159 (4): 699-711

Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, Elbourne D, et al (2010) CONSORT 2010 Explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* 340: c869

Molony V, Kent JE (1997) Assessment of acute pain in farm animals using behavioral and physiological measurements. *J Anim Sci* 75: 266–272

Moore SA (2016) Managing neuropathic pain in dogs. *Front Vet Sci* 3: 12

Moreno-Betancur M, Carlin JB, Brilleman SL, Tanamas SK, Peeters A, Wolfe R (2018) Survival analysis with time-dependent covariates subject to missing data or measurement error: Multiple Imputation for Joint Modeling (MIJM). *Biostatistics* 19 (4): 479-496

Morton DB, Griffiths PHM (1985) Guidelines on the recognition of pain, distress and discomfort in experimental animals and a hypothesis of assessment. *Vet Rec* 116: 431-436

Morton CM, Reid J, Scott EM, Holton LL, Nolan AM. (2005) Application of a scaling model to establish and validate an interval level pain scale for assessment of acute pain in dogs. *Am J Vet Res*. 66:2154–66

Moser P (2019) Out of Control? Managing Baseline Variability in Experimental Studies with Control Groups. In: Bespalov A, Michel M, Steckler T (eds). *Good*

Research Practice in Non-Clinical Pharmacology and Biomedicine. Handbook of Experimental Pharmacology vol 257: 101-117

Munn Z, Porritt K, Lockwood C, Aromataris E, Pearson A (2014) Establishing confidence in the output of qualitative research synthesis: the ConQual approach. *BMC Med Res Methodol* 14: 108

Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E (2018) Guidance for authors when choosing between a systematic or scoping review approach. *BMC Med Res Methodol* 18: 143

Murphey ED (2019) The AVMA animal health studies database. *Top Compan Anim Med* 37: 100361

Murrell JC, Psatha EP, Scott EM et al. (2008) Application of a modified form of the Glasgow pain scale in a veterinary teaching centre in the Netherlands. *Vet Rec* 162: 403– 408

Nair AS, Diwan S (2020) Pain scores and statistical analysis – the conundrum. *Ain-Shams J Anesthesiol* 12 (35)

Navarro E, Mainau E, Manteca X (2020) Development of a facial expression scale using farrowing as a model of pain in sows. *Animals* 10: 2113

Nimon KF (2012) Statistical assumptions of substantive analyses across the general linear model: a mini-review. *Front Psychol* 3: 322

Noble CE, Wiseman-Orr LM, Scott ME, Nolan AM, Reid J (2018) Development, initial validation and reliability testing of a web-based, generic feline health-related quality-of-life instrument. *J Feline Med Surg* 21 (2): 84-94

Olsen MF, Bjerre E, Hansen MD, Hilden J, Landler NE, Tendal B, Hróbjartsson A (2017) Pain relief that matters to patients: systematic review of empirical studies assessing the minimum clinically important difference in acute pain. *BMC Med* 15: 35

Orth EK, Gonzalez FJN, Pastrana CI, Berger JM, le Jeune SS, Davis EW, McLean AK (2020) Development of a donkey Grimace Scale to recognize pain in donkeys (*Equus asinus*) post castration. *Animals* 10:1411

Oyama MA, Ellenberg SS, Shaw PA (2017) Clinical trials in veterinary medicine: a new era brings new challenges. *J Vet Intern Med* 31: 970-978

Page MJ, Higgins JPT, Clayton G, Sterne JAC, Hróbjartsson A, Savovic´ J (2016) Empirical evidence of study design biases in randomized trials: systematic review of meta-epidemiological studies. PLoS ONE 11: e0159267

Pereira-Morales S, Arroyo-Novoa CM, Wysocki A, Sanzero Eller L (2018) Acute pain assessment in sedated patients in the Postanesthesia Care Unit. *Clin J Pain* 34 (8): 700-706

Peters MDJ, Godfrey CM, Khalil H, McInerney P, Parker D, Baldini Soares C (2015) Guidance for conducting systematic scoping reviews. *Int J Evid Based Healthc* 13: 141-146

Peters MDJ, Marnie C, Colquhoun H, Garritty CM, Hempel S, Horsley T, Langlois ET et al. (2021) Scoping reviews: reinforcing and advancing the methodology and application. *BMC* 10: 263

Polit DF, Beck CT (2006) The content validity index: are you sure you know what's being reported? Critique and recommendations. *Res Nurs Health* 29 (5): 489-497

Prasad K (1996) The Glasgow Coma Scale: a critical appraisal of its clinimetric properties. *J Clin Epidemiol* 49 (7): 755-763

Price J, Marques JM, Welsh EM, Waran NK (2002) Pilot epidemiological study of attitudes towards pain in horses. *Vet Rec* 151: 570-575

Prinja S, Gupta N, Verma R (2010) Censoring in clinical trials: review of survival analysis techniques. *Indian J Community Med* 35 (2): 217-221

Raja SN, Carr DB, Cohen M, Finnerup NB, Flor H, Gibson S, Keefe FJ, Mogil JS et al (2020) The revised International Association for the Study of Pain definition of pain: concepts, challenges, and compromises. *Pain* 161 (9): 1976-1982

Ramírez-Maestre C, Martínez AEL, Zarazaga RE (2004) Personality characteristics as differential variables of the pain experience. *J Behav Med* 27: 147-165

Ready LB, Edwards WTY (1992) Management of acute pain: a practical guide. *IASP Publications*, Seattle

Rehal S, Morris TP, Fielding K, Carpenter JR, Phillips PPJ (2016) Non-inferiority trials: are they inferior? A systematic review of reporting in major medical journals. *BMJ Open* 6: e012594

Reid J, Nolan AM, Hughes J, Lascelles D, Pawson P, Scott ME (2007) Development of the short-form Glasgow Composite Measure Pain Scale (CMPS-SF) and derivation of an analgesic intervention score. *Anim Welfare* 8:97–104

Reid J, Scott ME, Nolan AM, Wiseman-Orr L (2013) Pain assessment in animals. *In Practice* 35: 51-56

Reid J, Nolan AM, Scott ME (2018) Measuring pain in dogs and cats using structured behavioural observation. *The Vet J* 236: 72-79

Reijgwart ML, Schoemaker NJ, Pascuzzo R, Leach MC, Stodel M, de Nies L, Hendriksen CFM et al (2017) The composition and initial evaluation of a grimace scale in ferrets after surgical implantation of a telemetry probe. *PLoS ONE* 12 (11): e0187986

Rialland P, Authier S, Guillot M, del Castillo JRE, Veilleux-Lemieux D, Frank D, Gauvin D, Troncy E (2012) Validation of orthopedic postoperative pain assessment methods for dogs: a prospective, blinded, randomized, placebo-controlled study. *PlosOne* 7 (11): e49480

Rietmann TR, Stauffacher M, Bernasconi P, Auer JA, Weishaupt MA (2004) The association between heart rate, heart variability, endocrine and behavioural pain measures in horses suffering from laminitis. *J Vet Med A* 51: 218-225

Robertson-Plouch C, Stille JR, Liu P, Smith C, Brown D, Warner M, et al. (2019) A randomized clinical efficacy study targeting mPGES1 or EP4 in dogs with spontaneous osteoarthritis. *Sci Transl Med* 11: eaaw9993

Robinson KL, Bryan ME, Atkinson ES, Keeler MR, Hahn AW, Bryan JN (2020) Neutering is associated with developing hemangiosarcoma in dogs in the Veterinary Medical Database: an age and time-period matched case control study (1964-2003). *Can Vet J* 61 (5): 499-504

Rose S, Van Der Laan MJ (2009) Why match? Investigating matched case-control study designs with causal effect estimation. *Int J Biostat* 5 (1): art 1

Rufiange M, Rousseau-Blanc F, Pang DSJ (2019) Incomplete reporting of experimental studies and items associated with risk of bias in veterinary research. *Vet Rec Open* 6 (1): e000322

Rusticus SA, Lovato CY (2014) Impact of sample size and variability on the power and Type I error rates of equivalence tests: a simulation study. *Practical Assessment, Research, and Evaluation* 19: article 11

Safarkhani M, Moerbeek M (2017) A comparison of within-subjects and between-subjects designs in studies with discrete-time survival outcomes. *Open J Stat* 7: 305-322

Sanford J, Ewbank R, Molony V, Tavenor WD, Uvarov O (1986) Guidelines for the recognition and assessment of pain in animals. *Vet Rec* 118: 334-338

Schulz KF, Grimes DA (2002) Unequal group sizes in randomised trials: guarding against guessing. *Lancet* 359 (9310): 966-970

Sedgwick P (2012) Log transformation of data. *BMJ* 12 (345): e6727

Shibasaki WM, Martins RP (2018) Simple randomization may lead to unequal group sizes. Is that a problem? *Am J Orthodont Dentof Orthop* 154: 600–605

Singla NK, Meske DS, Desjardins PJ (2017) Exploring the interplay between rescue drugs, data imputation, and study outcomes: conceptual review and qualitative analysis of an acute pain data set. *Pain Ther* 6: 165-175

Slingsby L (2010) Considerations for prospective studies in animal analgesia. *Vet Anaesth Analg*. 37:303–305

Soriano Pastor JF, Monsalve Dolz V, Ibáñez Guerra E, Gómez Carretero P 2010 Personality and coping in neuropathic chronic pain: a predictable divorce. *Psicothema* 22 (4): 537-542.

Sotocinal SG, Sorge RE, Zaloum A, Tuttle AH, Martin LJ, Wieskopf JS, Mapplebeck JCS et al (2011) The Rat Grimace Scale: a partially automated method for quantifying pain in the laboratory rat via facial expressions. *Molecular Pain* 7: 55

Spacek A (2006) Modern concepts of acute and chronic pain management. *Biomedicine & Pharmacotherapy* 60 (7): 329-335

Srithunyarat T, Höglund OV, Hagman R, Olsson U, Stridsberg M, Lagerstedt A-S, et al (2016) Catestatin, vasostatin, cortisol, temperature, heart rate, respiratory rate, scores of the short form of the Glasgow composite measure pain scale and visual analog scale for stress and pain behavior in dogs before and after ovariohysterectomy. *BMC Res Notes* 9 (1):381

Srithunyarat T, Hagman R, Höglund OV, Stridsberg M, Olsson U, hanson J, Nonthakkotr C et al (2017) Catestatin, vasostatin, cortisol and pain in dogs suffering from traumatic bone fractures. *BMC Res Notes* 10:129

Steinmetz S, Tipold A, Loscher W (2013) Epilepsy after head injury in dogs: a natural model of posttraumatic epilepsy. *Epilepsia* 54: 580–588

Sterne JAC, White IR, Carlin BJ, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR (2009) Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 338: b2393

Streiner DL, Norman GR (1995) Health measurement scales: a practical guide to their development and use. Oxford, *Oxford Medical Publications* 145-146

Streiner DL, Norman GR (2008) Health Measurement Scales: A Practical Guide to Their Development and Use, Fourth Ed. *Oxford University Press*, Oxford

Stuart EA (2010) Matching methods for causal inference: a review and a look forward. *Stat Sci* 25 (1): 1-21

Sullivan GM, Feinn R (2012) Using effect size – or why p value is not enough. *J Grad Med Educ* 4 (3): 279-282

Tait N, Firth A, Reid J, Scott EM, Nolan AM (2011) Pilot study to investigate the responsiveness of the short form of the Glasgow Composite Measure Pain Scale (CMPS-SF) in Dogs. In *BSAVA Congress 2011*, Birmingham, 31 March – 3 April 2011

Tate AJ, Fischer H, Leigh AE, Kendrick KM (2006) Behavioural and neurophysiological evidence for face identity and face emotion processing in animals. *Phil Trans R Soc B* 361: 2155-2172

Testa B, Reid J, Scott ME, Murison PJ, Bell AM (2021) The short form of the Glasgow composite measure pain scale in post-operative analgesia studies in dogs: a scoping review. *Front Vet Sci* 8: 751949

Thurstone LL (1928) Attitudes can be measured. *Am J Sociology* 33: 529-554

Tovakol M, Dennick R (2011) Making sense of Cronbach's alpha. *Int J Med Educ* 27 (2): 53-55

Tschoner T (2021) Methods for pain assessment in calves and their use for the evaluation of pain during different procedures – a review. *Animals* 11: 1235

Unverzagt S, Prondzinsky R, Peinemann F (2013) Single-center trials tend provide larger treatment effects than multicenter trials: a systematic review. *J Clin Epidemiol* 66: 1271-1280

Viscardi A, Hunniford M, Lawlis P, Leach MC, Turner PV (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

Wang B, Wang H, Tu XM, Feng C (2017) Comparison of superiority, non-inferiority, and equivalence trials. *Shanghai Arch Psychiatry* 29 (6): 385-388

William AC (2002) Facial expression of pain: an evolutionary account. *Behav Brain Sci* 25 (4): 455-488

Wright JG, Young NL (1997) A comparison of different indices of responsiveness. *J Clin Epidemiol* 50 (3): 239-246

Zamanzadeh V, Ghahramanian A, Rassouli M, Abbaszadeh A, Alavi-Majd H, Nikanfar AR (2015) Design and implementation content validity study: development of an instrument for measuring patient-centered communication. *J Caring Sci* 4 (2): 165-178