Methods and Technologies for Improving Feasibility and Accuracy of Preclinical Systematic Reviews of *in vivo* Pain Neurobiology Data

Nadia Soliman

A thesis submitted to Imperial College London for the degree Doctor of Philosophy

Imperial College London

Faculty of Medicine, Department of Surgery and Cancer

Declaration of Originality

I declare that the work presented here is my own. All information sources have been appropriately referenced. All work was conducted and reported in accordance with Imperial College London's integrity policies. This thesis has not been submitted for another degree or any other professional purposes.

Funding

This work was funded by the Biotechnology and Biological Sciences Research Council (grant number BB/M011178/1).

Copyright

'The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non-Commercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.'

Relevant Publications and Presentations

Publications

1. BAHOR, Z., LIAO, J., CURRIE, G., AYDER, C., MACLEOD, M., MCCANN, S. K., BANNACH-BROWN, A., WEVER, K., **SOLIMAN, N**., WANG, Q., DORAN-CONSTANT, L., YOUNG, L., SENA, E. S. & SENA, C. 2021. Development and uptake of an online systematic review platform: the early years of the CAMARADES Systematic Review Facility (SyRF). BMJ Open Science, 5, e100103.

2. BANNACH-BROWN, A., HAIR, K., BAHOR, Z., **SOLIMAN, N**., M., M. & LIAO, J. 2021. Technological advances in preclinical meta-research. BMJ Open Science, 0, e100131.

3. FINN, D. P., HAROUTOUNIAN, S., HOHMANN, A. G., KRANE, E., **SOLIMAN, N.** & RICE, A. S. C. 2021. Cannabinoids, the endocannabinoid system, and pain: a review of preclinical studies. PAIN, Articles in Press.

4. HAROUTOUNIAN, S., ARENDT-NIELSEN, L., BELTON, J., BLYTH, F. M., DEGENHARDT, L., DI FORTI, M., ECCLESTON, C., FINN, D. P., FINNERUP, N. B., FISHER, E., FOGARTY, A. E., GILRON, I., HOHMANN, A. G., KALSO, E., KRANE, E., MOHIUDDIN, M., MOORE, R. A., ROWBOTHAM, M., **SOLIMAN, N**., WALLACE, M., ZINBOONYAHGOON, N. & RICE, A. S. 2021. IASP Presidential Task Force on Cannabis and Cannabinoid Analgesia: research agenda on the use of cannabinoids, cannabis, and cannabis-based medicines for pain management. PAIN, Articles in Press.

5. **SOLIMAN, N**., HAROUTOUNIAN, S., HOHMANN, A. G., KRANE, E., LIAO, J., MACLEOD, M., SEGELCKE, D., SENA, C., THOMAS, J., VOLLERT, J., WEVER, K., ALAVERDYAN, H., BARAKAT, A., BARTHLOW, T., BOZER, A. L. H., DAVIDSON, A., DIAZ-DELCASTILLO, M., DOLGORUKOVA, A., FERDOUSI, M. I., HEALY, C., HONG, S., HOPKINS, M., JAMES, A., LEAKE, H. B., MALEWICZ, N. M., MANSFIELD, M., MARDON, A. K., MATTIMOE, D., MCLOONE, D. P., NOES-HOLT, G., POGATZKI-ZAHN, E. M., POWER, E., PRADIER, B., ROMANOS-SIRAKIS, E., SEGELCKE, A., VINAGRE, R., YANES, J. A., ZHANG, J., ZHANG, X. Y., FINN, D. P. & RICE, A. S. C. 2021. Systematic review and meta-analysis of cannabinoids, cannabis-based medicines, and endocannabinoid system modulators tested for antinociceptive effects in animal models of injury-related or pathological persistent pain. Pain, 162, S26-s44.

6. **SOLIMAN, N**., HOHMANN, A. G., HAROUTOUNIAN, S., WEVER, K., RICE, A. S. C. & FINN, D. P. 2019a. A protocol for the systematic review and meta-analysis of studies in which cannabinoids were tested for antinociceptive effects in animal models of pathological or injury-related persistent pain. Pain reports, 4, e766-e766.

7. **SOLIMAN, N**., OKUSE, K. & RICE, A. S. C. 2019b. VGF: a biomarker and potential target for the treatment of neuropathic pain? PAIN Reports, 4, e786.

8. **SOLIMAN, N**., RICE, A. S. C. & VOLLERT, J. 2020. A practical guide to preclinical systematic review and meta-analysis. PAIN, 161, 1949-1954.

9. ZHANG, X. Y., VOLLERT, J., SENA, E. S., RICE, A. S. & **SOLIMAN, N**. 2021. A protocol for the systematic review and meta-analysis of thigmotactic behaviour in the open field test in rodent models associated with persistent pain. BMJ Open Science, 5, e100135.

Presentations

1. **Introduction to Preclinical Systematic Reviews**, University of Aberdeen, April 2020

2. **Introduction to Preclinical Systematic Reviews**, Pain Research Forum, May 2020

3. **Cannabinoid Research for Pain: Crowd Science Approaches to Collect and Aggregate Evidence**, Society for Paediatric Anaesthesia 34th Meeting (Virtual), October 2020

4. **Crowd Science for Collecting Data in Preclinical Systematic Reviews**, Evidence Synthesis Ireland, November 2020

5. **The Importance of Preclinical Systematic Reviews: Opportunities for Research and Researchers**, Washington University in St Louis, November 2020

6. **Systematic review and meta-analysis of studies in which cannabinoids were tested for antinociceptive effects in animal models of pathological or injury-related persistent pain,** International Association for the Study of Pain Virtual Series on Pain, March 2021

7. **Benefits, Challenges and Emerging Technologies for Systematic Reviews and Meta-Analyses of Preclinical Pain Studies**, International Association for the Study of Pain World Congress 2021, June 2021

Abbreviations

3Rs	Replacement, reduction, and refinement
95 % CI	95 % confidence interval
ABHD6	Abhydrolase domain containing 6
ARRIVE	Animal Research: Reporting of In Vivo Experiments
CAMARADES	Collaborative Approach to Meta Analysis and Review of Animal Data from Experimental Studies
CB	Cannabinoid
CBD	Cannabidiol
CFA	Complete Freund's Adjuvant
CRediT	Contributor Roles Taxonomy
DAGL	Diacylglycerol lipase inhibitor
EDA	Experimental Design Assistant
Eppi-Centre	The Evidence for Policy and Practice Information and Co-ordinating Centre
FAAH	Fatty acid amide hydrolase
FABP	Fatty acid-binding protein
IASP	International Association for the Study of Pain
ICJME	International Committee of Medical Journal Editors
MGL	monoacylglycerol lipase
ML	Machine learning
NAAA	N-acylethanolamine-hydrolysing acid amidase
NC3Rs	National Centre for the Replacement, Refinement, & Reduction of Animals in Research
NP	Neuropathic pain
PAM	Positive allosteric modulator
PEA	Palmitoylethanolamide
PPAR	Peroxisome proliferator-activated receptor
SD	Standard deviation

SE	Standard error
SEM	Standard error of the mean
SMD	Standardised mean difference
SWiM	Synthesis Without Meta-analysis
SYRCLE	SYstematic Review Center for Laboratory Animal Experimentation
SyRF	Systematic Review Facility
THC	delta-9 tetrahydrocannabinol
TRPA1	Transient receptor potential ankyrin 1
TRPV1	Transient receptor potential vanilloid receptor 1

Abstract

The failure to translate preclinical success to the clinic has led researchers to question the predictive validity of animal models with inadequate study design, lack of rigour and opaque reporting potentially responsible for translational failures. Preclinical systematic reviews aim to answer a range of research questions to improve research value and translational ability of results. However, there are several challenges which limit their feasibility including the exponentially increasing, large volume of literature and resource intensive methods.

The aims were to (i) address key areas of neurobiology to provide empirical evidence to improve and guide future preclinical research and (ii) evaluate methods to improve the feasibility, efficiency, and accuracy of preclinical systematic reviews, including crowd science, machine learning for study selection, and semi-automation of data extraction (the application Graph2Data).

I performed a narrative review supported by a systematic search and study selection to explore the role of VGF in neuropathic pain. The "systematic review and mete-analysis of studies in which cannabinoids were tested for antinociceptive effects in animal models of pathological or injury-related persistent pain" quantified anti-nociceptive efficacy and assessed the reliability of the data. Crowd science and machine learning were employed and assessed. Graph2Data was further developed and integrated into the online Systematic Review Facility (SyRF).

The reviews demonstrate that improved rigour and transparency of design, conduct, analysis, and reporting is required and the difference between the animal and clinical population highlights the importance for development of better validated animal models. Increasing biological variation will likely improve the generalisability of results and bridge the gap between animal models and the clinical population. Crowd science and machine learning are beneficial, mutually supportive, viable strategies for improving feasibility. The continued development and incorporation of machine assisted methods in open-source, online systematic review software is necessary for accurate and timely evidence synthesis.

Acknowledgements

I have had the privilege to serve as a British Army Officer and my return to academia has undoubtedly been shaped by lessons learned during my military service.

My PhD journey got off to a shaky start, so I am relieved to still be here. I am indebted to Professor Andrew Rice for the opportunity and his unwavering support throughout. He has generously given me the opportunities to develop and grow intellectually, allowed me to explore and challenge and build a network of talented researchers. The Pain Group is wonderfully diverse, and I would like to thank them all for their constructive criticisms, collaboration, and camaraderie.

I would also like to thank my secondary supervisors, Professors Malcolm Macleod and James Thomas for their continued support. Both introduced me to the world of meta-research and the complexity of automation technologies. With Professor Macleod come CAMARADES. The most welcoming of teams, they have always generously given their time to help and ensured that every visit to Edinburgh is not all work! A special thank you to Chris Sena, for his patience, commitment, and software development skills which undoubtedly led to the successful completion of the Cannabis & Pain review.

One of the challenges of being a mature student and having left a professional career is the niggling doubt that you have done the wrong thing which is compounded by the research rollercoaster and by 'being at the bottom' again. Professor Stephen Curry gave me the courage to share my leadership experiences at a Department for Life Sciences, Imperial College London, Women's Week talk. Since then, I have been working closely with Professor Marcus Munafò who has assisted me in the development and delivery of leadership training for academics. Leadership development is an area that I can help others and add value to improve our research culture so that researchers and research can thrive.

Lastly, a big thank you to my friends and family. I can't list everyone but would like to mention the mighty 11 platoon for always being fabulous, pushing boundaries, and smashing through ceilings, they are the best friends anyone could wish for. And finally, to Richard, for his support, encouragement, having a lot of patience and keeping the life in work-life balance!

Table of Contents

Declaration of Originality **2**
Funding **2**
Copyright **2**
Relevant Publications and Presentations **3**
Publications 3
Presentations 4
Abbreviations **5**
Abstract **7**
Acknowledgements **8**
List of Figures **14**
List of Tables **18**
1. Introduction **21**
1.1. Challenges of Translational Pain Research 21
1.2. Limitations of Preclinical Study Design 22
1.2.1. Experimental and Reporting Bias 23
1.3. Waste in Research and Systematic Reviews 23
1.4. Importance of Preclinical Systematic Reviews 25
1.4.1. Internal Validity 25
1.4.2. External validity 26
1.4.3. Evidence-based decisions for clinical trials 26
1.4.4. Publication Bias 27
1.4.5. Replacement, Refinement and Reduction of Animals in Research 27
1.5. The Practical Challenges of Systematic Reviews 27
1.6. Improving feasibility 30
1.6.1. Automation Technologies 30
1.6.2. Crowd Science 32
1.7. Aims and Objectives ..... **Error! Bookmark not defined.**
2. Evidence Synthesis: Narrative vs Systematic Reviews **35**
2.1. Introduction 35
2.1.1. The Different Roles of Narrative and Systematic Reviews 38
2.1.2. Strengths and Limitations of a Narrative and Systematic Review: Methodology, Quality and Reproducibility 39
2.2. Aims 41
2.3. Methods 41

2.3.1. ‘VGF: a biomarker and potential for the treatment of neuropathic pain?’ (Soliman et al., 2019b) 41
2.3.2. ‘Cannabinoids, the endocannabinoid system, and pain: a review of preclinical studies’ (Finn et al., 2021) 42
..... 45
2.4. Discussion 46
2.4.1. Systematic Review Processes Reduce Selection and Confirmation Bias in Narrative Reviews 46
2.4.2. Narrative Reviews can be Complimented by Systematic Review Processes 47
2.4.3. Methodological Quality and Hierarchy of Evidence Synthesis 48
2.4.4. Reducing Reporting and Citation Bias 49
2.4.5. Feasibility 51
2.4.6. The Case for Mixed Methods and Systematic Reviews 51
2.5. Conclusion 52
2.6. Recommendations 52
2.7. Future Research Questions 52
3. VGF, a biomarker and potential target for the treatment of neuropathic pain? 53
3.1. Background 53
3.2. Introduction 54
3.3. Aims 56
3.4. Methods 56
3.4.1. Literature Search 56
3.4.2. Study Selection 57
3.4.3. Data Extraction 57
3.5. Results 57
3.5.1. Study Characteristics 59
3.5.2. Internal Validity / Risk of bias 64
3.6. Discussion 64
3.6.1. Localisation and Upregulation of VGF 64
3.6.2. Pro-nociceptive effects of exogenous VGF-derived peptides 68
3.6.3. Identification of receptors, their expression and their role in pro-nociceptive signalling 70
3.7. Limitations 74
3.8. Future Research Questions 76
3.9. Conclusion 77
4. Cannabis & Pain Preclinical Systematic Review - Part 1: Mechanisms of a Systematic Review 78
4.1. Introduction 78

4.2. Methods & Results 79
4.2.1. Project Timelines 79
4.2.2. Collaboration & Communication 80
4.2.3. The CAMARADES Systematic Review Facility (SyRF) 85
4.2.4. Systematic Review Methods 86
4.2.5. Preparing Data for Synthesis 102
4.2.6. Meta-analysis 102
4.2.7. Discussion 103
5. Cannabis & Pain Preclinical Systematic Review – Part 2: The Results and Further Analyses . 106
5.1. Introduction 106
5.2. Methods 107
5.2.1. Risk of Bias 107
5.2.2. Meta-Analysis 108
5.2.3. Publication Bias 111
5.3. Results 112
5.3.1. Study Characteristics 113
5.3.2. Risk of Bias 116
5.3.3. Meta-analysis of the antinociceptive efficacy of treatment with a cannabinoid, cannabis-based medicine, or endocannabinoid system modulator 117
5.3.4. Effects of Study Characteristics on Antinociceptive Efficacy 118
5.3.5. Antinociceptive Efficacy of Drug Classes in Different Model Types 124
5.3.6. Antinociceptive Efficacy of Drug Classes in Different Model Types Assessed using Mechanical Limb Withdrawal 131
5.3.7. Exploring Sources of Heterogeneity with Meta-Regression 134
5.3.8. Impact of Methodological Quality 151
5.3.9. Prevalence of Publication Bias at the Study Level 152
5.4. Discussion 155
5.4.1. Antinociceptive Efficacy of Cannabinoids, Cannabis-based Medicines, and Endocannabinoid System Modulators 156
5.4.2. External Validity 157
5.4.3. Internal Validity 159
5.4.4. Publication Bias 160
5.4.5. Utility of Preclinical Systematic Reviews 161
5.4.6. Limitations 162
5.5. Future Research 164
5.6. Conclusion 166

6. Crowd Science: Challenging Traditional Ways of Working 167
6.1. Introduction 167
6.2. Aims and Objectives 168
6.3. Methods 168
6.3.1. Ethics 168
6.3.2. Recruitment 169
6.3.3. Training 169
6.3.4. Communication 170
6.3.5. Efficiency 170
6.4. Results 171
6.4.1. Crowd Recruitment, Training and Contribution 171
6.4.2. Communication 172
6.4.3. Efficiency 172
6.4.4. Crowd feedback 175
6.5. Discussion 184
6.5.1. A small number of contributors make the largest share of the contributions 184
6.5.2. Efficiency Savings May be Possible 185
6.5.3. Mutual Benefits 186
6.5.4. Crowd Science Facilitates Community Engagement, Networking, and Collaborative Working 188
6.5.5. Crowd Science Requires Close Project Leadership and Management 188
6.6. Future Directions 189
6.7. Conclusion 192
7. Machine Learning for Study Selection 193
7.1. Introduction 193
7.1.1. Screening 193
7.1.2. Machine learning for Study Selection 194
7.2. Aim 198
7.3. Methods 198
7.3.1. The Tool 198
7.3.2. Machine Learning for Study Selection; Training and Validation 199
7.3.3. Machine Learning for Error Analysis 200
7.4. Results 201
7.4.1. Machine Learning could not be used for prioritised screening 201
7.4.2. Machine Learning for Error Analysis 201
7.5. Discussion 204

7.5.1. Error Analysis 204
7.5.2. Considerations for Use 205
7.5.3. Limitations 208
7.6. Future Research Questions 208
7.7. Conclusion 209
8. Graph2Data: semi-automation of data extraction from graphs 210
8.1. Introduction 210
8.1.1. Why a meta-analysis? 210
8.1.2. Extracting Data from Graphs for Meta-Analyses 210
8.2. Aim 212
8.3. Methods 213
8.3.1. Update and Enhance Graph2Data 213
8.3.2. Randomised Controlled Trial 214
8.4. Results 215
8.4.1. Identification of graph types 215
8.4.2. Development of Graph2Data 215
8.4.3. Randomised Controlled Trial Protocol 219
8.5. Discussion 225
9. Discussion 227
9.1. The Role of Animals in Pain Research 230
9.2. Methodological Quality: poor reporting practises? 230
9.3. Animal Models for Pain Research 234
9.3.1. Poor Animal Models 234
9.3.2. Outcome Measures 235
9.3.3. Environmental Factors 237
9.3.4. Paradigm Shift Needed 238
9.4. Benefits of Systematic Reviews 239
9.5. The Feasibility Challenges of Systematic Reviews 242
9.5.1. Machine Learning for Study Selection 243
9.5.2. Crowd Science 244
9.6. Limitations 246
9.6.1. The Reviews 246
9.6.2. Investigation of Crowd Science and Machine Learning 248
9.7. Conclusion: Future Research & Future of Systematic Reviews 249
References 253

List of Figures

Figure 1.1. The 5 Pillars of Research Waste *(Moher et al., 2016). Evidence confirms imbalanced research question selection (Pillar 1), poor study design and execution (Pillar 2), poor regulation and evaluation (Pillar 3), non-publication/publication bias (Pillar 4) and poor reporting (Pillar 5) as sources of avoidable research waste. 24*

Figure 1.2. Example of the Systematic Review workflow. *The platforms that can be used at each stage are also represented (Soliman et al., 2020). 28*

Figure 1.3. Number of Pain Articles Published Yearly. *Graph depicting the increasing number of pain articles published each year. PubMed search was conducted using the MeSH term 'Pain' up to the year 2019. ..... 29*

Figure 1.4. *Automation technologies are being developed for the different stages of the review process. Machine learning and text mining have improved the feasibility and efficiency of the early stages of the process. Technological developments for the latter stages are in their infancy. 30*

Figure 2.1. The cumulative impact of reporting and citation bias on the evidence base for antidepressants (de Vries et al., 2018). *(a) represents the initial cohort of completed trials (53 positive and 52 negative or questionable). Each trial is a single-coloured circle, and the colour represents the results or presence of spin. (b) 52 of the positive trials were published, and only 25 of the negative trials were published and by (c) 10 of the negative trials become 'positive' in the literature by omitting the negative outcomes or switching the status of primary and secondary outcome measures. (d) Of the remaining 15 negative trials, the abstract of five trials concluded the treatment was effective and a further five had mild spin by suggesting that the treatment is numerically superior to placebo. Only four trials remained 'negative' and reported that the treatment was not more effective than placebo. In (e) the size of the circles represents the relative number of citations received by that category of studies which shows that positive trials were cited three times more frequently than negative trials. 50*

Figure 3.1. The vgf gene and its derived peptides. *The VGF polypeptide is the precursor of several biologically active peptides. The gene contains specific amino acid sequences, as shown, of the pro VGF polypeptide and C-terminal VGF-derived peptides which are highly conserved between species. 55*

Figure 3.2. Study flow diagram. *Screening and identification of publications investigating role of VGF in pain pathways. Appendix A references included studies. Updated with the additional search results since the publishing of Soliman et al. (2019b). 58*

Figure 3.3. The role of VGF-derived neuropeptides in (A) central and (B) peripheral sensitisation and (C) neuropathic pain. *(A) Microglia in the spinal dorsal horn are activated by VGF-derived neuropeptides following nerve injury through their interaction with either or potentially both the C3a1 or gC1q receptors. In turn, these activated microglia release several pro-inflammatory cytokines, chemokines and other agents which interact with neurons associated with pain transmission and increase the excitability of the neurons contributing to NP. Pro-inflammatory mediators released e.g., TNFα, IL-1β, IL-6, nitric oxide (NO), ATP and prostaglandins (PGs), initiate a self-propagating mechanism of enhanced cytokine expression by microglial cells. This leads to an increase in intracellular calcium, and activation of the p38 and MAPK pathway. (B) Damaged and spared fibres are pictured together. After peripheral nerve injury, non-neuronal cells including macrophages accumulate around the damaged cells. The upregulation and release of VGF-derived neuropeptides activate the macrophages through their interaction with either or potentially both the C3a1 or gC1qR receptors. The macrophages secrete factors (such as TNFα, IL-1β, IL-6, chemokine (C–C motif) ligand 2 (CCL2), PGs and nerve growth factor (NGF)) which elicit peripheral sensitisation and cause chronic neuroinflammation which maintains sensory abnormalities. (A) and (B) were created with Biorender.com. (C) The graph abstract highlights the potential role of vgf and VGF-derived peptides in the initiation, development, and maintenance of neuropathic pain. 73*

Figure 4.1. Timeline of the Cannabis & Pain Preclinical Systematic Review. *Each stage is presented with approximate start and finish dates and number of weeks that each stage took to complete. Protocol; protocol development; Screen; manual screening of 10,816 articles retrieved in the search; Extract; annotation and data extraction of the 473 included studies; Reconcile; process of identifying agreement and reconciling disagreement between two independent reviewers; Analysis; data preparation and synthesis and writing of the original manuscript; Report; first submission, peer review and acceptance in the journal PAIN. 80*

Figure 4.2. *Visualisation of the overlap between control and treatment group distributions of the overall SMD effect size of 1.32 (Magnusson, 2020, Soliman et al., 2021).* *83*
Figure 4.3. *A Venn diagram representing the three different components of the search strategy. The overlap provides the potentially relevant studies.* *87*
Figure 4.4. Screenshot of the SyRF user interface for study selection. *On the left-hand side are the prompts for inclusion/exclusion criteria, the green writing denotes the inclusion criteria, the red writing the exclusion criteria. The title, authors, journal and date and abstract are presented in the middle with the decision buttons 'include', 'exclude', and 'next' on the right-hand side.* *89*
Figure 4.5. *The study tab contains annotation questions that relate to the whole study e.g., the questions relating to the non-behavioural experimental procedures and risk of bias.* *94*
Figure 4.6. *The disease model tab contained questions relating to the model induction including type of model and whether it was surgically or pharmacological induced.* *95*
Figure 4.7. *The treatment tab contained questions relating to the intervention and included intervention name, dose, route of administration and dosing regime.* *96*
Figure 4.8. *The outcome assessment tab contained questions relating to the pain-associated behavioural outcome assessment used including type of assessment, the structure of the data presented i.e., average and variance type, direction of effect and the method employed.* *97*
Figure 4.9. *The cohort assessment tab contained questions relating to the animals' species, strain, sex, age, weight, and the sample size and combined the previously entered information; disease model, intervention, and outcome(s).* *98*
Figure 4.10. *The Experiment tab allowed for the entry of outcome data extracted from a graph, table or text and included the timepoint, mean and variance.* *99*
Figure 4.11. *Screenshot of the reconciliation version of SyRF which allowed the presentation of two reviewers' completed annotation and data extraction forms side by side for the identification of agreements and disagreements.* *101*
Figure 5.1. *A flow diagram of articles identified from the bibliographic search of 3 electronic databases, PubMed, Web of Science and Ovid Embase, conducted on 9 April 2019. The diagram provides the breakdown of records through de-duplication, screening, eligibility until final inclusion in both qualitative and quantitative analysis. Reported in accordance with the PRISMA guidelines. W of S, Web of Science.* *113*
Figure 5.2. *The reporting of methodological quality criteria (A) and a summary bar plot showing the proportion of studies with a given risk of bias for each methodological quality criteria (B) for the 374 included studies.* .. *117*
Figure 5.3. *A caterpillar plot of the 1,544 nested comparisons extracted from the 374 studies included in the meta-analysis. Hedge's g standardised mean differences (SMD) were calculated for each comparison. Effect sizes were pooled using the random effects model and heterogeneity estimated with the restricted maximum likelihood model. Overall effect size=1.321. Q=4101.26, d.f. 1543, p <0.0001.* I^2 *= 61.58 %.* *118*
Figure 5.4. *Forest plot of drug classes assessed in rat (A) and mouse (B) models of injury related or pathological persistent pain. The size of the square represents the weight (%). The weight is the influence that individual subgroup has on the pooled result. N denotes the number of animals that contribute to that subgroup. K denotes the number of comparisons that comprise each subgroup.* *119*
Figure 5.5. *Forest plots of study design characteristics of experiments in which treatments were assessed for antinociceptive efficacy in rat models of persistent or injury-related persistent pain. Model type (A), strain (B), and sex (C) and outcome assessment type (D) account for a significant proportion of heterogeneity. The size of the square represents the weight (%). The weight is the influence that individual subgroup has on the pooled result. N denotes the number of animals that contribute to that subgroup.* *122*
Figure 5.6. *Forest plots of study design characteristics of experiments in which treatments were assessed for antinociceptive efficacy in mouse models of persistent or injury-related persistent pain. Model type (A), strain (B), sex (C), outcome assessment type (D) all account for a significant proportion of heterogeneity. The size of the square represents the weight (%). The weight is the influence that individual subgroup has on the pooled result. N denotes the number of animals that contribute to that subgroup.* *123*
Figure 5.7. *Model-averaged predictor importance plot (without consideration for interactions between predictors) for rats in which cannabinoids, cannabis-based medicines and endocannabinoid system modulators assessed for antinociceptive efficacy in animal models of injury-related of pathological persistent pain. Outcome assessment type; 1.00; Model type; 1.00; Strain; 1.00; Drug class; 0.99; Strain; 0.98.* *141*

Figure 5.8. *Model-averaged predictor importance plot (without consideration for interactions between predictors) for mice in which cannabinoids, cannabis-based medicines and endocannabinoid system modulators assessed for antinociceptive efficacy in animal models of injury-related of pathological persistent pain. Drug Class; 1.00; Strain; 1.00; Model type; 1.00; Sex; 1.00; Outcome assessment type; 0.99.* *151*
Figure 5.9. *Effect sizes (and variance) associated with the reporting of methodological quality criteria.* *152*
Figure 5.10. (Incorrect) Assessment of publication bias*. (A) visual inspection of the funnel plot suggests asymmetry. Filled circles denote the published studies and the vertical dotted line the overall effect size. (B) Trim and fill analysis imputed 93 missing studies (open circles). The contours denote the significance of individual study effects.* *153*
Figure 5.11. *SMD vs SE funnel plots vs sample-size based precision plots (A) illustration of the distortion of SMD vs SE funnel plot. (B) plotting precision as 1/√n.* *153*
Figure 5.12. *Assessment of publication bias (study level). Visual inspection of the funnel plot does not suggest asymmetry.* *154*
Figure 5.13. Assessment of publication bias based upon individual outcome comparisons*. Visual inspection of the funnel plot suggests asymmetry. Filled circles denote the published studies and the vertical dotted line the overall effect size. (B) Trim and fill analysis imputed 910 theoretically missing outcomes (open circles).* *155*
Figure 6.1. *Locations of crowd scientists pinned to a world map.* *171*
Figure 6.2. *Crowd scientists' feedback on joining and contributing to the project. Survey results based upon the feedback of 32 participants.* *177*
Figure 6.3. *Crowd scientists' feedback on the impact that contributing to the project has had on them. Survey results based upon the feedback of 32 participants.* *178*
Figure 7.1 Schematic of the use of ML in an iterative manner for error analysis. *Each step re-run until the stopping criteria was met (no more changes in screening decisions).* *200*
Figure 8.1*.* ***Different graph types that feature in preclinical research publications.*** *(A) Bar plot, (B), line graph, (C) dot plot, (D) box plot, (E) scatter plot and (F) area under the curve. Graphs pictured were plotted for the Graph2Data non-inferiority trial (Cramond et al., 2019).* *213*
Figure 8.2. *The SyRF – Graph2Data workflow. An image(s) of a graph is selected from the PDF of a study. The graph can be tagged with an associated outcome measure. The graph is then displayed as an image. Upon clicking the image, the Graph2Data application is launched. Once data extraction is complete, the data are then passed in to SyRF and populate a table in SyRF. A thumbnail image of the selected graph and corresponding coordinates is also 'saved' on SyRF which allows for Graph2Data application to be re-opened and amend the data extraction, if necessary.* *216*
Figure 8.3. *SyRF and Graph2Data component workflow. A reviewer selects to open the study PDF. SyRF makes an API call to the server to load a JSON document which displays the PDF on the screen. The reviewer is then able to use the Graph Selector and on clicking the image of the graph, Graph2Data is launched. Once data extraction is completed, the data is emitted to SyRF, and a table of outcome data is populated. When the reviewer saves the session, SyRF makes an API call, and the session is saved on the server.* *217*
Figure 8.4. The Graph2Data user interface. *The series of selection boxes on the left-hand side need to be completed prior to being able to start collecting data. Under outcome, the data type and structure can be selected from a drop-down list e.g., mean and standard deviation. Cohorts can be entered manually or pre-populated based upon previous entries into SyRF. Some graphs have time series data and there is the option to select whether to use the application to calibrate the axes or extract the values manually. Lastly, is the option to calibrate the axes by selecting the corresponding minimum and maximum values and inputting their values. Selecting 'Start Collecting Data' builds the table on the right-hand side and the individual data points can be selected to populate the table.* *218*
Figure 8.5. The schematic for the randomised controlled trial of Graph2Data: 1. Studies. *Studies will be randomly assigned to a cohort using multivariate random sampling to ensure even distribution between the cohorts of both journal and year.* ***2. Cohort.*** *Cohort A contains half of the available studies and cohorts B and C each have one quarter of the available studies.* ***3. Treatment****. The cohorts are organised to aid the between and within group comparisons. Within each cohort studies are assigned a treatment; either the control (traditional methods of data extraction) or the Graph2Data tool (intervention). A) Cohort A. Each study is allocated to both control and intervention for data extraction. B) Cohort B. Data is extracted from these studies using the current methods only (control). C) Cohort C. data is extracted from these studies using Graph2Data*

(intervention). ***4. Reviewers****. The dashed lines from the reviewers denotes that the type of study the reviewer is directed to will depend upon what they have already completed. This is to ensure that reviewers are presented evenly with both an intervention and control study thereby reducing the impact of a single reviewer on the results. 219*

Figure 9.1. Improving the feasibility, accuracy, and efficiency for the potential of preclinical systematic reviews to be realised. *(A) This pyramid reflects our current research culture. The first step is making systematic reviews possible with infrastructure e.g., SyRF and development of automated technologies including Graph2Data for data extraction from graphs and text mining and machine learning for screening and annotation. Second, is make it easy. The use of new technologies will be easier if integrated into a systematic review platform with a user-friendly interface thereby negating the need for subject matter expertise to use. Thirdly, performing preclinical systematic reviews need to become normative. Building a community through crowd science can create and facilitate the opportunities to share, learn and perform systematic reviews together. Fourth, it needs to be rewarding or rewarded. As the uptake increases, more researchers will be able to shape their use and gain to inform their research practises. Concurrently, development of new technologies will increase. Concomitantly, if we researchers say that we value being able to synthesise evidence in a timely manner and identify research conducted with the highest rigour from which we can build upon it will lead to further reward through scientific progress. For the last criteria, policy; the conservative approach is that policy comes once all other steps have been developed. However, this will come at a continued research waste cost. (B) A more radical approach is to have policy on the bottom. Make preclinical systematic reviews a condition of funding. Ensure funding is made available for their conduct and continued development. It is likely that doing so will accelerate development of new technology to improve infrastructure etc. and increase uptake. Similarly, each step of the pyramid should be valued and rewarded. 251*

List of Tables

Table 1.1. *Automation of the Stages of Systematic Reviews.* *32*
Table 2.1. *Different review types, their aims, and methods (adapted from Grant and Booth (2009).* *36*
Table 2.2. *Comparison of Systematic and Narrative Review Aims and Features.* *39*
Table 2.3. *Comparison of the Underlying Methodology of each Stage of the Systematic and Narrative Review Processes.* *40*
Table 2.4. Summary of the effects of cannabinoids, cannabis-based medicines, and endocannabinoid system modulators on pain-related behaviour in the most frequently used rodent models of inflammation (Finn et al., 2021). *K; number of comparisons, N; number of animals within each subgroup; NT; not tested, ns; not significant.* *44*
Table 2.5. Summary of the effects of cannabinoids, cannabis-based medicines, and endocannabinoid system modulators on pain-related behaviour in the most frequently used rodent models of neuropathic pain (Finn et al., 2021). *K; number of comparisons, N; number of animals within each subgroup; NT; not tested, ns; not significant.* *45*
Table 3.1. *Study Characteristics for in vivo experiments assessing the effects of exogenous VGF-derived peptides. TLQP-21 is the most frequently use VGF-derived peptide. Its effect most commonly assessed in mice following administration into the spinal cord. Pain associated behaviours were most frequently assessed in the warm-water tail immersion assay.* *60*
Table 3.2. Study Characteristics measuring expression of VGF in animal nerve injury models. *VGF was upregulated in animal models of peripheral nerve injury and inflammatory pain. (Sample size refers to the number of animals which contributed tissue samples) Updated from (Soliman et al., 2019b).* *61*
Table 3.3. *Study Characteristics of in vivo experiments investigating antagonism of the effect of VGF-derived neuropeptides.* *62*
Table 3.4. *Characteristics of included in vitro studies.* *62*
Table 3.5. Comparison of data from the tail immersion assay. *The tail immersion assay was most frequently used to measure pain-associated behaviours following the administration of exogenous VGF-derived peptides.* *63*
Table 3.6. Methodological quality of studies evaluated. *To reduce selection bias, only two publications reported randomisation. Conversely, no publications reported allocation concealment or reported conducting a sample size calculation. Minimising the impact of performance and detection bias, blinding was reported as being conducted in 5 publications. To reduce the impact of attrition bias, the reporting of animal exclusions was less prevalent and reported in only one publication. In contrast, all publications reported compliance with animal welfare regulations and statement of potential conflicts of interest are more frequently reported.* *64*
Table 3.7. Summary to date of the identification and validation of VGF as a biomarker for neuropathic pain*. The identification and correlation to disease and pharmacodynamics have only be confirmed in animal models.* *74*
Table 4.1. *Glossary of terms used in the systematic review and meta-analysis of cannabinoids, cannabis-based medicines and endocannabinoid system modulators assessed for antinociceptive efficacy in animal models of pathological or injury-related persistent pain (Soliman et al., 2021).* *81*
Table 4.2. *Differences and overlap between treatment and control groups for drug classes assessed for antinociceptive efficacy in rats (A) and mice (B) which corresponds to the forest plots of figures 3 and 4 in the publication, figure 5.4 in the thesis. N, number of animals (Magnusson, 2020, Soliman et al., 2021).* *84*
Table 5.1. *Summary of the drug classes assessed for anti-nociceptive effect in animal models of injury-related or pathological persistent pain. FAAH, fatty acid amide hydrolase, PPAR, peroxisome proliferator-activated receptor, THC, delta-9 tetrahydrocannabinol, CBD, cannabidiol, FABP, fatty acid-binding protein, NAAA, N-acylethanolamine-hydrolysing acid amidase, MAGL, monoacylglycerol lipase, PAM, positive allosteric modulator, TRPV1, transient receptor potential vanilloid receptor 1, TRPA1, transient receptor potential ankyrin 1, ABHD6, abhydrolase domain containing 6.* *114*
Table 5.2. *Summary of the model types used to assess the anti-nociceptive effect of cannabinoids, cannabis-based medicines, and endocannabinoid system modulators.* *115*

Table 5.3. *The number of studies that conducted further experimentation to gain further understanding of the cannabinoids, cannabis-based medicines, and endocannabinoid system modulators in conjunction with anti-nociceptive effect.* *116*
Table 5.4. *Drug class effects in rat neuropathic pain associated models. Rows highlighted blue show effect, highlighted orange no effect. 'x' denotes was not tested. N comparisons; number of comparisons, SMD; standardised mean difference, SE; standard error, CI; confidence interval.* *125*
Table 5.5. *Drug class effects in rat inflammation models. Rows highlighted blue show effect, highlighted orange no effect. 'x' denotes was not tested. N comparisons; number of comparisons, SMD; standardised mean difference, SE; standard error, CI; confidence interval.* *126*
Table 5.6. *Drug class effects in mouse neuropathic pain-associated models. Rows highlighted blue show effect, highlighted orange no effect. 'x' denotes was not tested. N comparisons; number of comparisons, SMD; standardised mean difference, SE; standard error, CI; confidence interval.* *128*
Table 5.7. *Drug class effects in mouse inflammation models. Rows highlighted blue show effect, highlighted orange no effect. 'x' denotes was not tested. N comparisons; number of comparisons, SMD; standardised mean difference, SE; standard error, CI; confidence interval.* *130*
Table 5.8. Drug class effects assessed by mechanical limb withdrawal in rats*. (A) most frequently assessed other model types (B) most frequently assessed neuropathic pain associated models. Rows highlighted blue show effect, highlighted orange no effect. 'x' denotes was not tested. N comparisons; number of comparisons, SMD; standardised mean difference, SE; standard error, CI; confidence interval.* *132*
Table 5.9. Drug class effects assessed by mechanical limb withdrawal in mice*. (A) most frequently assessed other model types (B) most frequently assessed neuropathic pain associated models. Rows highlighted blue show effect, highlighted orange no effect. 'x' denotes was not tested. N comparisons; number of comparisons, SMD; standardised mean difference, SE; standard error, CI; confidence interval.* *133*
Table 5.10. *The rat coefficient estimates from multivariable meta-regression of the variables drug class, model type, outcome assessment, animal strain and sex. The redundant predictors were removed from the model. Significance levels; *** P< 0.001, ** P< 0.01, * P< 0.05.* *135*
Table 5.11. Multicollinearity in the regression model*. Rows highlighted amber represent VIF values between 1 and 5 and indicate moderate correlation. Rows highlighted red represent VIF values above 5 and indicate severe correlation. The redundant predictors were removed from the model.* *137*
Table 5.12. *The rat coefficient estimates from the multi-model inference meta-regression of the variables drug class, model type, outcome assessment, animal strain and sex. Significance levels; *** P< 0.001, ** P< 0.01, * P< 0.05.* *139*
Table 5.13. *The mouse coefficient estimates from multivariable meta-regression of the variables drug class, model type, outcome assessment, animal strain and sex. The redundant predictors were removed from the model.* *142*
Table 5.14. *Multicollinearity in the mouse multivariable regression model. Rows highlighted amber represent VIF values between 1 and 5 and indicate moderate correlation. Rows highlighted red represent VIF values above 5 and indicate severe correlation. The redundant predictors were removed from the model.* *145*
Table 5.15. *The mouse coefficient estimates from the multi-model inference meta-regression of the variables drug class, model type, outcome assessment, animal strain and sex. Significance levels; *** P< 0.001, ** P< 0.01, * P< 0.05.* *148*
Table 5.16*. Research priorities for assessing anti-nociceptive efficacy of cannabis-based medicines, cannabinoids, and endocannabinoid system modulators in animal models of persistent pain.* *165*
Table 5.17*. Research priorities relating to the improved rigor and transparency of design, conduct and reporting of preclinical trials assessing anti-nociceptive efficacy of cannabis-based medicines, cannabinoids, and endocannabinoid system modulators in animal models of persistent pain.* *165*
Table 6.1. *The number of crowd scientists that enrolled for training, completed the training, and contributed to the project at each stage of the review process.* *172*
Table 6.2. *Time appreciation of both the screening and data extraction phases of the review. Includes comparison of the real time data of the crowd compared to theoretical timeframe of 2 specialist reviewers.* *173*
Table 6.3. *Crowd scientists' inter-rater reliability calculated using Cohen's Kappa and % agreement.* *174*
Table 6.4. *Estimated cost of crowd scientist labour based upon the hourly rate of a student demonstrator at Imperial College London.* *175*

***Table 6.5**. Free text feedback from crowd scientists relating to their experience contributing to the project. . 179*
***Table 6.6.** Crowd scientists' feedback on the usability of the SyRF application. 180*
***Table 6.7.** Free text feedback from crowd scientists relating to their experience using SyRF. 181*
***Table 6.8.** Guiding Principles for Crowd Science. 189*
***Table 7.1**. Summary of ML methods that can be employed to reduce the screening burden. 197*
***Table 7.2.** The difference between the number of included and excluded studies between human and machine decisions. 0; Exclude, 1; Include. Green denotes agreement, red disagreement. 202*
***Table 7.3. ML algorithm performance for iterative error analysis.** The number of differences represents the number of studies for which the human and machine decisions differ in either direction. The next two columns represent the number of studies that were changed from exclude to include and include to exclude. The final two columns represent the total number of records excluded and included after each round. 203*
***Table 8.1.** Frequency of graph types presenting in vivo experimental data. 215*
***Table 9.1.** Comparison highlighting the differences between the two reviews undertaken during this PhD. ... 229*
***Table 9.2.** Comparison of the prevalence of reporting of methodological quality criteria in the VGF and Cannabis & Pain reviews. Additional comparison with another preclinical systematic review from the pain research field. 231*
***Table 9.3.** Utility of preclinical systematic reviews and meta-analysis based upon the Cannabis & Pain review. 240*

1. Introduction

This PhD aimed to address two inter-related problems: first, the failure to translate preclinical success to the clinic has led researchers to question the predictive validity of animal models. Secondly, that preclinical systematic reviews are vital research tools that aim to answer a range of research questions to improve research value and translational ability of results. However, there are several challenges which limit their feasibility including the exponentially increasing, large volume of literature and resource intensive methods. Bringing the two together, we need feasible, accurate and efficient methods for the timely synthesis of evidence so that we can make evidence-based decisions to address our research challenges such as improving predictive validity of animal models. Here, I first discuss the challenges of translational pain research and then introduce the challenges associated with preclinical systematic reviews. The aims of this PhD therefore were to assess methods and automation technologies to improve the feasibility, accuracy and efficiency of preclinical systematic reviews while answering questions of neurobiological importance.

"Pain is an unpleasant sensory and emotional experience associated with, or resembling that associated with, actual or potential damage" (Raja et al., 2020). Pain can serve an adaptive role, however for individuals living with chronic pain it can have a limiting and detrimental effect on function, social and psychological wellbeing. There are various estimates on the prevalence of chronic pain; Globally, estimates suggest that 20 % of adults suffer from chronic pain and 10 % of the adult population are newly diagnosed each year (Goldberg and McGee, 2011). A systematic review and meta-analysis of population studies in the UK estimates that chronic pain affects between one-third and one-half (~ 28 million) of the population (Fayaz et al., 2016). Similarly in the United States a Centers for Disease Control and Prevention analysis of 2016 National Health Survey data estimates that 20.4 % (50 million) of U.S. adults have chronic pain (Dahlhamer et al., 2018). It diminishes quality of life and is the leading cause of disability worldwide (Rice et al., 2016). Current analgesics often lack mechanistic rationales and are characterised by limited efficacy and unwanted side effects (Finnerup et al., 2015, Aldington and Eccleston, 2019), thus there is an urgent unmet medical need for improved treatment.

1.1. Challenges of Translational Pain Research

Animal experiments have contributed considerably to our understanding of human health and disease. They are often used to identify and validate potential targets and investigate the safety and/or efficacy of interventions intended for human use. Data from preclinical trials are also used to justify assessment of a candidate drug in clinical trials. Despite recent collaborative efforts between academic institutions and industry to develop better analgesics, only a few novel candidates have

shown promise in animal models and these have failed to demonstrate efficacy in clinical trials (reviewed by Borsook et al. (2014), Mao (2012), and Attal and Bouhassira (2019)). There are many possible reasons for the disparity between animal model and clinical trial results including the complexity and inadequacies of clinical trials. However, the failure to translate preclinical success to the clinic coupled with the multifaceted complexity of pain has led researchers to question the predictive validity of animal models (du Sert and Rice, 2014, Berge, 2011, Yezierski and Hansson, 2018).

1.2. Limitations of Preclinical Study Design

Pain cannot be directly measured in animal models; researchers rely on surrogate behaviours. An animal model consists of the subject i.e., the species, strain, sex, age of the animal, the assay i.e., the disease or injury modelled and the measure i.e., type of pain-associated behavioural assessment. How closely an animal model resembles the human condition (construct validity) is important and there are several study characteristics that can impact pain-related behavioural outcome measures and the assessment of novel analgesic efficacy. Careful consideration should be given to the choice of species, strain, sex, and age in relation to the clinical condition being modelled, as results and conclusions will be dependent on these variables. Common causes of reduced external validity are highlighted below:

(i) **Induction of disease in otherwise healthy animals.** Disease or injury is frequently induced in young, healthy animals. This contrasts with the clinical situation in which disease or injury predominantly occurs in the elderly who have co-morbidities.

(ii) **The effect of treatment is assessed in a homogenous group of animals versus heterogenous patient groups.** Animal cohorts are genetically very similar compared to the widely heterogenous clinical population. Within the animal population, strain (Mogil et al., 2005), age (Pickering et al., 2006) and sex (Craft et al., 2004, Dahan et al., 2008, Mogil and Chanda, 2005) have been shown to impact the development and sensitivity to pain. In general, female rodents are reported to be more sensitive than males (Mogil et al., 2000). Similarly in patients, both sex (Mogil, 2012b), and ethnicity (Alabas et al., 2013) can affect pain tolerances and response to analgesics can differ between the sexes (Niesters et al., 2010). To add, recent pain related systematic reviews have shown that researchers predominantly use male animals (Currie et al., 2019, Federico et al., 2020, Soliman et al., 2021) which is worrisome because sex related differences are known and there is a higher prevalence of chronic pain conditions in women (Mogil, 2012b).

(iii) **Insufficient similarity to the human condition**. Animal models do not effectively simulate multidimensional clinical pain conditions including the complex psychological component. They also do not mimic the slow progressive and degenerative nature of many human chronic diseases, nor do they replicate the complexity of co-morbidity or polypharmacy. For example, neuropathic pain associated animal models' behavioural responses are characterised by sensory gain. Whereas patients with neuropathic pain are more frequently characterised by sensory loss and spontaneous pain, sensory gain is less frequent (Maier et al., 2010, Vollert et al., 2017).

(iv) **Differences in outcome measures and the timing of outcome assessment.** The presence or absence of pain in animal models cannot be directly measured and can only be inferred from the observation of surrogate behaviours (du Sert and Rice, 2014). Reflex withdrawal responses are the most commonly assessed outcome in preclinical studies (Mogil and Crager, 2004); a measure of hypersensitivity which are appropriate for some pain types e.g. inflammation, but not the spontaneous pain, or pain as a result of sensory loss. Lastly, the duration of animal studies is usually brief (up to a few weeks) which does not adequately reproduce the impact of prolonged nociception on clinical pain.

(v) **Unrealistic treatment strategy for translation to the clinic.** Animal models can lack applicability to a clinical setting. For example, drugs are often tested prophylactically or in early stages of disease onset which does not represent the clinical situation where the drugs are often required in the late stages of disease.

1.2.1. Experimental and Reporting Bias

The theory that human physiology and disease are not sufficiently exemplified by animal models is one reason for failure to predict efficacy of potential therapeutics (van der Staay et al., 2009). However, another reason for the attrition and failed translation is that the animal literature is affected by biases in the design, conduct, analysis, and reporting of experiments. Poor experimental design and lack of transparent reporting in many preclinical biological fields including pain can lead to producing results that are not reproducible (van der Worp et al., 2010, Rice et al., 2008). Researchers have postulated that inadequate study design (Andrews et al., 2016b), lack of scientific rigour (Currie et al., 2019, Hirst et al., 2014, Crossley et al., 2008, Rooke et al., 2011), and transparent reporting (Rice et al., 2008, Clark, 2016) of preclinical studies may be responsible for the translational failures.

1.3. Waste in Research and Systematic Reviews

Chalmers and Glasziou (2009) estimated that 85% of biomedical research funding was being avoidably wasted at each stage of the process (Figure 1.1.).

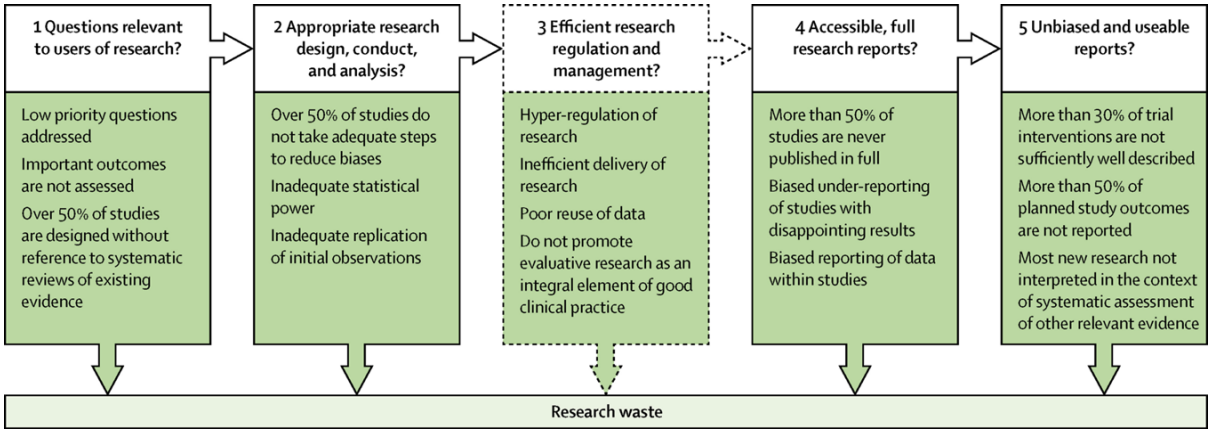


Figure 1.1. The 5 Pillars of Research Waste (Moher et al., 2016). Evidence confirms imbalanced research question selection (Pillar 1), poor study design and execution (Pillar 2), poor regulation and evaluation (Pillar 3), non-publication/publication bias (Pillar 4) and poor reporting (Pillar 5) as sources of avoidable research waste.

Several studies show that researchers, research funders, regulators, sponsors, and publishers of research fail to systematically identify and use existing research when preparing to start, fund, regulate, sponsor, or publish the results of new studies (Robinson and Goodman, 2011, Clarke, 2004, Chalmers, 2005); failure to meet requirement of pillar 1, 'asking questions relevant to users of research.' Evidence synthesis is critical to address translational challenges and reduce research waste. Empirical evidence from systematic reviews and meta-analyses can inform experimental design refinement and improve methodological quality thereby reducing waste and increasing the value of preclinical research.

Systematic reviews are vital research tools, important for evidence-based research planning and decision making. A systematic review identifies and summarises all available and relevant studies to answer a specific research question. They are an effective method for the consolidation of the high volume, rapidly accruing and often conflicting research on a specific topic. Designed to address a specific research question they use predefined methods to identify, select and critically appraise all available and relevant research evidence to answer the question in an unbiased manner (Egger et al., 2001). Where appropriate, a meta-analysis can be conducted whereby the quantitative data is extracted from the relevant studies and statistical techniques are used to integrate the individual studies' quantitative data thus providing an objective sense of the metadata. By combining the data from individual studies through a meta-analysis the power and precision of the results can be increased (Noordzij et al., 2009).

To date (March 2021), a PubMed search ("Pain" and "Preclinical systematic review" with non-human animal filter on) shows only twelve preclinical systematic reviews (Fernandes et al., 2016, Brito et al.,

2015, Guo et al., 2019, Tetreault et al., 2020, Farrell et al., 2014, Mohammad et al., 2016, Suokas et al., 2014, Federico et al., 2020, Hooijmans et al., 2015, Currie et al., 2019, Siqueira-Lima et al., 2017, Soliman et al., 2021) have been conducted in the pain field. Given the breadth of pain aetiologies and translational failures it is critical we evaluate and assess the evidence-base in a robust and systematic manner. To do so, it is imperative to understand what the barriers are to performing systematic reviews of the preclinical evidence and devise methods and technologies to overcome the challenges.

1.4. Importance of Preclinical Systematic Reviews

A systematic review and meta-analysis can assess the quality or rigour of experimental design, conduct, analysis and reporting as well as the statistical reliability of all available and relevant data, they provide more than a simple summary of the information as originally presented (Sena et al., 2014). Over the past 30 years, The Cochrane Collaboration and others have developed systematic review processes for research evidence synthesis. The well-established framework for evidence-based healthcare to guide clinical practice and healthcare policy has fundamentally revolutionised clinical medicine at all levels. It should be noted that for research to be informative the reporting of research should be of a high standard and complete (Moher et al., 2016). The value of a systematic review is inherently reliant upon complete reporting thereby ensuring that prior research can be robustly assessed and future direction reliably informed (Landis et al., 2012).

To address translational challenges, preclinical systematic reviews offer a framework by which the internal and external validity can be assessed, knowledge gaps can be identified, publication bias determined, discrepancies between preclinical and clinical trial design addressed, inform 3R (replacement, reduction, and refinement of animal use in research) decisions as well as generate new hypotheses for future research:

1.4.1. Internal Validity

It has been well documented that lack of scientific rigour and transparent reporting across many preclinical fields of biological research may also be responsible for the lack of translation (Andrews et al., 2016a). Systematic reviews can demonstrate the importance of maximising the internal validity of animal experiments. Within a systematic review, quality assessments give structured insight into whether the included studies' conclusions are at risk of bias. New hypotheses must be based on reliable data that has been accurately reported. Internal validity ensures that changes observed in outcomes are due to an induced change in the independent variable(s) rather than confounding factors (Higgins et al., 2011). The internal validity may be compromised by a range of biases: systematic differences in characteristics between study groups at baseline (selection bias), how groups are handled in terms of the care for the animals or the administration of the intervention

(performance bias), in the way outcomes are measured and interpreted (detection bias) and how animal exclusions/loss and deviations from protocol are handled (attrition bias). Several preclinical systematic reviews and meta-analyses have provided empirical evidence highlighting that inadequate methodological approaches are associated with bias (Hirst et al., 2014, Crossley et al., 2008, Rooke et al., 2011). Of concern, is that the low prevalence of the reporting of measures to mitigate bias tend to be associated with higher estimates of treatment effects (Crossley et al., 2008, Hirst et al., 2014). Thus, there is a critical need for transparency of reporting all experimental details so that the quality of research can be assessed (Sena and Macleod, 2007).

1.4.2. External validity

Systematic reviews can also be used to support evidence-based choice of study design characteristics including the advantages and disadvantages of specific animal models. They can provide evidence as to which model is most suitable for a new experiment or may show that available animal models are inadequate and there is a need for development. Results and conclusions will be dependent on the choice of species, strain, sex and age, therefore researchers must exercise care when selecting animals for their experiments and relate them to the clinical condition being modelled (e.g. (Wallace et al., 2007b, Mogil and Chanda, 2005, van der Graaf et al., 2013, Pickering et al., 2006)). Teamed with differences in outcome measures and the timing of assessments between animal studies and clinical trials there are many challenges to overcome to produce animal models with high construct validity. It is therefore important to understand how animal models are currently being used to identify ways in which study design can be improved in effort to increase translation and meet patients' needs for safe and effective treatments.

1.4.3. Evidence-based decisions for clinical trials

Systematic reviews can be used to support evidence-based translation, establishing whether there is enough evidence of necessary quality to justify a clinical trial and inform aspects of clinical trial design important to translation (Sena and Macleod, 2007, Hooijmans et al., 2012, Wever et al., 2012). Retrospective preclinical systematic reviews for interventions that failed in clinical trial have demonstrated that prospective systematic reviews of the animal literature would have concluded that efficacy is not evident and therefore a clinical trial would not have proceeded (reviewed Pound and Ritskes-Hoitinga (2020); thereby preventing harm to trial participants and reducing research waste. How the decision is made to take a drug to clinical trial is not clear, and the pipeline is often not linear. For prospective preclinical systematic reviews to reliably inform clinical trial decision making, the design, conduct, analysis, and reporting of animal studies needs to improve in conjunction with the feasibility of preclinical systematic reviews.

1.4.4. Publication Bias

The ability of systematic reviews to help improve the translation of animal studies is likely to be undermined by the presence of publication bias. Publication bias is the phenomenon that studies reporting statistically significant data are more likely to be published than studies reporting neutral or negative data (Song et al., 2010). Therefore, the publicly available research does not represent the total quantity of research conducted. Meta-analyses have suggested that publication bias can lead to an overestimation of treatment effects (Sena et al., 2010) therefore the failure to publish neutral or negative data presents ethical concerns (ter Riet et al., 2012).

1.4.5. Replacement, Refinement and Reduction of Animals in Research

Systematic reviews also offer the means to support scientific and technological developments that replace, reduce, or refine (3Rs) the use of animals in research. Systematic reviews alone are unlikely to remove the need to use animals, however, they may prevent the unnecessary duplication of experiments, make use of the data that is already available without conducting new experiments, or show that proposed experiments would not substantially add to the current knowledge. Conversely, systematic reviews can identify gaps in the knowledge base and generate hypotheses for future research. Systematic reviews also contribute indirectly by supporting the development and validation of replacement alternatives e.g. a systematic review conducted by de Vries et al. (2015b) provides a comprehensive overview of the full range of possibilities of using tissue engineered constructs as a replacement for laboratory animals.

Systematic reviews have made a major contribution to the growing body of evidence that animal experiments of poor design and at risk of bias will generate unreliable data and therefore are unlikely to lead to clinical benefit. The publication of preclinical systematic reviews has initiated a cultural change toward improving internal and external validity within animal experiments thereby improving their predictive value and reducing research waste.

1.5. The Practical Challenges of Systematic Reviews

The methods for the conduct of systematic reviews are well developed (Figure 1.2.), but by design they are resource intensive (Tricco et al., 2008) and the time they take to complete may weaken their usefulness (Shojania et al., 2007). To improve the feasibility of preclinical systematic reviews, methods and technologies that are easy to disseminate broadly and simple to use have been developed. The use of bespoke online software for conducting systematic reviews is standard practice (e.g. Covidence, DistillerSR, EPPI-Reviewer) and a comprehensive comparison of available systematic review software is provided by Bahor et al. (2021). There are two leading groups addressing the challenges and complexities of research involving animals who have adapted the systematic reviews and meta-analysis techniques for use in preclinical research: CAMARDES (Collaborative Approach to Meta

Analysis and Review of Animal Data from Experimental Studies) at the University of Edinburgh, UK and SYRCLE, (Systematic Review Center for Laboratory Animal Experimentation) Radboud University, NL.

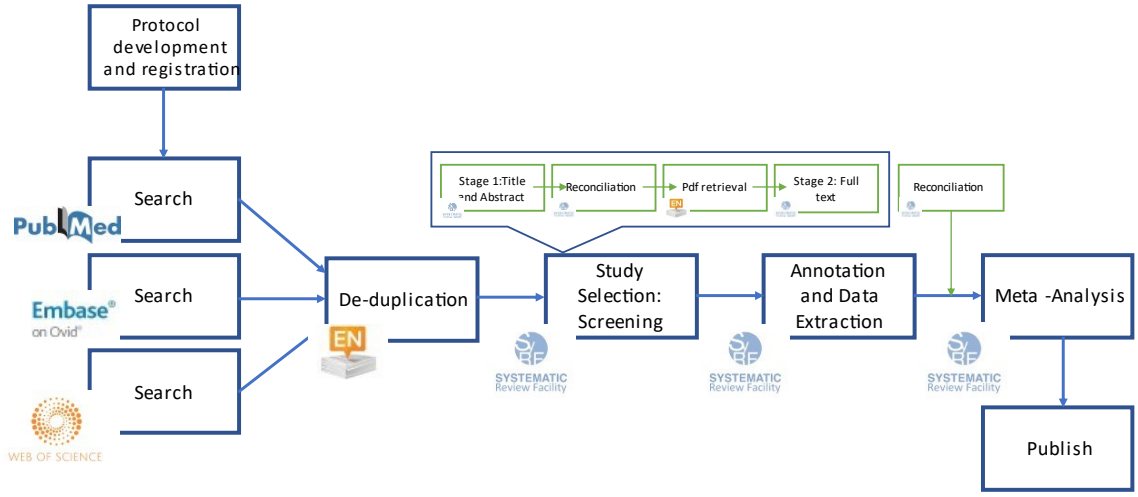


Figure 1.2. Example of the Systematic Review workflow. The platforms that can be used at each stage are also represented (Soliman et al., 2020).

CAMARADES have developed the Systematic Review Facility (SyRF;(http://syrf.org.uk/); a fully integrated online platform for performing systematic reviews of preclinical studies. SyRF is a free resource that can be accessed globally with an email address and internet connectivity. SyRF includes a secure screening database, data repository and analysis application. The application allows for multiple reviewers to contribute to the project simultaneously and is flexible enough to support the width and breadth of preclinical systematic reviews; it allows researchers to build annotation forms and extract information form a wide variety of experimental designs (Bahor et al., 2021). There are also educational resources on how to conduct and report a systematic review. In addition, they have developed Learn to SyRF; an online platform which researchers can use to create project specific training modules therefore enabling reviewers to learn, practice and demonstrate their reviewing skills prior to contributing to a review. CAMARADES and SYRCLE also offer guidance and support to researchers.

There remain, however, several challenges that are greater and unique to the preclinical domain in comparison to the clinical. The preclinical domain is characterised by high-volume and rapidly accruing data from studies with large variation in the design, conduct, analysis, and reporting. In addition, preclinical systematic reviewers can be interested in extracting a more heterogeneous breadth of data e.g., behavioural outcomes, electrophysiological recordings, protein expression levels etc. to address a range of research questions. The problems are being exacerbated by the

exponentially increasing number of publications in databases (Bastian et al., 2010). There are over 30 million articles in PubMed. A search of the MeSH term 'pain' for the year 2019 returned 12,138 articles which equals 33 articles a day, over 1 article an hour (Figure 1.3). In neuropathic pain research, the number of articles retrieved by a systematic search rose from 4568 in 2012 to 8029 in 2015 (Currie et al., 2019). In contrast, only 129 papers were identified for inclusion in a clinical neuropathic pain systematic review (Finnerup et al., 2015).

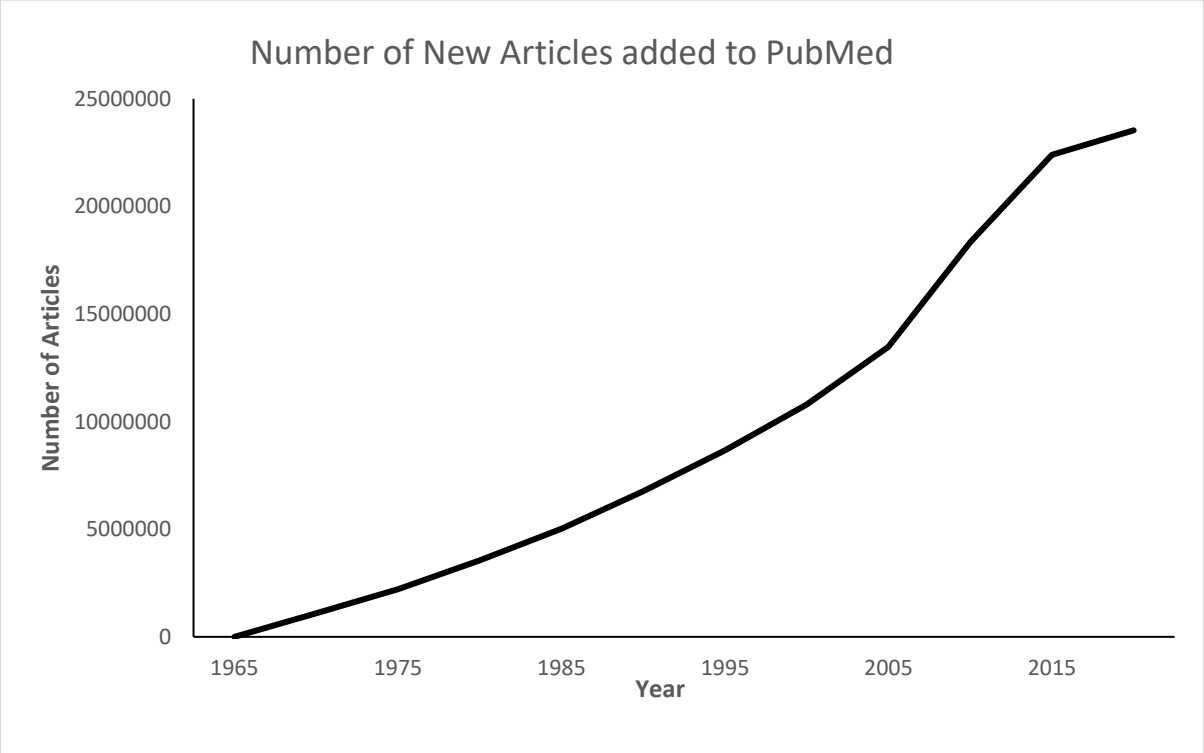


Figure 1.3. **Number of Pain Articles Published Yearly.** Graph depicting the increasing number of pain articles published each year. PubMed search was conducted using the MeSH term 'Pain' up to the year 2019.

The challenges of conducting a systematic review are likely exacerbated by the short time frames available within research grant cycles and limited resources. Systematic reviews are traditionally conducted by a small specialist systematic review team, often independent of primary researchers. This has two implications; firstly, a large workload shared across a small team increases the time that it takes to conduct a review. Secondly, it is necessary to close the circle and engage primary researchers so that their subject matter expertise can be garnered, and the potential of systematic reviews can be realised. Primary research should be informed by systematic reviews; research should be designed, conducted, analysed, and reported in such ways that it can then be used in meta-research thereby adding to an evidence base which can be built upon.

When assessing what is already known in a field, justifying and designing experiments, and making evidence-based funding and policy decisions, it is essential that scientists, funders and policy makers can assimilate this volume of rapidly accruing data in a comprehensive and unbiased manner.

1.6. Improving feasibility

Depending on the research question(s) and the size of the review there are several rate limiting stages e.g., study selection, full text retrieval, annotation and data extraction, and despite the increased ease provided by systematic review platforms, the stages still require considerable labour resource to complete. The need in both the clinical and preclinical domains for timely evidence synthesis is driving development of automation technologies to improve the feasibility, accuracy, and efficiency of these systematic review stages (Figure 1.4). The use of text mining tools and machine learning algorithms are becoming an increasingly popular approach to reduce human effort, cost, and time to complete (Howard et al., 2016b, Tsafnat et al., 2014, O'Mara-Eves et al., 2015, Cramond et al., 2019, Bannach-Brown et al., 2018).

1.6.1. Automation Technologies

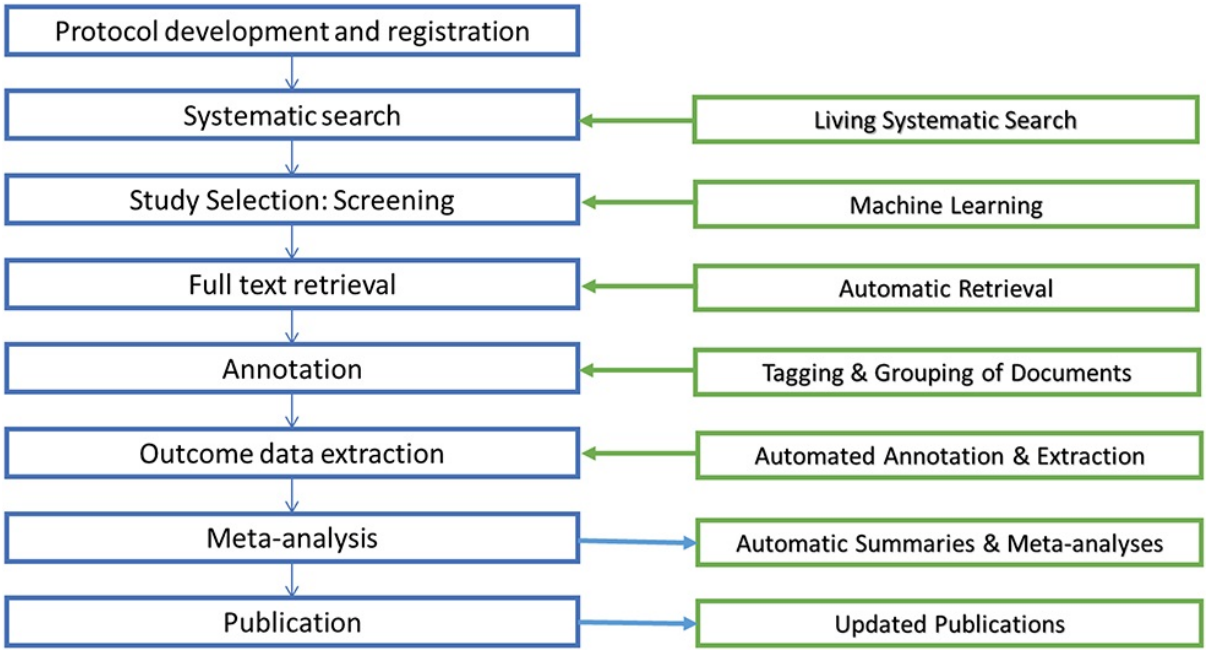


Figure 1.4. Automation technologies are being developed for the different stages of the review process. Machine learning and text mining have improved the feasibility and efficiency of the early stages of the process. Technological developments for the latter stages are in their infancy.

Text mining technologies can facilitate the identification of relevant literature, its categorisation and summarisation. The aim of text mining is to retrieve information from unstructured text and present

it to users in a usable and concise format (Ananiadou et al., 2006). Greater efficiency in the systematic review process is most likely to be found in the identification of relevant studies rather than in their quality assessment and synthesis as such, the use of text mining to date has been focused on the early stages of the systematic review, the identification of relevant studies and the initial descriptive stage. Text mining comprises of three activities; information retrieval, information extraction and data mining, activities that map well into the processes of systematic reviews; the retrieval of relevant studies, describing the study characteristics, identifying patterns within the studies. Text mining offers the possibility of extracting precise data from a retrieved document set, and of finding useful associations among disparate data, leading to the discovery of new knowledge.

Machine learning is the science of computers performing tasks without being explicitly programmed to do so. Systems can learn from data, identify patterns, and make decisions with minimal human intervention. It is in this context that machine learning algorithms can be employed to learn how to categorise studies and any systematic reviews can also take advantage of generic classifiers and review specific classifiers (O'Mara-Eves et al., 2015). Both generic and review specific classifiers are developed by humans and machines operating in mutually beneficial ways. Both classifiers are trained from "gold standard" classifications which are usually generated by human experts. It is possible to build a model that directly predicts the probability of the study's eligibility and relevance to the review however, lack of training data can be a limitation. In addition, it is difficult to assess reliability because the machine will not have had the benefit of 'seeing' the full variety of relevant studies. There are techniques for using machine learning in new reviews where humans and computers operate in a workflow which prioritises the relevant studies for manual screening while simultaneously improving the accuracy of the machine model (Wallace et al., 2010). Generic classifiers are highly accurate with respect to the task they have been trained to perform but they will not be able to classify studies in a way that directly aligns with the specific inclusion criteria defined by individual reviews. By contrast, review specific classifiers will tend to be highly attuned to the context of the criteria for review, but may be less accurate due to the limited extensiveness of training material (Thomas et al., 2017).

The development and use of automation technologies in preclinical systematic reviews have been reviewed extensively by Bannach-Brown et al. (2021) and are summarised in Table 1.1. It is due to these advances that large preclinical systematic reviews are now deemed feasible however challenges remain, notwithstanding the high tempo of evidence production and urgent need for further automation assistance, the challenges include, the evaluation, integration and validation of these tools, the maintenance and future development of automation technologies, and funding to support community expansion and continued collaborative development (Bannach-Brown et al., 2021).

Table 1.1. Automation of the Stages of Systematic Reviews.

Systematic Review Stage	Tool/Software
Search: automation citation retrieval and de-duplication	Application Programming Interface (API) allows two applications to talk to each other. APIs are available bibliographic databases including PubMed, Web of Science (Baker, 2018), biorxiv and medrxiv (McGuinness et al., 2021). There are several tools available for the identification and removal of duplicates that are retrieved due to database cross-over including within reference managers (Endnote and Zetoro) and as stand-alone e.g. Automated Systematic Search deduplicator (ASySD)(Hair et al., 2021).
Study Selection: screening and categorisation	Machine learning for citation screening and regular expression and other text-mining techniques for categorisation of articles.
Full Text Retrieval	Reference manager software e.g., Endnote or Zetoro.
Annotation e.g., Risk of Bias	The use of text mining to automate the identification of blinding, randomisation, and sample size calculation (Bahor et al., 2017).
Outcome Data Extraction	Graph2Data, a tool developed by Eppi-Centre and CAMARADES in collaboration on the SLIM (Systematic Living Information Machine) grant, which streamlines and semi-automates axes identification and data point identification, reducing the time required for data extraction, as well as reducing extraction and transcription errors associated with manual methods.

This presents an opportunity to take advantage of emerging technologies for the development of living systematic reviews, systematic reviews that are continually updated, incorporating relevant new evidence as it becomes available (Elliott et al., 2017) through fully automated workflows. Living systematic reviews will require automation tools that are reliable and require minimal manual labour at each stage. Several machine learning approaches have been evaluated in the context of systematic reviews of several medical problems (Cohen et al., 2012, Cohen et al., 2006, Howard et al., 2016a). In the preclinical domain there have fewer attempts to apply such techniques (Bannach-Brown et al., 2018, Howard et al., 2016b) and evaluation is ongoing (Bannach-Brown et al., 2021). It is not clear which technologies and strategies work best and is therefore imperative that the performance of these tools are evaluated and their utility assessed against specific user cases. Evaluation will provide reviewers with the most efficient and effective tools for their needs and will also provide insight for the continued development of such tools.

1.6.2. Crowd Science

Crowd science is a form of distributed collaboration, the principles of which are closely related to citizen science. Citizen science involves the public collaborating with professional scientists on

research projects. Citizen science has led to several benefits and similar benefits including improving the feasibility of systematic reviews, may be realised by employing a crowd of scientists. Citizen science has enabled research that would not otherwise be possible (Cooper et al., 2010, Kawrykow et al., 2012), has led to reduction of time and cost of research projects (Sauermann and Franzoni 2015), and the engagement of a crowd has improved diversity both in terms of the demographics of the crowd and specialist skills (Jeppesn and Lakhani 2010, Pimm et al 2014). In particular, online citizen science projects have demonstrated that they are an effective form of public engagement, providing volunteers with an opportunity to learn (Masters et al., 2016) and with an easily accessible means to make an authentic contribution to a research project, which can in turn improve public engagement and the advocacy of scientific findings (Forrester et al., 2017, Straub, 2016). Therefore, the instigation of a preclinical systematic review to be conducted using crowd science can create similar opportunities but within a research field.

Preclinical systematic reviews are resource intensive; the study selection, annotation and data extraction stages are time consuming and error prone and require considerable manpower to be completed within a timely manner. CAMARADES have previously used crowdsourcing; they recruited scientists external to their research group as additional reviewers and provided training prior to their contribution to the project (Hair et al., 2019). This allowed for the considerable workload to be spread across more people, thereby reducing the individual burden. This approach can be scaled thereby improving the feasibility of a project of any size. In addition, this approach offers scientists the opportunity to contribute to a research project, engage deeply with the literature thereby informing their own research practises and experiencing first-hand the utility of preclinical systematic reviews.

1.7. Thesis Overview

Preclinical systematic reviews have an important role to play in understanding animal research and the limited translation from animal trials to clinical research and practise. However, the feasibility of conducting a preclinical systematic review may be limiting their uptake. There is an increasing breadth of pain research being conducted. However, if research is not suitably guided by theories of aetiology, pathogenesis, and human experience of pain it may lead to further translational gaps. It is impossible for any consumer of research to assess the contribution of a new study unless its results are set systematically in the context of a review (Clarke et al., 2002). It is incumbent upon us to develop a system of preclinical research that is continuously informed and builds upon existing knowledge. This will allow for the identification of the most relevant models to the different human pain conditions so that the most promising therapeutic candidates can be reliably identified. By answering key questions of neurobiological interest, I present here methodological and technological approaches for improving feasibility, efficiency, and accuracy of preclinical evidence synthesis.

This thesis reports:

How narrative reviews can be informed by systematic review processes to improve their utility and reduce bias (Chapter 2). Using the example of two published reviews, 'VGF: a biomarker and potential target for the treatment of neuropathic pain?' (Soliman et al., 2019b) and 'Cannabinoids, the endocannabinoid system, and pain: a review of preclinical studies (Finn et al., 2021) I explore the strengths, limitations and appropriate uses of the two review types. In Chapter 3, I present the findings of the VGF review, and discuss the importance of pain related biomarkers for prognostics, nociception, prevention, and diagnosis, in addition to the need to be able to synthesise emerging evidence in an efficient manner, concurrent to its production to produce robust, contextualised understanding upon which we can build.

Chapters 4 and 5 report the systematic and meta-analysis of cannabis-based medicines, cannabinoids and endocannabinoid system modulators tested for antinociceptive efficacy in animal models of persistent pain (Soliman et al., 2021), referred to as the Cannabis & Pain review. The review provided the bases for exploring and reflecting upon the utility of different systematic review and meta-analytical approaches used in preclinical evidence synthesis. The review also provided the opportunity to assess methodological and technological approaches for improving the feasibility, accuracy, and efficiency of systematic reviews, the methods, and results for each are discussed in subsequent chapters.

Chapter 6 reports the use of crowd science for study selection, annotation, and data extraction. Crowd performance in the Cannabis & Pain review was assessed to determine the potential for efficiency savings; in terms of increasing the speed and accuracy of a review and reducing the cost.

Chapter 7 reports the use of machine learning for study selection. The chapter discusses the opportunities, strength, and limitations of machine learning algorithms for study selection and classification to reduce resource requirements and/or increase capacity of reviews.

Chapter 8 reports the continued development of a semi-automated outcome data extraction tool, Graph2Data (Cramond et al., 2019) for the abstraction of data presented in graphs.

2. Evidence Synthesis: Narrative vs Systematic Reviews

2.1. Introduction

Reviews and evidence syntheses are fundamental to good scientific practise. Science should be cumulative and therefore in depth understanding on the current state of knowledge is a requirement prior to designing and conducting further experimental research (Ioannidis et al., 2014, Moher et al., 2016). There are several ways of conducting research synthesis and this chapter aims to discuss and compare the methodologies and their purposes, processes, strengths, and limitations. It includes two published mixed methods reviews as case studies which I have authored (Soliman et al., 2019b, Finn et al., 2021).

Summarising evidence is a challenge across scientific fields, building upon prior knowledge is becoming increasingly complex. Reviews aim to improve understanding of concepts and summarise current knowledge. Publications of preclinical research in many fields often have large variation in how research was designed, conducted, analysed, and reported. Species, models, interventions, and outcome assessments can differ. This makes it challenging to make objective sense of the evidence base particularly if there are conflicting results. Reviewers need to make decisions about how to address the variation and uncertainty and without a structured framework it is unlikely that these decisions will be applied consistently throughout. A reader will be left with an impression of the evidence that they will use to answer a research question of interest. Therefore, it is imperative that reviews are effective and well-conducted to create a firm foundation for advancing knowledge.

There are several different review and evidence synthesis methods available to researchers, with Grant and Booth (2009) suggesting 14 different types which have both distinct and overlapping characteristics for the search, appraisal, synthesis, and analysis (Table 2.1). However, this chapter aims to focus on traditional narrative or literature reviews and systematic reviews contextualised to preclinical literature. The question of review type has been discussed vehemently within clinical research from which this chapter draws upon, but the conversation does not appear have taken hold yet within the preclinical domain.

Table 2.1. Different review types, their aims, and methods (adapted from Grant and Booth (2009).

Review Type	Aims	Methods
Critical review	Extensive literature review with critical evaluation of quality	Identifies most significant literature. Narrative synthesis and seeks to identify conceptual contribution to embody existing or derive new theory
Literature review	Examine recent or current literature	May or may not include a comprehensive search. Narrative synthesis and analysis may be chronological, conceptual, thematic.
Mapping review	Maps and categorises existing literature from which to commission further reviews	Search determined by scope and time. Graphical and tabular characterises quantity and quality of the literature.
Meta-analysis	Statistical combination of quantitative data to provide a more precise effect of the results	Comprehensive search. Quality is assessed and may determine inclusion/exclusion. Graphical and tabular synthesis with narrative commentary.
Mixed methods review	One component is a review of the literature but combines different review approaches	Comprehensive search. Quality is assessed. Components will be presented both narratively and with graphical and tabular synthesis.
Overview	Summary of the literature to describe its characteristics	Depends on whether it is a systematic overview as to whether the search is comprehensive. May include a quality assessment. Narrative synthesis with tabular formats. Analysis may be chronological, thematic, conceptual etc.
Qualitative systematic review	Integration and comparison of qualitative findings from qualitative studies	Selective or purpose sampling. Quality assessed. Qualitative narrative synthesis and thematic analysis.
Rapid review	Assessment of what is known about policy or practice issue	Comprehensive search. Time limited quality assessment. Narrative and tabular synthesis.

Scoping review	Preliminary assessment of potential size and scope of available literature. Aims to identify nature and extent of evidence base	Search determined by time and scope. No formal quality assessment. Tabular synthesis with narrative commentary. Provides current state of knowledge and priorities for future research
State-of-the-art review	Addresses current situation and may offer new perspectives and suggestions for future research	Comprehensive search. No formal quality assessment. Narrative synthesis.
Systematic review	Systematically search, appraise, and synthesise all evidence relevant to a specific research question	Comprehensive search. Quality assessment which may determine inclusion/exclusion. Narrative and tabular synthesis. Presents what is known, uncertainty and gaps in the evidence base, and provides recommendations for future research.
Systematised review	Includes elements of a systematic review provides what is known, uncertainty and methodological limitations	May or may not include a comprehensive search. May or may not assess quality. Narrative synthesis with tabular formats.
Umbrella review	Compiles evidence from multiple reviews into one accessible and useable document.	Identification of component reviews. Quality assessment of the reviews. Graphical and tabular synthesis with narrative commentary. Provides what is known, recommendations for practise, what remains unknown and recommendations for future research.

Narrative reviews are useful for pulling lots of information relating to a particular topic together whereas systematic reviews aim to answer a narrowly defined research question through structured methods (more detailed definitions and descriptions are described below). The two review types serve different purposes. It is therefore pertinent to ask which review methods are most appropriate to each need. Neither a systematic review with their narrow scope nor the individuality of narrative reviews can satisfy the different review needs so this chapter does not intend to argue the hierarchy of evidence synthesis but how the two methods serve different needs, can be used in mutually supportive manners, and demonstrate that methodological quality is critical to both. Lastly, the case will be made that researchers need to be cognisant that review type selection and methodological quality can influence whether we continue research on strong foundations or contribute to corrupting scientific processes.

2.1.1. The Different Roles of Narrative and Systematic Reviews

'Review' is a generic term that can refer to any summary of literature whereas 'synthesis' is the technique for producing knowledge that can be greater than the sum of its parts (Campbell et al., 2003). The Medical Subject Headings (MeSH) note describes a narrative review as "published materials that provide an examination of recent or current literature. Review articles can cover a wide range of subject matter at various levels of completeness and comprehensiveness based on analyses of literature that may include research findings" (MeSH, 2019b). Systematic reviews are described as reviews of primary literature that "attempt to identify, appraise and synthesise all the empirical evidence that meets specified eligibility criteria to answer a given research question. Its conduct uses explicit methods aimed at minimising bias in order to produce more reliable findings regarding the effects of interventions for prevention, treatment and rehabilitation that can be used to inform decision making" (MeSH, 2019a).

Collins and Fauser (2005) argue that narrative reviews provide historical context, background knowledge, evolving concepts and conflict that require the flexibility of choice about the inclusion. This is echoed by Greenhalgh et al. (2018) in which the role of narrative reviews is to provide interpretation and critique to deepen understanding. Whereas systematic reviews address narrowly focussed questions, and their key role is to summarise data. The rigour of a systematic review is needed to objectively determine the effectiveness of an intervention with inferences adhering to the cited evidence and free from opinion. The differences between the aims and features of a systematic and narrative review are summarised in Table 2.2.

Table 2.2. Comparison of Systematic and Narrative Review Aims and Features.

	Systematic Review	**Narrative Review**
Definition	Overview of primary research on a narrowly focused research question that identifies, selects and critically appraisals all relevant literature to answer the research question in an unbiased manner	Qualitatively summarises evidence on a topic using expert opinion and experience to collect and interpret relevant research
Aims	Answer a specific research question in an unbiased manner	Provide a summary or overview of a topic
Key Features	A priori declaration of research question and methods Systematic search and study selection Assessment of internal validity Presentation and inference of results based on the included studies	Introduction Exploration of themes Discussion
Timelines	Months to years	Weeks to months
Value	Evidence-based conclusions	Summary of the literature

2.1.2. Strengths and Limitations of a Narrative and Systematic Review: Methodology, Quality and Reproducibility

Narrative reviews tend to cover a wide range of inter-related topics and often do not have or communicate methodological approaches to mitigating bias. They often lack specific criteria for selecting studies for inclusion and there are rarely any criteria for assessing a study's validity which can only serve to undermine the reliability of the conclusions of the review. Despite guidelines e.g., Green et al. (2006) that encourage authors to share their methods so that risks of bias can be mitigated, the methods are rarely described. Systematic reviews are considered the best way to synthesise evidence. The strength of a systematic review includes a narrow-focused research question, comprehensive, systematic search, a criterion-based selection of all relevant evidence, with rigorous appraisal of validity, and if appropriate a quantitative analysis to provide evidence-based conclusions (Cook et al., 1997). The explicit framework and predefined methodology of a systematic review includes steps to minimise reviewer bias at all stages of the process using objective, reproducible criteria to select relevant literature and assess their validity. In contrast, narrative reviews rarely have or report a predefined methodology or structure, which makes reproducing the

review very difficult if not impossible. If a review cannot be reproduced it is important to question what value it has and how does this relate to efforts to reduce research waste (Moher et al., 2016). Table 2.3 compares the underlying methodology of each review type.

There are many reasons for performing a systematic review including providing empirical evidence for decision-making. If appropriate, a meta-analysis combines the results of single studies to provide a more precise summary of treatment effect (Schlesselman and Collins, 2003). A limitation, however, is that the narrow focus and prescribed methods of the review and prohibit comprehensive coverage of a topic. Systematic reviews have been criticised for being too specific and inflexible (Pawson et al., 2005), however these criticisms have been made in the context of the complexity implementation of health and social care interventions. Therefore, it is important not to generalise to preclinical systematic reviews where the systematic review aims can be different or assessment of intervention efficacy is less complex (due to controlled experimental environments), and the results will not be used to guide complex healthcare policy.

Table 2.3. Comparison of the Underlying Methodology of each Stage of the Systematic and Narrative Review Processes.

	Systematic Review	**Narrative Review**
Research question	Narrowly defined, focused research question or hypothesis to be tested.	May start with a specific research question but can also be more general discussion.
Search for relevant studies	An explicit search strategy is developed. Searches are conducted across several sources e.g., multiple electronic bibliographic databases and can be supported by manual searches to identify all relevant literature. Some reviews will include published, unpublished, and grey literature.	Will not aim to identify all relevant literature. Searches will be conducted for studies that are known to be pivotal to the subject matter experts. The search is not usually reported.
Study selection	Explicit inclusion/exclusion criteria with clearly stated PICO (Population, Intervention, Comparison and Outcome) elements. The process is conducted by 2 reviewers to reduce bias and inaccuracy.	Not usually reported. If reported, it is not conducted systematically.
Assessing study quality	Studies are critically appraised including methodology.	Not usually reported. If reported, it is not conducted systematically.

Synthesising studies	Qualitative and if appropriate quantitative summary; a meta-analysis resulting in a pooled estimate of intervention effectiveness.	Often a qualitative summary.

2.2. Aims

Two narrative reviews were instigated. Each differently aimed to demonstrate how systematic review methods can be employed to strengthen and reduce the bias within a narrative review.

2.3. Methods

The methods of each narrative review supported by systematic review processes are described below. Briefly, the VGF review (Soliman et al., 2019b) was a literature review supported by a systematic search and study selection. Second, the literature review Finn et al. (2021) discusses the biological evidence for the use of cannabinoids, cannabis-based medicines, and endocannabinoid system modulators for pain management and was uniquely supported by a complementary preclinical systematic review (Soliman et al., 2021). Contextually, both reviews are discussed in more detail in subsequent chapters (the VGF review in Chapter 3 and the Cannabis & Pain review in Chapter 5), here the focus is on the methods of the reviews.

2.3.1. 'VGF: a biomarker and potential for the treatment of neuropathic pain?' (Soliman et al., 2019b)

2.3.1.1. Introduction and Aim

VGF-derived peptides have a functional role in several disorders including obesity, dementia, depression, and pain. VGF expression is low under normal physiological conditions and is rapidly upregulated in sensory neurons following nerve injury and inflammation, therefore, may in part be responsible for the onset, development, and maintenance of neuropathic pain. The aim of this review was to identify, select and critically appraise the evidence base to improve the understanding of the molecular mechanisms and signalling pathways of VGF-derived neuropeptides and the physiological effects they mediate.

2.3.1.2. Methods & Results

- *Systematic Search*

To identify all potentially relevant publications for the review a PubMed search using the term "VGF" was instigated on the CAMARADES Systematic Review and Meta-Analysis Facility (SyRF). The search identified 424 potentially relevant articles.

- *Study Selection*

Title abstract and full text screening was performed to identify studies in which the role of VGF was investigated in pain pathways. Screening was conducted by only one reviewer (NS) and resulted in the inclusion of 14 studies.

- *Data Extraction*

Study characteristics were collected as well as reporting of measures to reduce the risk of methodological bias so that research quality could be assessed.

- *Analysis*

This is an emerging field, and it was anticipated that there would not be enough data for a meta-analysis. The reported results were qualitatively analysed and presented thematically. This included characteristic of *in vivo* experiments which assessed the role of exogenous VGF-derive peptides on pain-associated behavioural outcomes, *in vivo* experiments which investigated antagonism of the effect of VGF-derived neuropeptides, measuring expression of VGF in animal nerve injury models, the characteristics of *in vitro* studies and the methodological quality of the included studies. The discussion was informed by the included studies and fulfilled the expected role of a literature review; to describe what is known and provide critical analyses of the current work. The strength of the evidence through the identification of studies in agreement or in conflict was discussed across the major themes that relate to peptide signalling and included: localisation and upregulation, pronociceptive effects, and the receptors and their role in pronociceptive signalling. A limitations section was also included.

2.3.2. 'Cannabinoids, the endocannabinoid system, and pain: a review of preclinical studies' (Finn et al., 2021)

2.3.2.1. Introduction & Aim

The aim of the International Association for the Study of Pain's (IASP) Presidential Task Force (TF) for Cannabinoid and Cannabis Analgesia was to assess all the literature pertaining to the use of the cannabinoids, cannabis-based medicines, and endocannabinoid systema modulators for pain management. This included a review of the basic science which was conducted in two parts. Firstly, the narrative review describes the physiology of the endocannabinoid system, the value of studying cannabinoids and endogenous cannabinoid system modulators in animal models of pain, discusses the pain-related efficacy and additional pain-relevant effects and has identified directions for future research. This narrative review was supported by the results of the companion systematic review and meta-analysis of cannabinoids, cannabis-based medicines, and endocannabinoid system modulators tested for antinociceptive effects in animal models of injury-related or pathological persistent pain

(Soliman et al., 2021). It aimed to directly address research questions related to efficacy that could not be synthesised within a conventional narrative review but require meta-analytical techniques. The systematic review informed the section 'Antinociceptive efficacy of cannabinoids in animal models of pathological or injury-related persistent pain.'

2.3.2.2. Methods & Results

Within the narrative review, the key findings of the systematic review were summarised; this included the scope of the research and the most frequently reported study design characteristics as well as methodological quality of the animal literature. The systematic review also highlighted several differences between the preclinical and clinical assessment of candidate treatments including the type of candidates and the weaknesses of face and construct validity of the animal models (Finn et al., 2021, Soliman et al., 2021). A specific meta-analysis was performed to give a high-level summary of the efficacy of the different drug classes in the most frequently reported rodent inflammation and neuropathic pain models (as presented in Tables 2.4 and 2.5 from Finn et al. (2021)).

In rodent inflammation models (e.g., formalin, Complete Freund's Adjuvant (CFA), carrageenan and osteoarthritis), cannabinoid receptor type 1 (CB_1) and CB_2 receptor agonists and palmitoylethanolamide (PEA) consistently attenuated pain-related behaviours across the range of inflammation models (Table 2.4). The only exception was for carrageenan-induced inflammation in rats, for which CB_1 receptor agonists did not significantly attenuate pain-associated behaviours. The efficacy of FAAH inhibitors was mixed; pain-associated behaviours were significantly attenuated in formalin and CFA but not in osteoarthritis rodent models. In carrageenan models, a species difference was detected in which FAAH inhibitors significantly attenuated pain-related behaviours in mice but not rats. Tetrahydrocannabinol (THC) also significantly attenuated pain-associated behaviours in formalin, CFA, and carrageenan models. Like FAAH inhibitors, the efficacy of cannabidiol (CBD) was mixed; pain-related behaviours were significantly attenuated in formalin and CFA but not in carrageenan and osteoarthritis rodent models.

In rodent neuropathic pain models e.g., nerve injury, chemotherapy-induced peripheral neuropathy and diabetes, CB_1 and CB_2 receptor agonists, FAAH inhibitors and CBD consistently demonstrated antinociceptive efficacy. PEA also significantly attenuated pain-related behaviours in nerve injury and chemotherapy models. THC significantly attenuated pain-related behaviours in nerve injury models (Table 2.5).

Table 2.4. Summary of the effects of cannabinoids, cannabis-based medicines, and endocannabinoid system modulators on pain-related behaviour in the most frequently used rodent models of inflammation (Finn et al., 2021). K; number of comparisons, N; number of animals within each subgroup; NT; not tested, ns; not significant.

	Formalin				Complete Freund's Adjuvant				Carrageenan				Osteoarthritis			
Drug Class	K	N	Effect size	95 % CI	K	N	Effect size	95 % CI	K	N	Effect size	95 % CI	K	N	Effect size	95 % CI
Anandamide transport inhibitor	10	64	1.97	0.608 - 3.342	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT
CB_1 receptor agonist	59	496	1.12	0.731 - 1.502	21	144	0.74	0.378 - 1.107	ns	ns	ns	ns	3	16	1.22	0.378 - 1.107
CB_1 receptor inverse agonist	1	7	1.71	0.427 - 3.002	NT	NT	NT	NT	5	30	0.44	0.055 - 0.822	NT	NT	NT	NT
CB_1 receptor PAM	NT	NT	NT	NT	NT	NT	NT	NT	1	8	2.45	1.067 - 3.837	NT	NT	NT	NT
CB_2 receptor agonist	33	265	0.81	0.316 - 1.307	44	338	1.10	0.829 - 1.366	10	74	1.39	0.885 - 1.896	24	173	1.42	0.829 - 1.366
CB_2 receptor inverse agonist	3	22	1.05	0.653 - 1.444	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT
CBD	2	16	1.86	1.548 - 2.169	5	20	1.59	0.71 - 2.469	ns	ns	ns	ns	ns	ns	ns	ns
Dual FAAH/MAGL inhibitor	NT	NT	NT	NT	ns	ns	ns	ns	4	36	1.42	0.25 - 2.6	NT	NT	NT	NT
FAAH inhibitor	23	184	1.50	1.134 - 1.862	24	144	1.24	0.702 - 1.781	ns	ns	ns	ns	ns	ns	ns	ns
FABP inhibitor	ns	ns	ns	ns	15	98	0.78	0.194 - 1.373	6	36	1.42	0.947 - 1.895	NT	NT	NT	NT
Monoacylglycerol lipase inhibitor	8	69	2.28	1.382 - 3.175	ns	ns	ns	ns	8	60	2.67	0.967 - 4.378	ns	ns	ns	ns
NAAA inhibitor	NT	NT	NT	NT	3	30	1.63	1.235 - 2.025	12	72	1.91	0.923 - 2.897	NT	NT	NT	NT
Non-selective cannabinoid receptor agonist	38	253	1.56	1.019 - 2.11	ns	ns	ns	ns	45	311	0.74	0.202 - 1.268	ns	ns	ns	ns
PPAR-alpha agonist	21	143	1.73	0.701 - 2.761	2	16	2.53	0.851 - 4.205	26	173	3.26	1.51 - 5.006	9	75	0.98	0.851 - 4.205
THC	20	118	1.16	0.679 - 1.643	16	109	2.34	1.687 - 2.986	5	37	2.13	0.138 - 4.12	NT	NT	NT	NT

Table 2.5. Summary of the effects of cannabinoids, cannabis-based medicines, and endocannabinoid system modulators on pain-related behaviour in the most frequently used rodent models of neuropathic pain (Finn et al., 2021). K; number of comparisons, N; number of animals within each subgroup; NT; not tested, ns; not significant.

	Nerve injury				Chemotherapy				Diabetes			
Drug Class	K	N	Effect size	95 % CI	K	N	Effect size	95 % CI	K	N	Effect size	95 % CI
Anandamide transport inhibitor	1	5	7.95	3.26 - 12.643	NT	NT	NT	NT	NT	NT	NT	NT
CB_1 receptor agonist	102	756	1.17	0.842 - 1.5	31	173	1.47	1.007 - 1.937	30	183	1.74	1.234 - 2.248
CB_1 receptor inverse agonist	4	36	1.71	1.171 - 2.248	NT	NT	NT	NT	2	12	4.53	1.932 - 7.127
CB_2 receptor agonist	87	639	1.15	0.812 - 1.497	27	192	1.549995	1.117 - 1.983	10	88	1.29	0.772 - 1.818
CBD	9	72	1.86	1.307 - 2.406	10	76	1.31	0.594 - 2.021	7	49	1.08	0.257 - 1.896
Diacylglycerol lipase inhibitor	1	6	3.69	1.235 - 6.148	ns	ns	ns	ns	NT	NT	NT	NT
Dual FAAH/MAGL inhibitor	2	12	2.08	1.348 - 2.808	NT	NT	NT	NT	NT	NT	NT	NT
FAAH inhibitor	59	439	1.75	1.212 - 2.281	19	148	1.96	1.048 - 2.868	17	180	2.53	1.455 - 3.604
Monoacylglycerol lipase inhibitor	16	119	1.31	0.774 - 1.84	12	80	1.42	0.179 - 2.653	NT	NT	NT	NT
NAAA inhibitor	5	34	4.54	2.594 - 6.496	NT	NT	NT	NT	NT	NT	NT	NT
Non-selective cannabinoid receptor agonist	75	504	1.18	0.825 - 1.536	ns	ns	ns	ns	8	48	1.89	0.335 - 3.437
PPAR-alpha agonist	26	229	1.14	0.493 - 1.786	11	74	2.54	1.341 - 3.746	NT	NT	NT	NT
THC	9	86	1.26	0.838 - 1.677	NT	NT	NT	NT	NT	NT	NT	NT

2.4. Discussion

The two reviews demonstrate that systematic review processes are required to ensure that all relevant research is identified and that some research questions e.g., efficacy of interventions can only be answered by objective extraction of study results for combination in a meta-analysis. The two reviews demonstrate the value of performing mixed methods reviews. Mixed methods combine the strengths of both review types and allows for the freedom and exploration of a topic, including subject matter expert critique and interpretation while mitigating selection and confirmation bias.

2.4.1. Systematic Review Processes Reduce Selection and Confirmation Bias in Narrative Reviews

The conduct of a systematic search and study selection mitigated selection and outcome reporting bias in the review of VGF. Presenting study characteristics in a format that allowed for comparison prevented a subjective weighting of any study and ensured that the relationship between the study characteristics and study results was considered. An example of the utility of this format was that the tail immersion assay was the most frequently reported pain-associated behavioural assessment and several authors had compared the results of the different studies with the suggestion that the results have been independently replicated. However, the experimental design was of poor methodological quality and the individual studies differed including assessment of different peptides, doses, and duration.

Themes emerged from the included studies that allowed for the more conventional exploration and appraisal of the literature base, and each study where applicable informed the discussion. Each study's contribution and whether the findings are corroborated or conflicted were discussed. Confirmatory findings were identified, for example, seven of the included studies demonstrate upregulation of VGF after nerve injury through several types of experimental investigation methods. Also, six studies reported pronociceptive effects of exogenous VGF-derived peptides, and all experiments reported an increase in mechanical and thermal hypersensitivity. There were no studies that contradicted these results. This approach explicitly provides the reader with greater confidence in the conclusions drawn from the synthesis of the evidence and simultaneously questions the possibility of publication bias given that no negative or neutral results have been reported. As discussed within the manuscript, the results of any form of evidence synthesis are reliant upon the availability of all research and therefore the analysis may be susceptible to publication bias (Soliman et al., 2019b).

Similarly, there is clarity in terms of where the knowledge gaps are. For example, there is uncertainty around the expression and role of the receptors in pronociceptive signalling and how the VGF-derived peptides exert their effect, and it is yet to be confirmed whether together they have a role in the initiation, development, and maintenance of neuropathic pain. It is also evident that VGF expression has not been characterised in neuropathic pain patients which is critical for continued understanding, determination of potential use as biomarkers and whether targeting offers a viable therapeutic strategy. Lastly, the inclusion of assessing the

reporting of methods to reduce the risk of bias highlights potential weaknesses of the evidence base: the results presented in the studies may be at risk of bias and as such recommendations have been made to ensure the reliable conduct and reporting of animal research.

Narrative reviews can be useful for summarising the literature and providing guidance if they are of high methodological quality. The VGF review provides a search strategy that can be reproduced; a list of included studies is included, and the discussion is structured in such a way that each study's results are discussed. Transparent reporting within narrative reviews is necessary to determine if there has been a comprehensive search and critical appraisal of all relevant studies. Similarly, any synthesis methods should be described in detail to allow for assessment of potential sources of bias. Reviews that are not performed in a systematic way should not be relied upon to draw conclusions. Narrative reviews that claim an effect should be interpreted with caution as it is unlikely to be a research question that can be answered through conventional narrative review processes.

2.4.2. Narrative Reviews can be Complimented by Systematic Review Processes

The example of a narrative review (Finn et al., 2021) being supported by a systematic review (Soliman et al., 2021) from the IASP's Presidential TF demonstrates that the two review methods can be complimentary. Doing so explicitly recognises the strengths and limitations of each review type. Often due to the volume of literature and the strict reporting processes of a systematic review it is not possible to provide the important biological and pharmacological contexts as are discussed in narrative reviews. I did not replicate the analysis and presentation of the results as in the systematic review but conducted a high-level synthesis to improve accessibility of the detailed findings presented in the systematic review. This analysis provided strength to the narrative review. It would not be possible to assimilate and synthesise the efficacy data without a meta-analysis. Single studies often report efficacy across a range of pain-associated behavioural assessments and many of which are not adequately powered (Soliman et al., 2021). By combining the data from individual studies through meta-analysis the power and precision of the results can be increased (Noordzij et al., 2009). It was this that has provided critical insights into the efficacy of the different drug classes for different pain types. For example, mirrored in all the analyses is the mixed efficacy of FAAH inhibitors and a failure to attenuate pain-associated behaviours in rodent models of osteoarthritis (Finn et al., 2021, Soliman et al., 2021). In a commentary for the journal PAIN, Lichtman and Chapman (2011) cited the animal literature as showing promise of FAAH inhibitors as treatment for osteoarthritis associated pain and highlight that the clinical relevance of animal osteoarthritis models is debated therefore there is a need for clinical trial. However, the results from our systematic review highlight that the issue may be related to reliance of single animal studies without critical appraisal, not only the clinical relevance of the model. Narrative reviews of endocannabinoid system modulators including FAAH inhibitors as potential therapeutics (Lichtman and Chapman, 2011, Schlosburg et al., 2009, Kaur et al., 2020, Saario and Laitinen, 2007, Woodhams et al., 2017, Guindon and Hohmann, 2009) are characterised by discussion of the physiological framework for analgesia

before reviewing the evidence for analgesic efficacy using animal models of pain conditions. Consistently reviewers refer to single studies without critical appraisal as evidence of antinociceptive efficacy thereby not considering risk of bias and robust study design. The finding that FAAH inhibitors were not effective in animal models of osteoarthritis not only supports the necessity of efficacy related research questions to be addressed by meta-analytical approaches but supports the utility of prospective reviews of animal data prior to clinical trial: the FAAH inhibitor, PF-04457845, failed to demonstrate analgesic efficacy in a randomised placebo-controlled clinical trial of osteoarthritis patients (Huggins et al., 2012).

2.4.3. Methodological Quality and Hierarchy of Evidence Synthesis

A reason that systematic reviews are often considered superior is because that they are more likely to have a focused research question, the methods are explicitly stated and therefore are less likely to be biased. Greenhalgh et al. (2018) on reflection of a drive toward systematic reviews over narrative reviews in the clinical domain, argue that the belief that systematic reviews are synonymous with high quality has led to the proliferation of systematic reviews which can allow data aggregation to claim a more privileged position in the hierarchy than deserved. Throughout this chapter I have referred to the methodological weaknesses of conventional narrative reviews however, it should also be stated that even a review reported as systematic can be at risk of bias due to poor methodological quality and therefore may not be as reliable as perceived. The quality of all research outputs should be appraised.

In a survey of prevalence of narrative and systematic reviews in five major medical journals, Faggion et al. (2017) show that most published reviews are narrative (despite Greenhalgh et al. (2018) assertion about the proliferation of systematic reviews) and argue therefore that much of the information reported in these journals is at risk of bias and their findings not reproducible. While their analysis focused on clinical literature it will undoubtedly be the case for preclinical literature where in pain research there have been only 12 preclinical systematic reviews published (PubMed search as of 3 May 2021). While all reviews are at risk of bias, the difference between a narrative and systematic review is that systematic reviews with their defined structure can be audited and quality evaluated. For systematic reviews there are several guidelines to assist e.g. AMSTAR 2 (Shea et al., 2017), GRADE (Hooijmans et al., 2018, Atkins et al., 2004) and ROBIS (Whiting et al., 2016) but attention should also be paid to the research question, inclusion/exclusion criteria and analysis methods because these are areas where authors can introduce forms of bias and provide misleading conclusions under the auspices of a methodologically sound process. For example, in the IASP TF for Cannabinoid and Cannabis Analgesia an overview of systematic reviews of randomised controlled trials found that most of the reviews were of low or critically low quality and therefore could not be relied upon for evidence-based decision making (Moore et al., 2020). Indeed, Fountoulakis et al. (2014) based upon the meta-analyses of major depression studies argues that meta-analyses have the potential to provide the highest level of evidence for evaluating interventions but they are at danger of over-analysis, flawed analysis

and over-interpretation and the methodological flaws and limitations need to be recognised otherwise their utility and validity will be undermined.

In contrast, evaluating the methodological quality of a narrative review is very challenging if not impossible. Of concern is evidence presented by Faggion et al. (2017), that journals with the highest impact factor tend to publish more narrative reviews and that readers may not be aware of the importance of review methodology and may wrongly assume that a review published in a highly ranked journal is sufficient demonstration of quality of the article, yet it has been shown that there is not a correlation between journal impact factor and article quality (Macleod et al., 2015, Nestor et al., 2020). A study's quality should be determined on its own merits, and this applies to both primary and secondary research. The goal therefore is to ensure that all reviews are methodologically transparent so that readers can judge the quality of the review and if required the review can be reproduced.

2.4.4. Reducing Reporting and Citation Bias

All reviews and their conclusions are subject to the availability and the quality of evidence upon which they are based. Regardless of review type steps can be employed to reduce bias and an important consideration for all reviews are reporting and citations biases. These are comprised of publication bias, outcome reporting bias, spin, and citation bias. Publication bias refers to not publishing the results of an entire study and it is more likely that the results of those studies will be neutral (Sena et al., 2010, Macleod et al., 2004). Outcome reporting bias refers to the failure to publish negative outcomes of a study which may be facilitated by changing the status of primary and secondary outcomes (Dwan et al., 2013), although in preclinical research the primary outcome is rarely stated (Landis et al., 2012). Defined as spin, the reporting and interpretation of results can also mislead readers (Boutron et al., 2014, Boutron et al., 2010). Lastly, citation bias refers to the phenomenon that positive results receive more citations than negative studies thereby making it harder to identify negative studies and leads to a heightened visibility of positive results (Duyx et al., 2017). In a review of the evidence base for antidepressants, de Vries et al. (2018) illustrate the collective and compounding impact of these biases which results in negative study results being concealed from view (Figure 2.1). While this study focuses on the reporting of clinical trials it's a very effective presentation of the biases at work. de Vries et al. (2018) demonstrate the need to assess individual study results, rather than rely upon author conclusions. The need to conduct independent literature searches is also highlighted because reference lists are likely skewed towards more positive results. It is not clear how narrative reviewers assess study results and nor is it likely that they perform systematic searches therefore narrative reviews are susceptible to the biases as presented by de Vries et al. (2018) (Figure 2.1). Similarly, Leng et al. (2019) in a citation network analysis of the debate of the vasopressin-memory hypothesis show that although citations are not a measure of study quality, they reflect influence. Positive and supportive studies were cited more than negative studies and the reasons for this cannot be attributed to differences in study quality. They highlight that "structural forces" in science can lead to the persistence of some hypotheses while

concurrently allowing the systematic dismissal of negative evidence. It is therefore critical that researchers are cognisant of these biases and their impact and mitigate accordingly when reviewing or synthesising evidence so that their secondary research does not promulgate distorted interpretations of results. There are several requirements to mitigate against these biases including the preregistration of studies and publication of all results. In terms of reviewing and evidence synthesis, the most effective forms of mitigation are systematic searches to identify all relevant literature and assessment of individual study data, not reliance on what has been stated by authors or other researchers.

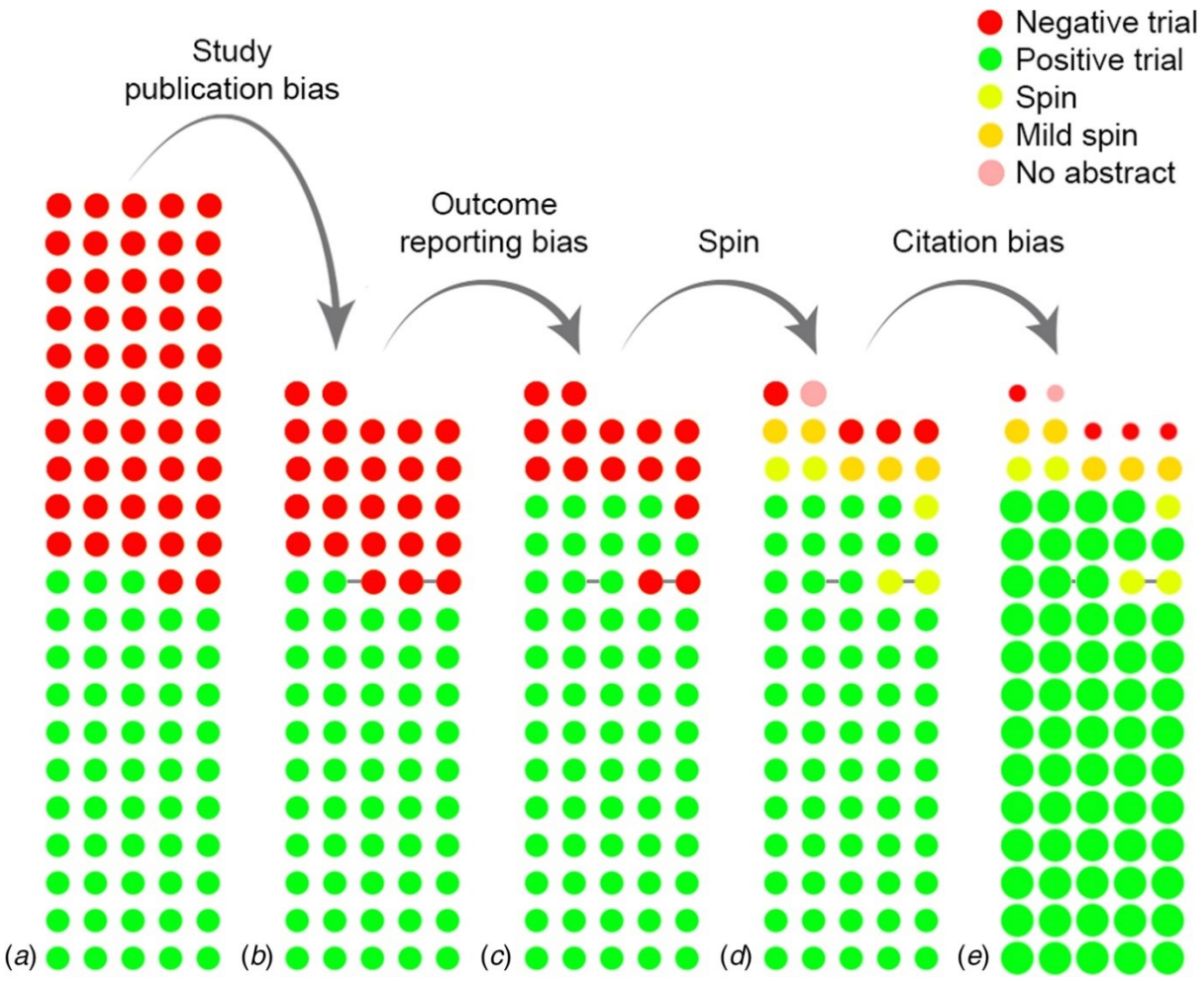


Figure 2.1. The cumulative impact of reporting and citation bias on the evidence base for antidepressants (de Vries et al., 2018). (a) represents the initial cohort of completed trials (53 positive and 52 negative or questionable). Each trial is a single-coloured circle, and the colour represents the results or presence of spin. (b) 52 of the positive trials were published, and only 25 of the negative trials were published and by (c) 10 of the negative trials become 'positive' in the literature by omitting the negative outcomes or switching the status of primary and secondary outcome measures. (d) Of the remaining 15 negative trials, the abstract of five trials concluded the treatment was effective and a further five had mild spin by suggesting that the treatment is numerically superior to placebo. Only four trials remained 'negative' and reported that the treatment was not more effective than placebo. In (e) the size of the circles represents the relative number of citations received by that category of studies which shows that positive trials were cited three times more frequently than negative trials.

2.4.5. Feasibility

It can take more time to conduct a narrative review supported by systematic review processes than a conventional narrative review, but the product is more comprehensive and reliable. All researchers are susceptible to bias and given the volume and breadth of literature pertaining to any scientific field it is not humanly possible to provide objective sense of it all. The utility of narrative reviews can be falsely assumed due to high citation which can also lead to citation bias and distortion of the scientific record. The question must be asked; what the added value is and how can the value and reliability be maximised to minimise research waste.

2.4.6. The Case for Mixed Methods and Systematic Reviews

Narrative reviews should not be eliminated completely. They have utility in facilitating discussion around topics but should not be used to inform strategies to address human disease. However, consumers of the review are better served when the choices made within the review, are explicit, transparent, clearly communicated, and reproducible. The combination of systematic review processes strengthens narrative reviews. A mixed methods approach is a good strategy for mapping themes as well as identifying knowledge gaps in the literature. It does not aim to provide a rigid or restrictive framework that prevents expression of expert opinion. There is utility in the sharing of expertise and experiential perspectives (Greenhalgh et al., 2018) and in fact all evidence synthesis activities are likely to be strengthened by being conducted by multi-disciplinary teams. The fundamental problem of not conducting a systematic search of the literature is that authors can be led to misunderstand the meanings of their data because they will not have comprehensively understood their true limitations and strengths. However, there needs to be clarity between empirical evidence based upon full appraisal of all available literature and opinions based upon experience. It is human nature that even empirical evidence can be interpreted differently through different experiential lenses.

Much argument for and against systematic and narrative reviews are centred around systematic reviews having a very limited scope which is to determine effectiveness of an intervention. In fact, preclinical systematic reviews can have a much broader remit e.g., from assessing methodological quality to study design characteristics required for external validity to the replacement, refinement, and reduction of animals in research. However, systematic review methods could be strengthened by adopting the strengths of narrative reviews. The copious tables of results and their formulaic nature can be difficult to read and if the intended audience do not find the format accessible it may undermine the utility of such a review. It is with this in mind that I have often made the case that a systematic review team should be multi-disciplinary including subject matter experts and the intended consumers of the research e.g., primary researchers in the case of preclinical systematic reviews. Yet, systematic reviews allow for the synthesis of the state of current knowledge so that future research can be guided in a more objective manner than narrative reviews can, putting future research in the context of existing knowledge (Higgins and Green 2008) and Faggion et al.

(2017) make the case that information serving as a basis for further research or decision making should be published only when a systematic and reproducible approach was used.

2.5. Conclusion

There are a lot of different types of review with distinct and overlapping methodologies. This chapter has not sought to discuss all the different review types but focus on the requirement for authors to declare *a priori* their aims and select and diligently utilise the appropriate methods. The use of mixed methods highlights that evidence synthesis has a spectrum of methods and there is not the dichotomy of a narrative or a systematic review. The selected methods should be based on the appropriateness in relation to the research questions and aims of the review. It should however be noted, that given the volume of literature a narrative review will always be stronger if all the relevant literature is considered through support of a systematic search.

It is also important to note that current partisan discussions about review types and their value is occurring in the clinical domain where systematic reviews have an established and critical role in evidence-based medicine. However, in the preclinical domain where the uptake of systematic reviews is still low, there is a requirement for their utility to be realised. The strength of argument for preclinical systematic reviews in the current climate of poor research methodological quality, translational failure, and tightening research budgets with requirement to reduce research waste is by far greater. Subject matter expertise and insight is invaluable but the most effective way for that to be brought to bear is through complimentary synthesis processes and open and transparent methods so that the quality of a review can be assessed, and the findings reproduced. Researchers must do all that they can to limit the pernicious threat of biases distorting the knowledge base and prevent the undermining of the validity of evidence-based decision making.

2.6. Recommendations

- Authors should conduct reviews supported by systematic processes; choose the most appropriate method for their research needs and ensure that they limit the promulgation of biased evidence and improve reproducibility.
- Journal editors should limit the amount of non-systematic evidence to reduce the probability of delivering biased information to readers.

2.7. Future Research Questions

- Identify barriers to performing mixed methods reviews in preclinical pain research to improve their uptake
- Assess reporting and citation bias in pain research e.g., trials of interventions for the treatment of neuropathic pain
- Use citation network analyses to map the instigation and proliferation of key hypotheses, determine influence and reliability

3. VGF, a biomarker and potential target for the treatment of neuropathic pain?

3.1. Background

Current pharmacological, behavioural, and surgical therapies for chronic pain have limited efficacy. There are several reasons for this including an insufficient understanding of the neurobiological mechanisms of chronic pain, poor translation of preclinical trial data, and challenges inherent to clinical trials and the patient population. Therefore, there is a need to further develop translation tools including the identification and validation of biomarkers which can be used to predict disease progression, select patients for clinical trials, confirm target engagement of novel therapeutics and predict therapeutic efficacy (Davis et al., 2020).

Given the heterogenous pathophysiology, sensory and emotional experiences of pain, patient stratification is of particular importance. Biomarkers could reduce the variability and response to an intervention assessed in clinical trials by matching patients with treatments that are most likely to show greatest benefit according to disease pathophysiology, phenotype, and drug mechanism, creating opportunity for personalised treatment. Analysis of clinical development success rates between 2011 and 2020 shows that trials employing patients by preselection biomarkers have a two-fold higher likelihood of subsequent approval for population-based use of the drug (15.9 %) than those that do not (7.6 %) (Thomas et al., 2021a). Hence, biomarkers for acute and chronic pain, predisposition to pain, predictors of pain chronification, recovery and treatment outcomes are currently being investigated in both preclinical animal models and human studies (Davis et al., 2020).

Data to date suggest that there are multiple genetic mechanisms in chronic pain (Ray et al., 2018, Bullock et al., 2019, Calvo et al., 2019, Mogil, 2012a, Denk and McMahon, 2012). It is likely that genetic and epigenetic variations play a critical role in the risk, initiation, development, and maintenance of chronic pain and therefore have potential as prognostic biomarkers, and their influence needs to be understood. Greater molecular neurobiological insight into human nervous tissue is needed for the development of new therapeutics. As experimental methods have evolved there is increasing capacity to investigate the role of genomics in pain disorders and how their molecular cascades can be effectively targeted.

To elucidate the mechanisms underlying peripheral neuropathic pain, Maratou et al. (2009) compared the gene expression in dorsal root ganglia of rats. Vgf was commonly upregulated in models of HIV (gp120 protein), HIV+antiretroviral (gp120+ddC), Varicella Zoster virus and spinal nerve transection. Since then researchers have been investigating the mechanism of VGF action and its role in pain. The aim of this narrative review was to employ systematic review methods to identify, select and critically appraise the literature which investigated the role of VGF in pain pathways. This chapter reports the review as published in *PAIN Reports* (Soliman et al., 2019b). Over 18 months have passed since publication, so I have also updated

the search which led to the identification and inclusion of one new relevant study. As discussed in the previous chapter, systematic search and study selection methods were used to ensure all relevant literature was included and to reduce biases usually associated with traditional narrative reviews.

3.2. Introduction

VGF (non-acronymic) is a granin-like neuropeptide precursor whose expression is robustly regulated by neuronal lesions and growth factors (Bartolomucci et al., 2011, Ferri et al., 2011, Levi et al., 2004). VGF-derived peptides have a functional role in several disorders including obesity, dementia, depression, and pain. This review focuses on the role of VGF in pain pathways. VGF was first identified due to its rapid induction in PC12 cells following treatment with nerve growth factor (Levi et al., 1985, Cho et al., 1989, Salton, 1991). Subsequent studies have identified numerous neurotrophins which also up-regulate VGF expression including brain-derived neurotrophic factor (BDNF) and neurotrophin-3, in targets such as cortical or hippocampal neurons (Alder et al., 2003). The *vgf* gene encodes a precursor protein of 615 (human) and 617 (rat, mice) amino acids (Levi et al., 2004, Salton et al., 2000). The VGF precursor protein sequence is highly conserved among rats and mice, with only 21 out of 617 amino-acids substitutions, none of which occur at the C-terminus. The precursor protein contains approximately a dozen cleavage sites and cleaves different peptides with specific neuronal bioactivities. Functional effects have been reported for several proteolytic products contained within the C-terminal 62 amino acid portion of VGF. Several of these VGF-derived peptides have been identified and are named by the first 4 N-terminal amino acids and their total length *(*e.g. TLQP-62, TLQP-21, HHPD-41, AQEE-11 and LQEQ-19, reviewed in Levi et al. (2004)(Figure 3.1). VGF has a tissue-specific pattern of expression and is robustly regulated and synthesised in immune cells, neuronal and neuroendocrine cells of both the central and peripheral nervous system (PNS) (Levi et al., 1985, Snyder and Salton, 1998, Salton, 1991, Pol et al., 1994, Busse et al., 2015, Busse et al., 2014, Cattaneo et al., 2010). In the peripheral nervous system (PNS), VGF is highly expressed in both neurons of the sympathetic ganglia and dorsal root ganglia of primary sensory neurons (Ferri et al., 1992). In the central nervous system (CNS), robust expression of VGF mRNA has been detected in the adult rat spinal cord and brain (Ferri and Possenti, 1992, Snyder et al., 1998, Snyder and Salton, 1998).

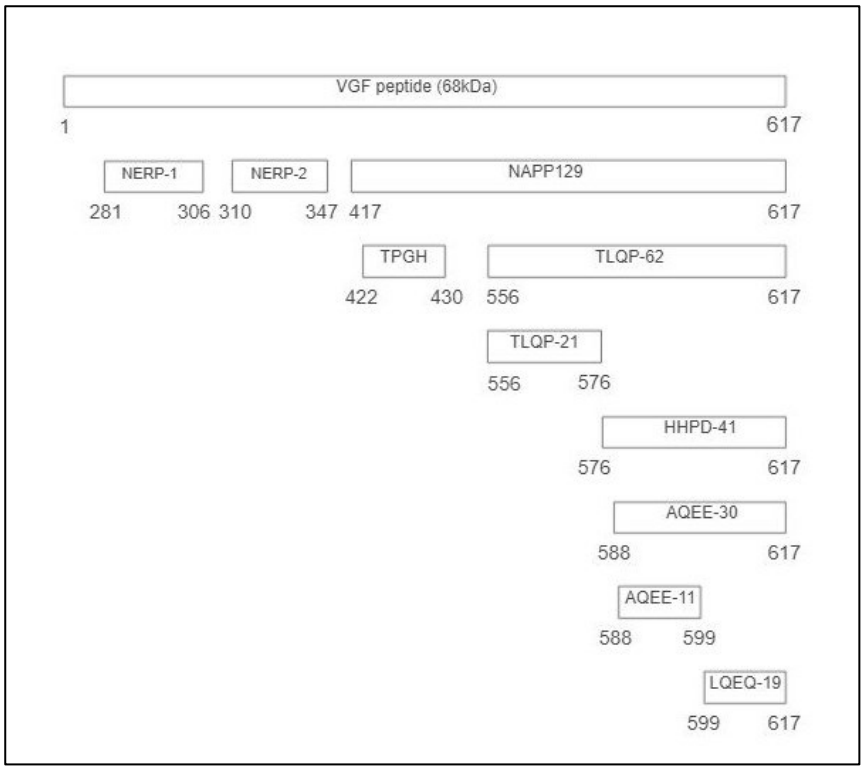


Figure 3.1. The *vgf* gene and its derived peptides. The VGF polypeptide is the precursor of several biologically active peptides. The gene contains specific amino acid sequences, as shown, of the pro VGF polypeptide and C-terminal VGF-derived peptides which are highly conserved between species.

Following peripheral tissue damage, nerve injury or inflammation, changes to protein expression and functional properties of sensory neurons leads to altered nociceptive processing in the spinal cord (Woolf and Costigan, 1999, Campbell and Meyer, 2006, White et al., 2007). Neuropathic pain (NP) is defined as "pain caused by a lesion or disease of the somatosensory system" (Treede et al., 2008, Jensen et al., 2011) and can result from various causes including disease, drugs, spinal cord injury and nerve trauma. The symptoms of NP are characterised by spontaneous pain produced in the absence nociceptor stimulation. NP can occur in the context of sensory loss (anaesthesia dolorosa) or sensory gain hyperalgesia, hyperpathia and allodynia which are triggered and maintained by a combination of peripheral and central mechanisms. Current animal models of injury-related or pathological persistent pain only reflect the sensory gain component. The occurrence of sensory abnormalities (gain and loss) varies between and within pathophysiological conditions (Merskey, 1994, Maier et al., 2010, Baron et al., 2010b). Each condition can be associated with multiple and various underlying pain generating mechanisms which are not fully understood. Current therapies have limited efficacies and unwanted side effects and therefore there remains a large unmet medical need to improve understanding and identify novel therapeutic targets (Finnerup et al., 2015, Finnerup et al., 2010).

In vivo evidence implicates the *vgf* gene and its products in pain signalling pathways. Increased levels of VGF mRNA and protein have been measured in the dorsal root ganglia (DRG) and spinal cord in animal models of pathological persistent pain (Moss et al., 2008, Riedl et al., 2009, Costigan et al., 2002, Valder et al., 2003).

The DRG contain the cell bodies of nociceptors and are the neurons which convey nociceptive signals. Therefore, they have been the focus of many gene expression studies in pain models, as transcriptional changes affect neuronal sensitivity. The VGF polypeptide is sorted into secretory granules and stored in the sensory terminals, processed into small peptides by endoproteolytic cleavage and released upon depolarisation activation (Levi et al., 2004, Riedl et al., 2009). Administration of exogenous VGF-derived peptides is reported to result in both mechanical and thermal hypersensitivity (Chen et al., 2013, Moss et al., 2008, Riedl et al., 2009). Similarly, several authors report endogenous VGF-derived peptides contribute to both inflammatory and nerve injury induced hypersensitivity (Chen et al., 2013, Fairbanks et al., 2014). Despite the regulation of VGF mRNA by peripheral nerve injury, the role of VGF in the initiation, development and maintenance of NP remains to be identified.

The most studied VGF-derived peptide is TLQP-21. It has been identified as a ligand for two receptors; the complement 3a receptor (C3aR1) and the globular head of the complement component C1q receptor (gC1qR). The C3aR1 is expressed on several immune cells including microglia and has been implicated in various immunomodulatory processes including CNS inflammation (Boos et al., 2004). It is hypothesised that TLQP-21 functions as an injury signal and leads to microglial activation (Doolen et al., 2017, Hannedouche et al., 2013). Increased expression of TLQP-21 following peripheral nerve injury or tissue damage may contribute to neuroimmune modulation of spinal neuroplasticity. The gC1qR is expressed on both microglia and macrophages (Chen et al., 2013). The subsequent release of bioactive molecules may, in part, be responsible for the mechanisms of nerve-injury induced hypersensitivity (Chen et al., 2013, Doolen et al., 2017). The relative contribution of these two receptors to the physiological effects of TLQP-21 is not yet known. However, greater understanding of the role of TLQP-21, microglia and macrophages in pain modulation will provide novel insights into pain mechanisms and has the potential to lead to the identification of novel therapeutic targets.

3.3. Aims

The review aimed to identify, select, and critically appraise all relevant research, and collect and analyse data from studies investigating the role of VGF-derived peptides in pain pathways. It is a demonstration of how systematic review methods can be used in support of a narrative review. It aimed to provide a less biased summary of research findings than a conventional narrative review (as discussed in Chapter 2) and allowed assessment of both the range of available evidence (external validity) and the likelihood that drawn conclusions are at risk of being confounded by bias (internal validity). It also provided direction for future research.

3.4. Methods

3.4.1. Literature Search

A living search for PubMed was set up using the search tool within the CAMARADES Preclinical Systematic Review Facility (SyRF) and the search term "VGF", publications were identified for screening. The live search adds potentially relevant publications as they are published allowing for continual, active monitoring of the evidence production. This assists a reviewer to determine how frequently new evidence should be sought and when new evidence should be incorporated into the review. The initial retrieval of search results was conducted on 30 April 2018. Subsequent updates were conducted on 29 August 2018 and 10 July 2019 prior to publication and most recently 4 June 2021 for this thesis chapter.

3.4.2. Study Selection

Studies were screened based on title and abstract by one review author (N.S.). Studies that described the role of VGF in pain were selected. Full texts of the studies were obtained on the basis that they may meet the pre-specified inclusion criteria (any study investigating the role of VGF in pain), and these comprised the systematic review dataset.

3.4.3. Data Extraction

Methodological quality was determined by the prevalence of reporting methods to mitigate the risk of bias. For each study detailing *in vivo* experimentation, quality was assessed (N.S.) against a checklist of factors relevant to animal models (van der Worp et al., 2010, Sena et al., 2007). This comprised of the following criteria:

(i) Random allocation to group

(ii) Allocation concealment

(iii) Blinded assessment of outcome

(iv) Sample size calculation

(v) Reporting of animal exclusions

In addition, journals commonly require authors to provide a statement of compliance with animal welfare regulations and declare possible conflict of interests, therefore the reporting of these parameters was extracted too.

3.5. Results

Four *in vitro* studies, ten *in vivo* studies (one additional study since the review was published in 2019) and one *ex vivo* study were selected for inclusion within this review (Figure 3.2). Excluded records did not discuss the role of VGF in a pain context or were not primary studies. The review focuses predominantly on the data presented within the *in vivo studies.* Summary information and study characteristics for the included studies are reported in Tables 3.1 - 3.4. Spared nerve injury is the most frequently reported model to determine expression levels of both VGF mRNA and protein following nerve injury (7 studies). Six of the studies report the pro-nociceptive effects of administered exogenous VGF-derived peptides. TLQP-21 is the most

commonly studied peptide; five studies assessed the effects of exogenous TLQP-21 on pain-associated behaviours. The most commonly reported outcome measure is tail flick latency from the warm-water tail immersion assay (3 studies).

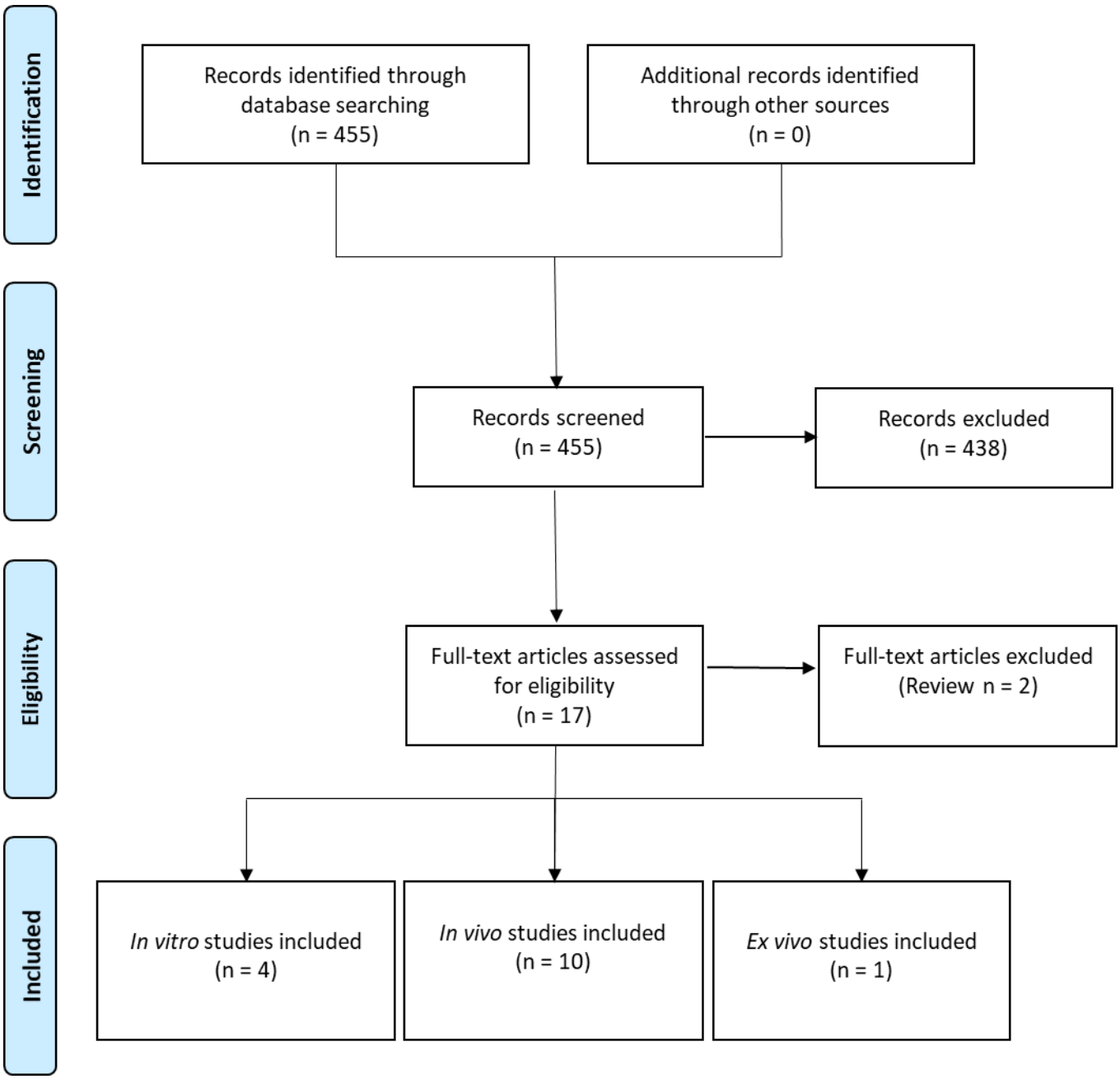


Figure 3.2. Study flow diagram. Screening and identification of publications investigating role of VGF in pain pathways. Appendix A references included studies. Updated with the additional search results since the publishing of Soliman et al. (2019b).

3.5.1. Study Characteristics

Six studies investigated the effect of exogenous VGF-derived peptides on pain-related behavioural outcomes (Table 3.1). Between the six studies, four different nerve injury models were used to assess changes in VGF expression; one study assessed changes in a model of Complete Freund's Adjuvant (CFA) induced inflammation (Table 3.2). Four studies used antagonism to verify the role of VGF-derive peptides in potentiating pain-associated behaviours or to derive the signalling pathways (Table 3.3). The most frequently used behavioural assessment following administration of exogenous VGF-derived peptides was the warm-water tail immersion assay (3 studies). Table 3.4 details the characteristics of the *in vitro* investigations in which VGF is implicated in pain signalling. Lastly, multiple authors make comparisons between the different studies and the results are summarised in Table 3.5.

Table 3.1. Study Characteristics for in vivo experiments assessing the effects of exogenous VGF-derived peptides. TLQP-21 is the most frequently use VGF-derived peptide. Its effect most commonly assessed in mice following administration into the spinal cord. Pain associated behaviours were most frequently assessed in the warm-water tail immersion assay.

Author	Year	Animal							Exogenous vgf	Dose (moles in a given volume)	Dose (mM)	Route of administration	Control
		Species	Strain	Sex	Age	Model	Sample size	Behavioural assessment					
Rizzi	2008	Mouse	CD1	Male	Unknown	Formalin	n = 9	Licking	TLQP-21	20, 40, 60, 80, 160 nmol/20 ul	1, 2, 4, 8 mM	Subcutaneous	Saline
						Naïve	n = 6		TLQP-21	20, 40 nmol/20 ul	2, 4 mM	Subcutaneous	Saline
						Formalin	n = 6		TLQP-21	1, 2, 4 nmol/1ul	1, 2, 4 mM	Intracerebroventricular	Saline
Moss	2008	Rat	Sprague Dawley	Unknown	Adult	Naïve	n = 8	Von Frey (Filaments) and Acetone	TLQP-62	25, 50 nmol/10 ul	2.5, 5 mM	Intrathecal	Saline
Riedl	2009	Mouse	Unknown	Male	Adult	Naïve	n = 6	Warm-Water Tail Immersion Assay	AQEE-30	0.003, 0.03, 0.3, 3 nmol/5ul	0.0006, 0.006, 0.06, 0.6 mM	Intrathecal	Saline
									AQEE-11	3 nmol/5ul	0.6 mM	Intrathecal	Saline
									LQEQ-19	3 nmol/5ul	0.6 mM	Intrathecal	Saline
Chen	2013	Rat	Wistar	Male	3 weeks	Naïve	n = 4	Von Frey (Electronic)	TLQP-21 (macrophage)			Intraplantar	Scrambled TLQP-21
Fairbanks	2014	Mouse	ICR/CD-1	Male	Adult	Naïve	n = 4-6	Warm-Water Tail Immersion Assay	TLQP-21	0.1, 1, 3 nmol/5ul	0.02, 0.2, 0.6 mM	Intrathecal	Unknown
Doolen	2017	Mouse	ICR/CD-1	Male	Unknown	Naïve	n = 4	Warm-Water Tail Immersion Assay	TLQP-21	1 nmol/5ul	0.2 mM	Intrathecal	TLQP-R21A

Table 3.2. Study Characteristics measuring expression of VGF in animal nerve injury models. VGF was upregulated in animal models of peripheral nerve injury and inflammatory pain. (Sample size refers to the number of animals which contributed tissue samples) Updated from (Soliman et al., 2019b).

Author	Year	Animal						Sample Location	Reported result
		Species	Strain	Sex	Age	Model	Sample size		
Moss	2008	Rat	Sprague Dawley	Unknown	Adult	Spared nerve injury	n = 8	L4/L5 DRG	3-fold upregulation of VGF mRNA
							n = 3	L4/L5 DRG	Significant increased expression of VGF protein in all cell bodies
							n = 4	Spinal cord	Increased expression of VGF protein in the spinal cord
Riedl	2009	Rat	Sprague Dawley	Male	Adult	Spinal nerve ligation		L4/L5 DRG	Significant increase in the number of vgf positive neurons in both injured L5 and non-injured L4 DRG
						Spinal nerve ligation	n = 4	Dorsal horn	Significant increase in vgf protein within the dorsal horn ipsilateral to SNL compared with sham or naïve animals
						CFA inflammation		L4/L5 DRG	CFA inflammation resulted in a more modest, yet still significant increase in vgf protein expression in the L5 DRG
Maratou	2009	Rat	Wistar	Male	Unknown	Spinal nerve transection	n = 3	DRG	vgf is significantly upregulated in the DRG of all models
						gp120+ddc	n = 3		
						VZV	n = 3		
Dawes	2014	Rat	Wistar	Male	Unknown	UVB irradiation	n=6	L4/L5 DRG	Significant upregulation of VGF FC= 2.7 (136.2/50.0) (Number of reads (UVB irradiation/control)
Doolen	2017	Mouse	ICR/CD1	Male	35 days	Spared nerve injury	n=14-16	Dorsal horn	Upregulation of C3aR1
Skorput	2018	Mouse	ICR	Male	Adult	Spared nerve injury	n=6	DRG	Increased expression of VGF protein in the spinal cord
Sun	2020	Rat	Sprague Dawley	Male	5 weeks old	Partial sciatic nerve ligation	n=3	L4/L5/L6 DRG	Upregulation of VGF mRNA
Sun	2020	Rat	Sprague Dawley	Male	5 weeks old	Partial sciatic nerve ligation	n=8	L4/L5/L6 DRG	Upregulation of VGF mRNA

Table 3.3. Study Characteristics of in vivo experiments investigating antagonism of the effect of VGF-derived neuropeptides.

Author	Year	Animal							Exogenous VGF	Dose (mM)	Route of administration	Antagonist	Dose	Route of administration
		Species	Strain	Sex	Age	Model	Sample size	Behavioural assessment						
Riedl	2009	Mouse	Unknown	Male	Adult	Naïve	n=6	Warm-Water Tail Immersion Assay	AQEE-30	0.06	Intrathecal	SB202190 (p38 MAPK)	0.1, 1,	Intrathecal
									LQEQ-19	0.02	Intrathecal		3 nmol/5ul	
Chen	2013	Rat	Wistar	Male	Unknown	PSNL	n=9	Von-Frey (electronic)	N/A	N/A	N/A	Anti-gC1qR	25 ug	Surgerical - left sciatic nerve
Doolen	2017	Mouse	ICR/CD1	Male	Unknown	Naïve	n=6	Warm-Water Tail Immersion Assay	TLQP-21	0.2	Intrathecal	SB290157 (C3aR1)	0.1, 0.3, 1 nmol/5ul	Intrathecal
						SNI	Unknown	Von-Frey (electronic/filaments)	N/A	N/A	N/A	SB290157 (C3aR1)	1 nmol/5ul	Intrathecal
Skorput	2018	Mouse	CD-1/ICR	Male	Adult	SNI	n=4	Von-Frey (electronic)	N/A	N/A	N/A	Anti-AQEE30	0.5, 1 ug	Intrathecal

Table 3.4. Characteristics of included *in vitro* studies.

Author	Year	VGF- derived neuropeptide studied	Cell type	Results
Kasuyama	2014	N/A	N/A	Clioquinol increases VGF mRNA levels, mediated by c-Fos
Chakraborty	2015	TLQP-21	N/A	Provides possible soultion structure
Akhter	2017	TLQP-21	Human neuroblastoma SH-SYFY	Identified HSPA8 as a cell surface binding protein of TLQP-21
Molteni	2017	TLQP-21	Chinese Hamster Ovary (CHO)	TLQP-21 activation (↑Ca2+) of a binding site on CHO cells. Receptor belongs to the family of Gq-coupled receptors.

Table 3.5. Comparison of data from the tail immersion assay. The tail immersion assay was most frequently used to measure pain-associated behaviours following the administration of exogenous VGF-derived peptides.

Author	Year	Animal (Species, Strain)	Sample size (per group)	Peptide	Dose (nmol /5ul)	Dose (mM)	Peak Hyperalgesia (time, min)	Magnitude (delta latency(s))	Duration (time, min)
Reidl	2009	Mouse, unknown	n = 6	IQEQ-19	3	0.6	10	1.7	60
				AQEE-30	0.003	0.0006	30	0.28	90
					0.03	0.006	30	1.11	90
					0.3	0.06	30	1.66	90
					3	0.6	30	1.28	90
Fairbanks	2014	Mouse, ICR/CD1	n = 4	TLQP-21	0.1	0.02	60	0.2	150
					1	0.2	60	1.82	150
					3	0.6	60	3.17	175
Doolen	2017	Mouse, ICR/CD1	n = 4	TLQP-21	1	0.2	60	2.62	120

3.5.2. Internal Validity / Risk of bias

Internal validity of the nine included studies which included animal experimentation was assessed against the reporting of five criteria considered relevant to animal models; blinded assessment of outcome, randomisation, allocation concealment, sample size calculation and the reporting of animal exclusions. The mean score for the selected publications was 1 out of 5. The overall reporting of quality measures for the included publications are presented in Table 3.6.

Table 3.6. Methodological quality of studies evaluated. To reduce selection bias, only two publications reported randomisation. Conversely, no publications reported allocation concealment or reported conducting a sample size calculation. Minimising the impact of performance and detection bias, blinding was reported as being conducted in 5 publications. To reduce the impact of attrition bias, the reporting of animal exclusions was less prevalent and reported in only one publication. In contrast, all publications reported compliance with animal welfare regulations and statement of potential conflicts of interest are more frequently reported.

Author	Year	Random allocation to group	Allocation concealment	Blinded assessment of outcome	Sample size calculation	Reporting of animal exclusions	Quality Score	Ethics statement	Conflict of Interest
Rizzi	2008	-	-	-	-	-	0	+	-
Moss	2008	-	-	+	-	-	1	+	+
Riedl	2009	-	-	-	-	-	0	+	-
Maratou	2009	-	-	+	-	+	2	+	-
Chen	2013	+	-	+	-	-	2	+	-
Fairbanks	2014	+	-	+	-	-	2	+	+
Doolen	2017	-	-	+	-	-	1	+	-
Dawes	2014	-	-	-	-	-	0	+	+
Skorput	2018	-	-	+	-	-	1	+	+
Sun	2020	-	-	-	-	-	0	+	+

3.6. Discussion

I performed a narrative review supported by systematic review methods to identify, select, and critically appraise all the published literature that explores the role of VGF in the initiation, development, and maintenance of NP. The search that I instigated is living, which means it is periodically updated as and when new articles are published. For this chapter, I returned to the search to find that 31 potentially relevant articles had been published in the preceding 23 months. Through screening I identified one new relevant study. The study (Sun et al., 2020) has added weight to the data presented by previous studies as it demonstrated that VGF is upregulated following nerve injury. Apart from the inclusion of one new study, the results, review and recommendations for future research remain unchanged from the published article (Soliman et al., 2019b). This combination of systematic review processes has strengthened the narrative review. The methods are explicit, transparent, and reproducible and threats of biases distorting the synthesis of the VGF literature have been limited. Summary tables of the studies were provided in the results section and each study has informed the thematic discussion below. The merits of this approach were discussed in more detail in Chapter 2.

3.6.1. Localisation and Upregulation of VGF

Large scale expression analyses have been used to study the changes that occur under chronic pain conditions, leading to the identification of genes which function in pain signalling, some of which are yet to

be fully characterised (Costigan et al., 2002, Valder et al., 2003, Wang et al., 2002). The prospective significance of VGF was first derived from proteomic differential expression analysis of an *in vitro* model of injured sensory neurons. VGF co-localises with Substance P (SP), calcitonin gene related peptide (cGRP), TrkA and P2X3 (Rizzi et al., 2008, Riedl et al., 2009). Subsequent experiments provide the evidence for upregulation of VGF in the DRG and dorsal horn following nerve injury and inflammation (Moss et al., 2008, Riedl et al., 2009, Maratou et al., 2009, Dawes et al., 2014, Skorput et al., 2018, Sun et al., 2020).

Moss et al. (2008) observed prolonged up-regulation of VGF mRNA and protein in the DRG neurons, central terminals and dorsal horn neurons in the spared nerve injury (SNI) animal model. Injury to the sciatic nerve was associated with a 3-fold up-regulation of VGF mRNA which was maintained for 3 weeks. In addition, both the amount of VGF mRNA per cell and the number of cells expressing VGF mRNA were significantly greater in the ipsilateral compared to the contralateral DRG. At 21 days post injury, a 20-fold increase of VGF positive cells was observed in the ipsilateral compared to the contralateral DRG.

Similarly, Riedl et al. (2009) demonstrate a rapid upregulation of VGF following nerve injury and inflammation. In axotomised neurons, following 24 h in culture a 45-fold increase in the expression of VGF protein was detected. These changes in protein expression were paralleled in the spinal nerve ligation (SNL) model corroborating earlier findings. In the L5 DRG, 24 h post injury, the number of VGF expressing neurons significantly increased; 80% in SNL animals compared to 37% in sham operated and 20% in naïve animals. Interestingly, immunohistochemical analysis showed that VGF was upregulated in injured as well as uninjured neurons. The increased number of VGF neurons in DRG of sham animals suggested that VGF may also be induced in inflammatory conditions in concordance with the associated tissue damage and inflammation following surgery. This is supported by evidence which shows that VGF is upregulated in the CFA model of inflammatory pain. CFA inflammation resulted in a significant increase in the number of VGF expressing neurons in the L5 DRG (28.2 ± 0.3% in CFA L5 DRG compared with 19.0 ± 2.8% in naive L5 DRG; p _ 0.05, t test).

More recently, Skorput et al. (2018) provide evidence for the presence of TLQP-62 in the spinal cord and its upregulation following peripheral nerve damage. Western blot analysis of lumber spinal cord lysates from naïve mice provide evidence for the bioavailability of C-terminal VGF peptides in the dorsal horn including that of TLQP-62. In concordance with other studies (Moss et al., 2008, Riedl et al., 2009), the authors also observed a nerve-injury induced increase in VGF levels in the DRG and spinal cord. Population analysis of anti-AQEE immunoreactivity in the DRG found increased immunoreactivity among DRG neurons of SNI mice relative to sham controls. In addition, the proportion of anit-AQEE30 labelled DRG neurons was higher in SNI mice compared with sham mice (SNI, 41.6 % ± 2.2 % vs sham 27.1 % ±4.0 %, n=6 per group, $P < 0.05$, unpaired Student *t* test). In addition, mice pre-treated with intrathecal anti-AQEE (1 µg) immediately before peripheral

nerve damage (SNI) demonstrated significantly higher paw withdrawal thresholds compared with controls in the electronic von-Frey test suggesting that VGF-derived peptides play a role in the initiation of nerve-injury induced hypersensitivity. Similarly, attenuation of mechanical hypersensitivity was observed in a separate cohort of mice that received a single intrathecal injection of anti-AQEE30 at the time of injury (Skorput et al., 2018).

Whole genome rat array studies were conducted by Maratou et al. (2009) to measure gene expression to identify genes that are directly relevant to neuropathic mechanical hypersensitivity. The authors compared the primary sensory neuronal gene expression profiles of three models of NP (peripheral nerve trauma (SNT), HIV infection and anti-retroviral induced neuropathy and varicella zoster infection). Although these have distinct aetiologies, they share a common outcome of mechanical hypersensitivity. Comparison of the three models identified 39 genes to be differentially regulated in the same direction suggesting these genes may be responsible for the common mechanical hypersensitivity phenotype. When mechanical hypersensitivity was well established the *vgf* gene was found to be upregulated as were NpY and Pap/Reg2 and therefore are potentially important for this phase of hypersensitivity. Similarly, Dawes et al. (2014) demonstrate that *vgf* with REG38 and CCL2 are amongst the most upregulated genes in the DRG in the ultraviolet B (UVB) inflammatory pain model. Riedl et al. (2009) and Moss et al. (2008) do not report measuring pain-associated behaviour before taking samples to measure VGF expression so a correlation cannot be drawn. More recently, Sun et al. (2020) analysed the transcriptomic changes in rat DRG following partial sciatic nerve ligation (pSNL). The authors confirmed hypersensitivity to mechanical and thermal stimuli post pSNL surgery and using RNA sequencing revealed 184 genes were downregulated and 414 genes upregulated including *vgf* in the DRG of rats 7-days post model induction. To further validate their data, they also used qualitative real-time polymerase chain reaction and confirmed the mRNA expression levels of VGF from the DRG of rats before, 7, 14 and 21 days after pSNL surgery. The results revealed that mRNA expression levels were significantly upregulated from 7 days after pSNL and were maintained at high levels at day 14 and 21 (Sun et al., 2020). Despite the differences between the studies, the evidence is compelling: VGF expression is induced and upregulated in sensory neurons following nerve injury and inflammation. Further investigation is required to identify candidate genes that are responsible specifically for the initiation, development and/or maintenance of NP.

Conversely, Lind et al. (2016) investigated the effects of electrical neuromodulation by spinal cord stimulation (SCS) on the cerebrospinal fluid (CSF) proteome of 14 patients responsive to SCS treatment and suggest that VGF upregulation may have an analgesic role. Most of the patients were diagnosed with radiculopathy but were a highly heterogeneous population in terms of the location of their nerve lesion(s) and where they perceive their pain. Two different proteomic mass spectrometry protocols were used to analyse the CSF samples; one taken following 48 h of the stimulator being turned off (SCSOFF) the second after normal use

for 3 weeks (SCSON). Each patient acted as their own control. The patients in the study reported pain relief following SCS treatment and most patients reported an increase in pain intensity scores when the device was turned off for 48 h, neither the patients nor the investigators were blinded. Additionally, there were possible confounding factors; firstly, the presence of paraesthesia and secondly, many of the patients had had lower back surgery which made sampling challenging. Therefore, sampling may have triggered inflammation affecting the proteome. The reported results indicate 86 CSF proteins were significantly altered in this cohort, of which VGF was in the top 12 and was increased by 38% in the SCSON condition. They report that the upregulated VGF-derived peptides were between the amino acid positions 195 and 309 corresponding to the NERP-1 chain (amino acid 281-306) (Toshinai and Nakazato, 2009). The authors have not stated how they validated their findings but speculate that VGF may have active fragments that moderate or counter the biological response of their precursors, similar to that of dynorphin, angiotensin, Substance P and nociceptin (Hallberg, 2015) and that during SCSON there is an increase of an analgesic VGF fragment which contributes to the long-term effects of SCS. While NERP-1 does not overlap with the TLQP-62 peptide, from which the pro-nociceptive effects have been observed, the role of NERP-1 in pain pathways is unknown and changes in the expression of NERP-1 may not be related to pain.

VGF is rapidly and robustly up-regulated by nerve growth factor (NGF) and brain derived neurotrophic factor (BDNF) which drive *vgf* gene transcription increasing VGF mRNA levels up to 50-fold in PC12 cells (Levi et al., 2004, Salton, 1991). NGF and BDNF regulate the development and maintenance of specific functions in different populations of peripheral and central neurons. Under normal physiological conditions NGF is synthesised by peripheral target tissues, whereas BDNF synthesis is highest within the CNS. We originally reported that following peripheral nerve injury, the expression of both NGF and BDNF increases rapidly (Meyer et al., 1992, Fukuoka et al., 2001). We also reported that peripheral nerve damage triggers NGF release by macrophages, mast cells and Schwann cells (reviewed by Pezet and McMahon (2006)), however, it is now understood from RNA sequencing that NGF transcripts are missing from these cells and that nerve injury does not cause NGF upregulation in peripheral sensory neurons (reviewed by Denk et al. (2017). More research is required to improve our understanding of the cell-type specific patterns of NGF however, its role in nociception is better understood (Denk et al., 2017). BDNF is up-regulated in tyrosine receptor kinase A (TrkA) expressing DRG cells. TrkA is a high affinity NGF receptor which mediates the effects of NGF. Additionally, NGF/TrkA signalling has been implicated in pain pathways particularly pain associated with inflammation (Mantyh et al., 2011, Ashraf et al., 2016). Anti-NGF treatment leads to reduction of hypersensitivity in peripheral nerve injury models (Pezet and McMahon, 2006) and many of the injury induced changes in the dorsal horn neuronal excitability are mediated by release of BDNF from microglia (Coull et al., 2005). In addition, BDNF is implicated in pain signalling through its interaction with the TrkB receptor (Reichardt, 2006). In the hippocampus, it has been demonstrated that TLQP-62 induces neuroplasticity through a BDNF-TrkB dependent mechanism (Alder et al., 2003, Bozdagi et al., 2008, Lin et

al., 2015, Thakker-Varia et al., 2014). TLQP-62 treatment potentiated glutamatergic responses in both the rat and mouse superficial dorsal horn (Skorput et al., 2018). However, TLQP-62 treatment in conjunction with the Trk inhibitor K252a, prevented the TLQP-62-induced increase of glutamatergic responses. These results demonstrate that the TLQP-62 induced increase in glutamatergic signalling in the dorsal horn is kinase dependent (Skorput et al., 2018). Therefore, the upregulation of VGF in sensory neurons following nerve injury is likely to result from a spike in activity and the upregulation of these neurotrophins in the DRG and dorsal horn (Costigan et al., 2002). The rapid and robust upregulation sets apart VGF from other peptides known to be increased after nerve injury (Fukuoka et al., 1998, Shortland et al., 2006, Zhang et al., 1997) and therefore release of VGF peptides from sensory neurons may provide an early signal for peripheral nerve injury and abnormal pain signalling. Similarly, the sustained increase of VGF suggests that it may have role in both the initiation and maintenance of neuropathic pain.

3.6.2. Pro-nociceptive effects of exogenous VGF-derived peptides

Six studies reported the pro-nociceptive effects of exogenous VGF-derived peptides (characteristics summarised in Table 3.1). Most assessed central effects following intrathecal administration to naïve animals (four studies). However, two studies did assess the peripheral effects of specific VGF-derived peptides with either subcutaneous or intraplantar administration. The most commonly assessed VGF-derived peptide is TLQP-21 (four studies), however, TLQP-62, AQEE-30, LQEQ-19 and AQEE-11 have also been assessed. All experiments report an increase in pain associated behaviours which are described as increased mechanical and thermal hypersensitivity.

Rizzi et al. (2008) were the first to provide evidence for the involvement of a VGF-derived peptide in pain modulation *in vivo* and suggested that TLQP-21 demonstrates a different action at peripheral and central levels of the nociceptive pathway. Firstly, peripheral injection of TLQP-21 to naïve mice was associated with increased licking responses suggesting an alteration in pain sensitivity. Secondly, TLQP-21 was observed to have a functional role in affecting formalin induced pain associated behaviour in mice. The formalin test is an acute inflammatory pain model: formalin injection to the paw induces a biphasic response. The first phase is the activation of nociceptors and corresponds to the release of peptides from nerve terminals. The second phase is posited to reflect ongoing peripheral activity and central sensitisation (Porro and Cavazzuti, 1993). In the second phase of the formalin model, the authors report that peripheral administration of the highest doses of 4 and 8 mM TLQP-21 results in worsening pain associated behaviours (increased licking responses) therefore suggesting that TLQP-21 may have a functional role in the inflammatory process induced by formalin. In support of TLQ-21 having a functional role in the inflammatory process, intraplantar injection of macrophages stimulated with TLQP-21, likely mediated through the gC1q receptor, evokes mechanical hypersensitivity in rats (Chen et al., 2013).

Conversely, intracerebroventricular (i.c.v) administration of 2 nM / 1 µl/ mouse TLQP-21 resulted in reduced licking response in the second phase, which the authors describe to be an analgesic effect (Rizzi et al., 2008) however, the lower dose of 1 mM and the highest dose of 4 mM did not have an effect. The authors postulate that TLQP-21 modulation of inflammatory pain depends upon the route of administration and that the analgesic effect of central administration could be a consequence of a modulatory role exerted on the descending inhibitory pathway of which the glutamatergic, adrenergic and serotoninergic systems could be the target of its action (Millan, 1999). However, this analgesic phenomenon does not appear to have been investigated further to provide confirmatory or concomitant evidence.

Acute intrathecal administration of TLQP-62 (2.5 and 5 mM) to naïve rats was associated with both mechanical (von Frey filaments) and cold (cutaneous acetone) behavioural hypersensitivity (Moss et al., 2008). Similarly, intrathecal administration of both AQEE-30 and LQEQ-19 (Riedl et al., 2009) and TLQP-21 (Doolen et al., 2017) to naïve mice, dose-dependently induced thermal hypersensitivity assessed in the warm-water tail immersion assay.

Authors have made comparisons about the magnitude and duration of effect of the VGF-derived peptides with TLQP-21 being described as having the greatest magnitude and duration of effect in the warm-water tail immersion assay (Table 3.5). However, the peptides were not assessed in the same experiment and therefore a prospective experiment is required to directly compare their actions and dose response to allow for conclusions to be drawn about magnitude and duration of effect.

Riedl et al. (2009) also evaluated several pharmacological agents known to interfere with pathways involved in nociceptive processing to determine whether they can inhibit the AQEE-30 and LQEQ-19 induced thermal hypersensitivity. These included inhibitors of nitric oxide synthase (7-NI, 10 nM; L-NAME, 100 nmol), NMDA receptor (MK801, 10 nmol), protein kinase C (GF109203X, 1 and 10 nmol), protein kinase A (KT5720, 1.7 nmol), and MAPK (U0126, 2.5 nmol; SB202190 0.1, 1 and 2.5 nmol; SB600125, 2.5 nmol). It's reported that only the p38 inhibitor SB202190, administered as a pre-treatment, dose-dependently reversed AQEE-30 (0.3 nmol) and LQEQ-19 (1 nmol) evoked thermal hypersensitivity (data of other inhibitors not shown) suggesting that the effects of VGF-derived peptides are mediated by p38. In corroboration of these findings Fairbanks et al. (2014) demonstrated TLQP-21 evoked p-38 MAPK dependent thermal hypersensitivity. The effects of intrathecal TLQP-21 were attenuated by treatment with either the p38 inhibitor (SB202190), COX inhibitor (indomethacin) or the lipooxygenase inhibitor (AA-861) assessed in the warm water tail immersion assay.

The *in vivo* evidence suggests that VGF-derived peptides may be involved in both nociception and inflammatory pain. VGF may have a direct or indirect action which leads to the release of other neuropeptides and neurotransmitters. It has been shown in cultured DRG cells that VGF-derived peptides are co-localised with substance P (SP)(Ferri et al., 1992). It is well established that SP has a role in nociception (Barber et al., 1979, Payan, 1989, Warden and Young, 1988) and in the pathophysiology of inflammatory

disease (O'Connor et al., 2004). VGF-derived peptides therefore may be part of the mixture of molecules produced by the DRG neurons which are secreted in response to nociceptive stimuli together with the release of SP (Malcangio et al., 1998) hence their proposed role in the inflammatory process. TLQP-21 stimulates COX activity (Fairbanks et al., 2014) leading to the synthesis of prostaglandins, important mediators of inflammation and pain. Prostaglandins increase levels of cyclic AMP and enhance nociceptor sensitisation by reducing the activation threshold of sodium channels (England et al., 1996) as well as sensitising primary afferent neurons to bradykinin and other mediators (Neugebauer et al., 1989).

3.6.3. Identification of receptors, their expression and their role in pro-nociceptive signalling

Despite the large volume of data on physiological effects of VGF-derived peptides, very little is known about the receptors with which VGF-derived peptides interact, their downstream signalling pathways and mechanisms of action. It has been hypothesised that VGF-derived peptides may lead to the sensitisation of sensory neurons by direct (autocrine) or indirect (paracrine) action. The complement component 3a receptor (C3aR) (Hannedouche et al., 2013, Molteni et al., 2017) and globular head of the complement component C1q receptor (gC1qR) (Chen et al., 2013) have been identified as receptors for TLQP-21; the receptors for the other VGF-derived peptides have not been identified. It is unanimously reported that TLQP-21 treatment leads to increased intracellular Ca^{2+} release. Ca^{2+} is a highly versatile second messenger involved in a variety of intracellular signalling pathways, including gene regulation, cell proliferation and death.

3.6.4. *The globular head of the complement component C1q receptor (gC1qR)*

Chen et al. (2013) report that gC1qR is expressed on both macrophages and microglia and the peripheral pro-nociceptive actions of TLQP-21 involve the activation of gC1qR on macrophages. *In vitro* they demonstrated that TLQP-21 treatment (not TLQP-62 or LQEQ-19) is associated with an increase in intracellular Ca^{2+} levels in cultured rat bone-marrow derived primary macrophages and brain derived primary microglia. However, TLQP-21 did not induce changes in intracellular Ca^{2+} levels in cultured DRG neurons suggesting that TLQP-21 does not directly hypersensitise sensory neurons (authors' did not publish the data) supporting the hypothesis that TLQP-21 effect is mediated by receptors expressed on macrophages and microglia (Chen et al., 2013). They also demonstrate that this increase in intracellular Ca^{2+} levels in macrophages is gC1qR dependent. Silencing of gC1qR with siRNA in macrophages significantly reduced the gC1qR protein expression and fewer macrophages responded to TLQP-21 treatment. Similar results were observed following pre-incubation of macrophages with neutralising gC1qR monoclonal antibodies resulting in significant reduction in response to TLPQ-21 (Chen et al., 2013).

Chen et al. (2013) also examined the role of gC1qR in pain pathways *in vivo*: firstly, intraplantar injection of TLQP-21 stimulated macrophages resulted in mechanical hypersensitivity in rats. Secondly, the gC1qR

antibody (Mab1) was applied to the site of nerve ligation in the partial sciatic nerve ligation (PSNL) model rats. Application of the gC1qR antibody delayed the onset of mechanical hypersensitivity associated with PSNL whereas, control IgG-treated rats demonstrated a reduction of punctate mechanical threshold. Therefore, macrophages stimulated by TLQP-21 via gC1qR may be responsible for the hypersensitivity of sensory neurons, characteristic of the mechanical hypersensitivity observed.

3.6.5. *The complement component 3a receptor (C3aR1)*

Hannedouche et al. (2013) first describe the identification of C3aR1 as a target for TLQP-21 in rodents. Doolen et al. (2017) provide functional and anatomical evidence that the spinal pro-nociceptive actions of TLQP-21 involve the activation of C3aR1 of microglia in the dorsal horn. They demonstrate extensive C3aR1 co-localisation with Iba-ir microglia in the dorsal horn (Iba1; a marker for macrophages and microglia). Their co-localisation analysis of C3aR1, GFAP, and NeuN labelling did not yield evidence for C3aR1 expression in astrocytes or neurons; therefore, TLQP-21 directly targets C3aR1 to activate spinal microglia rather than neurons. They also demonstrate that in Iba1-eGFP mice TLQP-21 elicits Ca^{2+} signalling in microglia with over 95% of eGFP-positive profiles responding to TLQP-21 suggesting that the Ca^{2+} release in response to TLQP-21 is a result of microglial activation. Additionally, in C3aR1 knock-out mice and wild type cultures exposed to R21A (an inactive peptide at C3aR1; a single amino acid substitution in the C-terminus of TLQP-21) (Cero et al., 2014), treatment with TLQP-21 did not lead to an increase in intracellular Ca^{2+} signalling, further reinforcing the evidence that TLQP-21 signalling is C3aR1 mediated (Doolen et al., 2017).

Doolen et al. (2017) also confirm *in vivo* the molecular and structural findings that TLQP-21 directly activates C3aR1 (Cero and Bartolomucci, 2014, Hannedouche et al., 2013). Peripheral nerve injury (spared nerve injury; SNI) increased both microglial C3aR1 expression and TLQP-21 evoked C3aR1 mediated Ca^{2+} signalling in the dorsal horn of the spinal cord. SNI also increased both the number of microglia with increased intracellular Ca^{2+} signalling and the peak amplitude indicating that nerve injury results in increased TLQP-21 signalling. Pharmacological inhibition of the C3aR1 with the antagonist SB290157 attenuated the hypersensitivity associated with SNI. In microglial cell culture, the increased Ca^{2+} signalling following treatment with TLQP-21 was also attenuated. Together these results provide the evidence that microglial C3aR1 signalling is required for nerve injury induced hypersensitivity (Doolen et al., 2017).

Despite the identification of two receptors, the signalling mechanisms of VGF-derived peptides are completely uncharacterised. Findings to date, describe a receptor within the CNS (C3aR1) that is activated by both a peptide (TLQP-21) and an immune mediator (C3a) which is mirrored in the periphery with gC1qR and C1q. Interestingly, gC1qR and C3aR1 are receptors for the complement proteins C1q and C3a respectively. They form part of the complement system which consists of enzymes, regulatory proteins and

receptors that regulate both innate and adaptive immune responses. It is possible that TLQP-21 interacts with both receptors. Since the TLQP-21 precursor VGF is up-regulated in sensory neurons within 24 h of nerve injury (Riedl et al., 2009) it is probable that TLQP-21 released from injured sensory neurons functions as a warning or protective signal that contributes to spinal neuroinflammation through activation of C3aR1 and peripheral neuroinflammation through gC1qR. It has not yet been investigated whether the gC1qR expressed on microglia in the CNS and whether C3aR1 expressed on macrophages in the periphery also play a role in TLQP-21/VGF-derived peptide signalling, however, the identification of these two receptors suggests a novel interplay of neuronal and immune signalling mediators in both peripheral and CNS diseases (Figure 3.3).

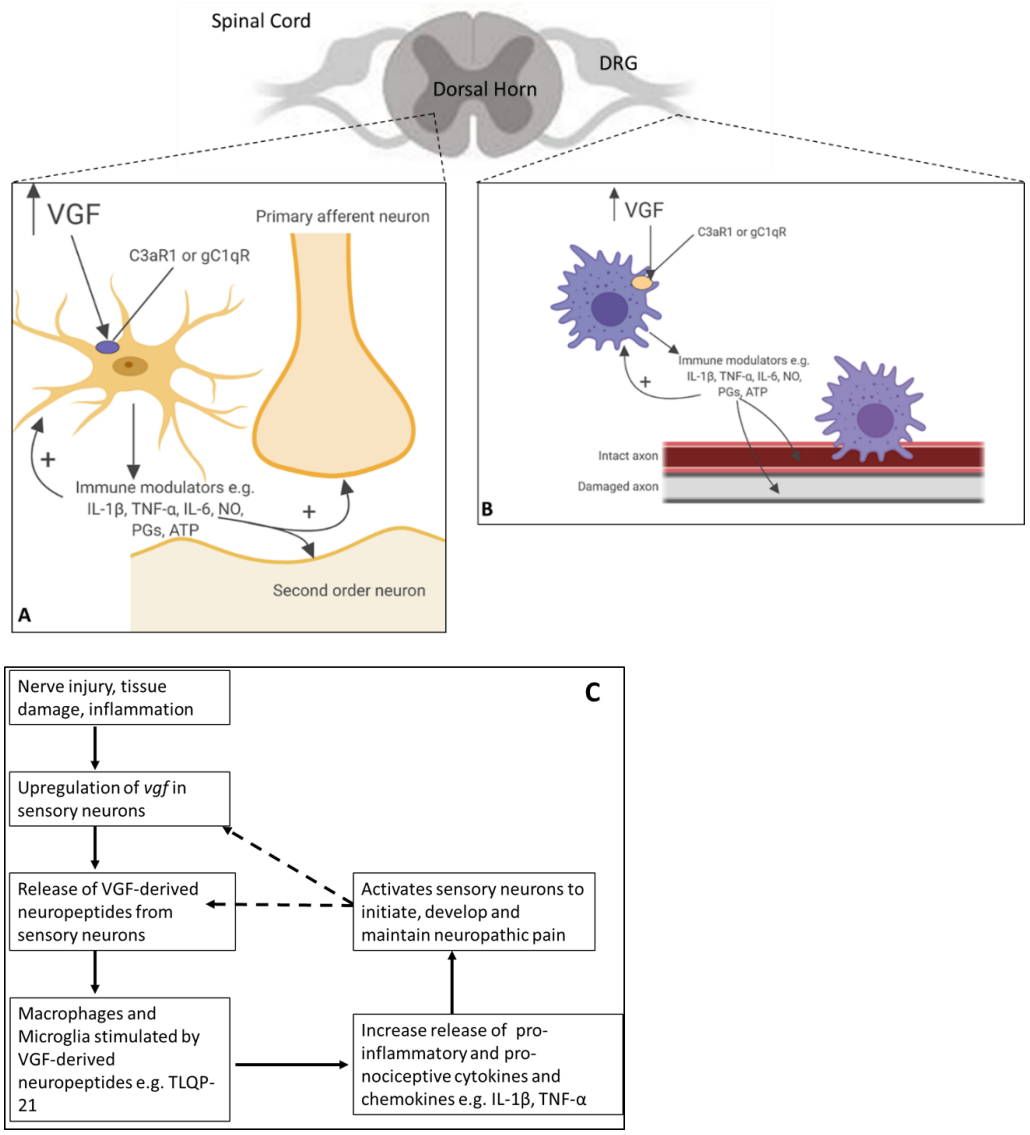


Figure 3.3. The role of VGF-derived neuropeptides in (A) central and (B) peripheral sensitisation and (C) neuropathic pain. (A) Microglia in the spinal dorsal horn are activated by VGF-derived neuropeptides following nerve injury through their interaction with either or potentially both the C3a1 or gC1q receptors. In turn, these activated microglia release several pro-inflammatory cytokines, chemokines and other agents which interact with neurons associated with pain transmission and increase the excitability of the neurons contributing to NP. Pro-inflammatory mediators released e.g., TNFα, IL-1β, IL-6, nitric oxide (NO), ATP and prostaglandins (PGs), initiate a self-propagating mechanism of enhanced cytokine expression by microglial cells. This leads to an increase in intracellular calcium, and activation of the p38 and MAPK pathway. (B) Damaged and spared fibres are pictured together. After peripheral nerve injury, non-neuronal cells including macrophages accumulate around the damaged cells. The upregulation and release of VGF-derived neuropeptides activate the macrophages through their interaction with either or potentially both the C3a1 or gC1qR receptors. The macrophages secrete factors (such as TNFα, IL-1β, IL-6, chemokine (C–C motif) ligand 2 (CCL2), PGs and nerve growth factor (NGF)) which elicit peripheral sensitisation and cause chronic neuroinflammation which maintains sensory abnormalities. (A) and (B) were created with Biorender.com. (C) The graph abstract highlights the potential role of *vgf* and VGF-derived peptides in the initiation, development, and maintenance of neuropathic pain.

In summary, peripheral tissue injury is associated with changes in protein expression in sensory neurons that may contribute to abnormal nociceptive processing. The publications within this review have focused on exploring the role of VGF-derived peptides *in vivo* but the expression of VGF in NP patients has not been characterised, summarised in Table 3.7. The presumed abundant and selective expression of the VGF-derived peptides in blood and CSF suggest a possibility that they could be utilised as biomarkers of NP. The understanding of the molecular mechanisms and signalling events by which VGF-derived active peptides exert their many physiological actions is in its infancy. Future work should aim to have a better understanding of the downstream consequences of cell treatment with TLQP-21 which should uncover proteomic changes in TLQP-21 treated cells and intracellular mechanisms of TLQP-21 actions and subsequent understanding may offer therapeutic strategies for NP. The identification of the two complement receptors with natural ligands of which little is understood suggest the possibility of a dual ligand mechanism of action.

Table 3.7. Summary to date of the identification and validation of VGF as a biomarker for neuropathic pain. The identification and correlation to disease and pharmacodynamics have only be confirmed in animal models.

Pain Disease state	Biomarker	Correlation to disease	Correlation with pharmacodynamic outcome	Correlation with pain state	Clinical efficacy
Neuropathic pain (Nerve injury animal models)	VGF expression	Elevated VGF	Antagonism of VGF-derived neuropeptides	No	No

3.7. Limitations

There are several methodological limitations. Firstly, the systematic search was conducted using one electronic database, PubMed. Searching multiple electronic databases is in line with common guidance for clinical systematic reviews (Lefebvre et al., 2008) and accounts for how databases differ in their coverage of journals and the indexing of studies. Searching multiple databases also mitigates the risk of selection bias thereby increasing robustness of the overall results. However, it has been suggested that performing searches of multiple databases has a modest impact on the results of systematic reviews of therapeutic interventions (Halladay et al., 2015, Hartling et al., 2016) yet how applicable these findings are to preclinical literature is not clear. Secondly, due to time and resource constraints the study selection and data extraction stages of this review were only conducted by a single reviewer (N.S.). If conforming to systematic review standards, to mitigate bias and ensure accuracy, those stages of the review should be conducted by two independent researchers and any disagreements reconciled by a third independent reviewer. It should also be noted that only what is reported within the included studies can be assessed. When considering gene and mRNA expression there is a wealth of data in open-source repositories including independently replicated

transcriptional information from RNA-sequencing studies (Crow and Denk, 2019). Therefore, depending on the research question it may be pertinent to use alternate sources of information i.e., databases and not rely only on studies identified in the systematic search. Lastly, the rate of publication of new primary research means that reviews can become outdated quickly; in the two years between the study being published and now, one study has been published, therefore the current rate of publication is manageable for researchers to maintain currency. The instigation of a living search will allow for the identification of new relevant literature and is the first step towards a living systematic review, a review that is continually updated as new evidence is published (Elliott et al., 2017). Given the spectrum of review types, this review could also be considered as a qualitative systematic review, because the way in which the narrative synthesis has been conducted is in line with the Synthesis Without Meta-analysis (SWiM) guideline (Campbell et al., 2020). The guidelines were published after this review and cover key features of synthesis including how studies are grouped, synthesis methods used, presentation of data and summary text and limitations of the synthesis. However, to mitigate bias, all analysis plans should be stated *a priori* in a registered protocol.

3.7.1. *Animal models of neuropathic pain (predictive value) and behavioural outcomes*

Animal experimentation has contributed extensively to our understanding of mechanisms of disease and development of novel therapeutics however, their predictive value of treatment effectiveness in humans remains controversial. Despite promising preclinical evidence, the lack of success in the clinic suggests there are limitations in their translatability (Perel et al., 2007). Pain is a subjective, multifaceted symptom which can only be measured by self-reporting (Vierck et al., 2008) and the presence or absence of pain in animal models cannot be directly measured and can only be inferred from the observation of surrogate behaviours (du Sert and Rice, 2014). In the assessment of VGF-derived peptide expression, SNI is the most commonly used model (5 studies). NP is highly heterogeneous, both within and across underlying conditions and the predominance of traumatic nerve injury models does not match the clinical situation (Finnerup et al., 2005). A more rational approach is to select models that more closely reflect the pathophysiological condition of humans. Hence, disease specific models have been developed including models of VZV infection (Fleetwood-Walker et al., 1999, Garry et al., 2005, Hasnie et al., 2007) and HIV-associated peripheral neuropathy (Wallace et al., 2007b, Wallace et al., 2007a). To add, reflex withdrawal assessments do not necessarily measure global pain and therefore can be misleading, if not just a measure of nociception. Predictive validity may also be improved by measuring changes in behaviours which are ethologically relevant to rodents and may be affected by pain e.g. rearing (Matson et al., 2007), feeding (Stevenson et al., 2006) and burrowing (Andrews et al., 2012).

3.7.2. *Internal Validity*

Characteristic of the field, the lack of reporting of methods undertaken to reduce the risk of bias suggests that the *in vivo* data reported in the studies within this review may be susceptible to bias. This review only determined the prevalence of reporting of methodological quality criteria and not how experimental bias

was mitigated. Experimental bias is often unintentional and can be because of low internal validity leading the scientist to incorrectly attribute an observed effect to a treatment or intervention (Higgins et al., 2011). SYRCLE's risk of bias assessment (Hooijmans et al., 2014) is designed to facilitate and improve the critical appraisal of evidence from animal studies. It provides the framework to assess the experimental bias mitigation methods researchers report and therefore provides a more robust means to improve the methodological quality of animal studies. Internal validity ensures that changes observed in outcomes are due to an induced change in the independent variable rather than confounding factors. The internal validity may be compromised by a range of biases: selection, performance, detection, and attrition bias. There are several mitigations that will reduce the risk of these biases; randomisation, allocation concealment, sample size calculation, blinded assessment of outcome and a pre-determined animal eligibility criterion. In addition, it is worth noting the small sample sizes reported. Small sample sizes are associated with low statistical power, inflated false discovery rate, inflated effect size estimation and low reproducibility (Colquhoun, 2014). It is likely that the authors would describe these studies as exploratory and therefore confirmatory studies for which the hypothesis is stated a priori, and the studies are adequately powered and use rigorous methodology are required (Kimmelman et al., 2014).

Several systematic reviews and meta-analyses have provided empirical evidence highlighting that inadequate experimental approaches are associated with bias in several preclinical fields (Hirst et al., 2014, Crossley et al., 2008, Rooke et al., 2011). Of concern, is that low prevalence of reporting of measures to reduce the risk of bias tend to give higher estimates of treatment effects (Crossley et al., 2008, Hirst et al., 2014). The introduction of the ARRIVE reporting guidelines in 2010 (Kilkenny et al., 2010a, McGrath et al., 2010, Kilkenny et al., 2010b) and the development of National Centre for the Replacement, Refinement and Reduction of Animals in Research Experimental Design Assistant gives pre-clinical researchers clear guidance on how to conduct animal experiments with appropriate rigour (du Sert et al., 2017).

In addition, as is often the case, the publications included in the review may be susceptible to publication bias. TLQP-21 is the most frequently researched VGF-derived peptide, particularly in recent years. It was not possible to estimate publication bias due to the low number of studies. However, publication bias deprives researchers of accurate data that is needed to generate new hypotheses and prevent a waste of resources in the case that a direction of research is chosen which has already been fully exploited but not published due to neutral or negative data. Thus, there is a critical and ethical need for transparency of reporting all experimental details to complete the story (Sena and Macleod, 2007).

3.8. Future Research Questions

- Characterisation of *vgf* upregulation and expression of VGF-derived neuropeptides in neuropathic pain patients to be able to assess utility as a biomarker
- Further elucidation of the molecular mechanisms and signalling events exerted by VGF-derived peptides. Including:

- Downstream consequences of cell treatment with TLQP-21 to uncover proteomic changes in TLQP-21 treated cells and the intracellular mechanisms of TLQP-21
- The role of the complement receptors and their natural ligands

3.9. Conclusion

Pain biomarkers have the potential to improve analgesic drug research and development and improve the prevention, diagnosis, and treatment of chronic pain patients. A persistent challenge in the management of NP is to target the specific mechanisms leading to a change from normal to abnormal sensory perception while ensuring that the defensive pain perception remains intact. Targeting VGF-derived peptides may offer this opportunity. The up-regulation of *vgf* and increased expression of VGF mRNA following nerve injury or inflammation suggests that VGF-derived neuropeptides have a role in enhancing nociception and present a nociceptive biomarker(s). Because of its rapid and then sustained up-regulation it also offers potential as a preventive and diagnostic biomarker and may have a role to play in the transition to the development and maintenance of chronic pain. The transition is considered a critical determining event and therefore identification of a biomarker at this stage offers the opportunity for therapeutic reversal to prevent the development of chronic pain. There are both central and peripheral mechanisms to VGF-derived peptide activity, yet it is not clear whether the mechanisms are distinct and/or have different roles in the initiation, development, and maintenance of NP. VGF expression levels are very low under normal physiological conditions suggesting that it does not play a role in nociception but its rapid upregulation in sensory neurons after nerve injury and inflammation along with the activation of microglia and macrophages may in part be responsible for the onset, development, and maintenance of neuropathic and inflammatory pain.

Pain biomarkers are offering the promise of improving analgesic drug development to bring new safe and effective treatments to patients sooner. For this promise to be realised molecules and mechanisms must be assessed and carefully validated in well-characterised patient populations. For that, as more advancements to the knowledge base are made, we, as a research community, need to be able to synthesise emerging evidence in an efficient manner, concurrent to its production to produce robust, contextualised understanding upon which we can build. Using systematic review methods in support of a narrative review is one approach for ensuring that all the relevant literature is appraised in an unbiased manner while providing current state of the art and science, critical analysis, and in-depth insights in the form of a traditional narrative review.

4. Cannabis & Pain Preclinical Systematic Review - Part 1: Mechanisms of a Systematic Review

4.1. Introduction

In 2018, The International Association for the Study of Pain (IASP) established a Presidential Task Force (TF) on Cannabis and Cannabinoid Analgesia to comprehensively investigate the evidence for the use of cannabinoids and cannabis-based medicines for pain management. This provided the opportunity to assess methods aimed to improve the feasibility and efficiency of the preclinical review process.

The TF was comprised of 4 work packages (WP) which assessed the (1) basic science, including medicinal chemistry, compound classification, pharmacology and nociceptive efficacy assessed in preclinical models (Finn et al., 2021, Soliman et al., 2021, Soliman et al., 2019a); (2) evidence synthesis of clinical efficacy which included a systematic review of reviews to date (Fisher et al., 2020, Moore et al., 2020); (3) evidence synthesis on harms (Mohiuddin et al., 2020); and (4) the social impact including changing policy and political practise (Haroutounian et al., 2020). A comprehensive research agenda and position statement was informed by the WP outputs (Haroutounian et al., 2021).

As part of WP1, it was my responsibility to lead the preclinical systematic review. The work package was comprised of senior scientists that did not have practical systematic review experience. Although my experience was limited, I had conducted a small preclinical systematic review of HIV and herpes-zoster associated neuropathic pain and had fostered good professional relationships with both CAMARADES and SYRCLE, leading groups in preclinical systematic review and referred to their resources for guidance.

The review employed machine learning for study selection and a crowd of scientists for study selection and annotation and data extraction stages of the review (both discussed in more detail in subsequent chapters). The review provided an overview of the available evidence of cannabinoid preclinical research. The anti-nociceptive effect of cannabinoids, cannabis-based medicines, and endocannabinoid system modulators assessed in animal models of pathological or injury-related persistent pain was quantified and an appreciation of the rigour of the data was provided. The empirical evidence was used to underpin practical recommendations for future research including the identification of knowledge gaps, areas to improve the internal and external validity of preclinical trials and factors influencing treatment efficacy (Soliman et al., 2021). The findings of the review will be discussed in detail in the next chapter.

The aim of this chapter is to discuss the project design, conduct and management. I will reflect on the project to share lessons learned pertinent to both the successful and timely completion of a large preclinical systematic review and how to foster successful collaborations. In addition, there are many decisions that are likely to impact the final conclusions of a systematic review, from the review design to the data preparation and analysis techniques. Here I will discuss the rationale for decisions made and how these decisions can

impact the overall results. It is therefore not structured in the usual format but the methods and results sections are combined and comprises the project timelines, collaboration, the online systematic review platform SyRF, and the stages of the systematic review process.

4.2. Methods & Results

4.2.1. Project Timelines

The TF convened in October 2018 with work starting in earnest in November 2018. The aim of the TF was to complete the full appraisal of the evidence base and present our findings at IASP's World Congress in June 2020, this therefore gave a projected timeframe of approximately 19 months for the preclinical systematic review. The project from inception to publication took 28 months; protocol development started in November 2018 and the manuscript was accepted for publication in March 2021. The timeline is depicted in Figure 4.1. The systematic search of the electronic databases was conducted on the 9 April 2019. The screening started on 20th April 2019 and took approximately 6 weeks to perform, the annotation and data extraction stages 26 weeks and reconciliation 7 weeks. Between screening and data extraction there was a 16 week delay due to necessary updates to the SyRF platform. There was also a delay between data extraction and reconciliation of 11 weeks while a platform was developed to facilitate the reconciliation process. Time on screening, annotation, data extraction and reconciliation was approximately 39 weeks (9 months), and 5 months for analyses and writing of the manuscript. Consequently, a systematic review and meta-analysis of this size requires approximately 22 months had it not been for the unavoidable and unpredicted delays for SyRF development. Further time savings could be made, in the protocol development and improved recruitment, training and contribution of crowd scientists, to achieve completion within the timeframe as initially set out. These timelines can be used as a planning estimates for future reviews.

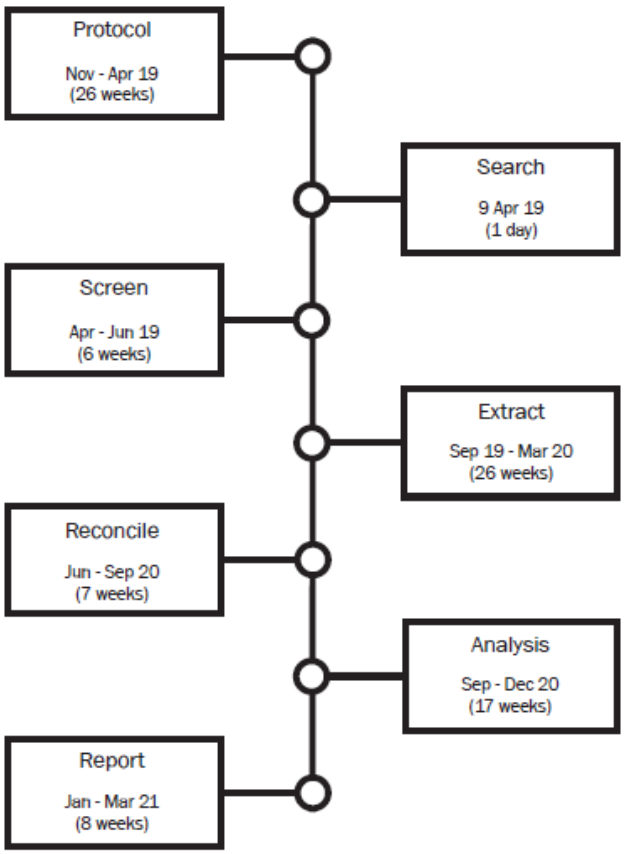


Figure 4.1. Timeline of the Cannabis & Pain Preclinical Systematic Review. Each stage is presented with approximate start and finish dates and number of weeks that each stage took to complete. Protocol; protocol development; Screen; manual screening of 10,816 articles retrieved in the search; Extract; annotation and data extraction of the 473 included studies; Reconcile; process of identifying agreement and reconciling disagreement between two independent reviewers; Analysis; data preparation and synthesis and writing of the original manuscript; Report; first submission, peer review and acceptance in the journal PAIN.

4.2.2. Collaboration & Communication

The success of any project is dependent on fostering a collaborative working group that effectively brings together the expertise of many disciplines to collectively meet the aims of the project. I led a large multi-disciplinary effort, many were senior scientists from multiple institutions across the globe. The research team consisted of the core 6 WP1 members but evolved throughout the review as subject matter expertise was sought at critical stages. The main author team consisted of senior scientists of pharmacological clinical and preclinical research expertise. They provided invaluable insights into the pharmacology of the interventions and the characteristics of experiments designed to assess their anti-nociceptive efficacy. This was critical to decisions surrounding how to categorise and perform the subgroup analyses. However, at times I needed methodological guidance and through the course of the project a further 7 scientists were consulted in their capacity to provide machine learning support, software development and SyRF platform user support, and meta-analytical guidance, plus there were the 100 crowd scientists that contributed to study selection and annotation and data extraction stages of the review. Effective communication was paramount; I wanted to

foster an open learning culture and would always feedback the guidance I received to the WP1 members. This allowed us to have many insightful discussions about the processes and how we would meet the aims of the review. The scientists of the main author team represented the primary researcher, and I was a conduit between them and our systematic review experts. The conversations highlighted disconnects between the aims, understanding and priorities of both groups. The primary researchers wanted to identify which interventions had the greatest anti-nociceptive potential and for which indications, and therefore were concerned primarily with study characteristics. They also expressed concern with the accessibility of meta-analysis for primary researchers including that primary researchers are likely to require more assistance to understand the methods and how to interpret the results. The systematic reviewers on the other hand tended to focus more on justifying the methods used, internal validity and used meta-analytical terms to describe their results. These differences may present a barrier to the uptake of preclinical systematic reviews. A bridge was to be built and a balance to be struck between the two groups: to improve accessibility of the review I included a glossary to provide explanations of the key systematic review and meta-analysis terms (Table 4.1).

Table 4.1. Glossary of terms used in the systematic review and meta-analysis of cannabinoids, cannabis-based medicines and endocannabinoid system modulators assessed for antinociceptive efficacy in animal models of pathological or injury-related persistent pain (Soliman et al., 2021).

Systematic Review	Use predefined methods to identify, select and critically appraise all available literature to address a specific research question
Meta-analysis	The statistical combination of quantitative results (pain-associated behavioural outcomes) of two or more studies. The methods included in the meta-analysis are the calculation of effect sizes, the pooling of the effects so that the range and distribution of effects can be observed
Study	In this instance a study refers to the publication. A publication can have multiple experiments in which an intervention is tested in a cohort of animals and a pain-associated behaviour is measured. There can be multiple outcome measures per cohort.
Comparison and nested comparison	The outcome measure of a treatment group compared to a control (vehicle-treated) group is a comparison. Often the same cohort of animals undergo multiple pain-associated behavioural outcome measurements. In these instances, the comparisons are combined to give one outcome statistic (a nested comparison) which represents the global measure of the outcomes in that comparison.
Effect size	For each comparison, an effect size is calculated using standardised mean difference (SMD). The difference between group means (mean of control group – mean of experimental group) is divided by the pooled variance which converts all outcome measures into a standardised scale. (A correction factor, 1 or -1 is used to define the direction of the effect size, whether the outcome is better or worse in comparison to the control).

Heterogeneity	Study heterogeneity denotes the variability in outcomes that are not due to measurement errors but other influencing factors (e.g. study characteristics). We have estimated heterogeneity using both Cochran's Q and I^2 and explored sources of heterogeneity with stratified meta-analysis.
Estimating heterogeneity with Cochran's Q	Q is an estimate of between-study heterogeneity and is calculated from effect sizes. It is based on a chi-squared distribution. A larger Q value denotes larger variation across studies rather than within subjects within a study. The p value of Q is used to indicate the presence or absence of heterogeneity.
Estimating heterogeneity with I^2	I^2 is the proportion of total variance between studies that is due to true differences in effect sizes, not differences that are due to chance. If $I^2 = 0$ % all variation is due to chance alone, 100% all variation is due to differences between the true effect sizes between studies. 0 – 25 % - very low heterogeneity 25 – 50 % - low heterogeneity 50 – 75 % - moderate heterogeneity >75 % - high heterogeneity
Stratified analysis	Studies which share a particular characteristic e.g. sex, strain, animal model will be more similar than studies that do not share the same characteristic. Stratified analysis allows us to partition the heterogeneity between groups of similar studies and between groups of studies to determine whether the differences are statistically significant.
Animal model	Whole in vivo animal models of pathological or injury-related persistent pain e.g. tissue injury, cancer, chemotherapy-induced, inflammation, or nerve damage. Persistent pain was defined as studied over a period of hours, days, weeks, or months.
Pain-associated behavioural outcome	These were when pain was declared the reason for assessment by the authors. Behavioural outcomes include: • Evoked limb withdrawal to mechanical, heat or cold stimuli • Spontaneous e.g. weight bearing difference, spontaneous foot lifting, grimace scale, nocifensive behaviour) • Complex e.g. open field test (thigmotaxis) and burrowing
Antinociception	Attenuation of pain-associated behaviour

It came to light that effect sizes are not widely understood so I included a figure and table to visualise the overlap between control and treatment groups effect size distributions to better illustrate the calculated effect sizes (Figure 4.2 & Table 4.2). Also, to tailor further towards the primary researcher, less focus was placed on internal validity and the strengths and limitations of the systematic review methods. The risk of bias assessment was placed at the end of the results and discussion sections because it was felt that it is known that the conduct and reporting of methodological rigour needs to improve. A compromise given that internal validity underpins the reliability of the data (Soliman et al., 2021).

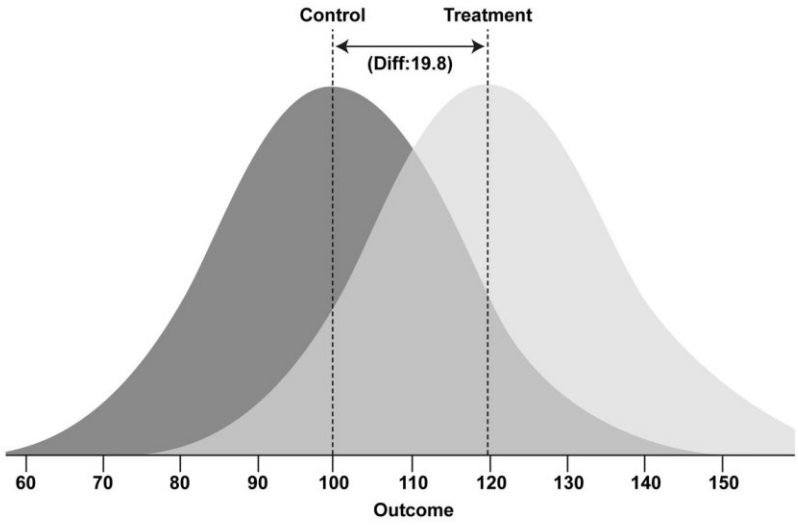


Figure 4.2. Visualisation of the overlap between control and treatment group distributions of the overall SMD effect size of 1.32 (Magnusson, 2020, Soliman et al., 2021).

Animal studies yield larger effect sizes than clinical studies. The challenge of how best to compare effect sizes between clinical and preclinical meta-analyses was therefore also discussed at length. It is to be expected that animal studies will yield larger effect sizes in comparison to clinical studies due to the more homogenous nature of the study population reducing the observed variance. It is not possible to directly compare the effect sizes from preclinical and clinical trials. In clinical settings, because there is a very broad natural heterogeneity of data, effect sizes are smaller whereas in animal experiments, where heterogeneity and variance are reduced, yield larger effect sizes. There is no scale for preclinical effect sizes to accurately determine small, medium, large effect (which is arbitrary)(Cohen, 1962), nor are there reliable techniques to compare animal and clinical effect sizes given the inherent differences.

Table 4.2. Differences and overlap between treatment and control groups for drug classes assessed for antinociceptive efficacy in rats (A) and mice (B) which corresponds to the forest plots of figures 3 and 4 in the publication, figure 5.4 in the thesis. N, number of animals (Magnusson, 2020, Soliman et al., 2021).

A Drug Class	Effect Size	% of treatment group with larger mean than control group	% overlap	N
NAAA inhibitor	1.59	94.4	42.7	78
Unknown mechanism of action	1.50	93.3	45.3	32
CB_2 receptor agonist	1.48	93.1	45.9	1671
THC	1.47	92.9	46.2	476
CB_1 receptor inverse agonist	1.44	92.5	47.2	131
PPAR-alpha agonist	1.40	91.9	48.4	1105
Non-selective cannabinoid receptor agonist	1.37	91.5	49.3	1379
Anandamide transport inhibitor	1.32	90.7	50.9	497
Monoacylglycerol lipase inhibitor	1.31	90.5	51.2	151
FAAH inhibitor	1.21	88.7	54.5	1270
CB_1 receptor agonist	1.18	88.1	55.5	2021
CBD	1.12	86.9	57.5	383
CB_2 receptor inverse agonist	0.96	83.1	63.1	8
FABP inhibitor	0.87	80.8	66.4	12
Hemp oil	0.09	53.6	96.4	12
GPR55 agonist	-0.38	35.2	84.9	8
PPAR-gamma antagonist	-0.72	23.6	71.9	26

B Drug Class	Effect Size	% of treatment group with larger mean than control group	% overlap	N
FAAH inhibitor/TRPA1 agonist	3.52	100	7.84	24
NAAA inhibitor	3.23	99.9	10.6	110
CB_1 receptor inverse agonist	2.69	99.6	17.9	66
PPAR-alpha agonist	2.60	99.5	19.4	508
Diacylglycerol lipase inhibitor	1.83	96.6	36	138
FAAH inhibitor	1.56	94.1	43.5	1058
Unknown mechanism of action	1.56	94.1	43.5	183
Monoacylglycerol lipase inhibitor	1.53	93.7	44.4	457
Dual FAAH/MAGL inhibitor	1.42	92.2	47.8	84
THC	1.39	91.8	48.7	283
CB_1 receptor agonist	1.38	91.6	49	887
CB_1 receptor PAM	1.08	86	58.9	61
CB_2 receptor inverse agonist	1.05	85.3	60	48
CBD	1.05	85.3	60	230
Non-selective cannabinoid receptor agonist	1.03	84.8	60.7	815
CB_2 receptor agonist	1.03	84.8	60.7	1448
Anandamide transport inhibitor	0.82	79.4	68.2	146
FABP inhibitor	0.65	74.2	74.5	252
ABHD6 inhibitor	0.27	60.6	89.3	33
FAAH inhibitor/TRPV1 agonist	-0.39	34.8	84.5	45

Faggion et al. (2017) illustrate that there is a generation of scholars skilled in the technical mechanisms of systematic reviews and critics may assume that they override and make redundant those with in depth understanding. However, it is my experience, that in the context of preclinical systematic reviews, subject matter expertise is required. An aim of many preclinical systematic reviews is to provide empirical evidence to inform a future research agenda: the target audience of these reviews are preclinical researchers. It is therefore imperative that the preclinical researchers are included in the review process, so that they can provide in-depth understanding and guidance on the needs and challenges of their community and ensure that findings are communicated in an accessible manner. I have previously discussed that limited uptake of preclinical systematic reviews may be a consequence of technical challenges however, the full potential of systematic reviews may yet to be realised by primary researchers. To encourage greater uptake collaboration between preclinical primary researchers and systematic reviewers is key so that the full utility of preclinical systematic reviews can be demonstrated.

4.2.3. The CAMARADES Systematic Review Facility (SyRF)

The review was conducted using the online systematic review facility (SyRF) an open-source platform designed to support the conduct of preclinical systematic reviews launched in 2016 by the Collaborative Approach to Meta-Analysis and Review of Animal Data from Experimental Studies (CAMARADES). We have recently published an article describing the development and uptake of SyRF which provides a detailed description of the architecture and its features (Bahor et al., 2021). Briefly, it is flexible to support a wide variety of experimental designs and can be adapted to different research questions. It provides the option to collaborate easily with many reviewers, efficiently manage the workflow of a project, screen, annotate, and extract outcome data into a secure database. Each project lead must design their own annotation and data extraction question forms, tailoring each step to the needs of their project. Beyond a strong collaborative relationship with CAMARADES, there were two main reasons for choosing to use SyRF; firstly, it is specifically designed for preclinical systematic reviews and has the flexibility for the incorporation of different annotation and data extraction types as well as for multiple models, treatments, and outcome assessments that preclinical studies are often characterised by. Secondly, it is easy to coordinate the contribution of many reviewers (crowd scientists).

SyRF was released with minimum features and is under continued development. Throughout I worked closely with the CAMARADES lead developer (Chris Sena; CS) to improve the functionality of the platform. During the building and testing of the annotation and data extraction form on SyRF I identified several issues that needed to be rectified before reviewers could contribute. Examples include missing core questions e.g., variance in the outcome tab and sample size in the cohort tab, the guidance notes were not appearing when hovering over a question and the features of the experiment tab were not properly enabled so the table for entering outcome data was not available. I also discovered a security issue; anyone could change the annotation questions thereby compromising, potentially corrupting, and undermining any review.

There were other features that are not yet developed for which I needed to identify a work around. Firstly, it is not possible for a reviewer to return to a study that they had started reviewing so it would be left incomplete, or the reviewer would be unable to rectify a mistake. To work around this, reviewers needed to save the URL to the study to return to it for completion or to rectify any mistakes. Secondly, a related issue was that when reviewers opened a new study to review, SyRF would allocate this study to that reviewer. However, reviewers may not have intended to review that study and/or if they did, they had no way of returning to it. This meant that studies were being held by reviewers, not being reviewed and the study could not be assigned to a new reviewer. To prevent studies being held unreviewed, we introduced a 'cap' to prevent reviewers starting too many studies and force them to contact me for the URL to review the studies that they had 'started'.

I describe below my use of the SyRF platform for the study selection, annotation and data extraction and reconciliation stages.

4.2.4. Systematic Review Methods

A protocol was developed in which the research questions were clearly defined the methods of the review articulated thereby mitigating the potential for introducing bias and ensuring transparency. The protocol was registered on PROSPERO (CRD42019124804) and published in PAIN Reports (Soliman et al., 2019a).

4.2.4.1. Systematic Search

Studies that assessed the antinociceptive efficacy of cannabinoids, cannabis-based medicines, endocannabinoid system modulators were identified through searching three electronic databases: Embase, PubMed and Web of Science on 9th April 2019. The step-by-step guide by Leenaars et al. (2012) was followed to transform the research question into a search strategy. Firstly, the research question was split into the critical search components, 'pain', 'cannabinoids', and 'animals' (Figure 4.3).

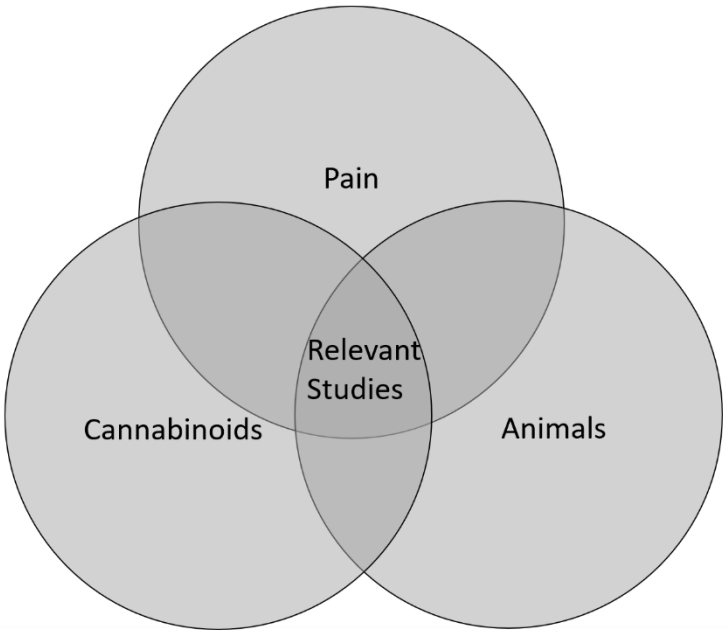


Figure 4.3. A Venn diagram representing the three different components of the search strategy. The overlap provides the potentially relevant studies.

For both the 'pain' and 'cannabinoids' components the standardised subject terms and free-text terms were identified and combined using the Boolean operator 'AND' (Table 4.3). Individual search strategies were constructed for each of the databases; the full search strategy can be found on the project's osf page). The results of each of the database searches were exported to Endnote 7.0, and duplicates identified and removed. The titles and abstracts of the identified studies were then uploaded to the SyRF platform for screening for inclusion.

Table 4.3. The general search terms used to identify the relevant literature.

Pain	Pain OR Hyperalgesia OR pain OR analgesia OR analgesic OR analgesics OR allodynia OR neuralgia OR hypersensitivity OR hyperalgesia OR hyperalgesic OR antinociception OR anti-nociception OR hypoalgesia OR hypoalgesic OR anti-hyperalgesia OR antihyperalgesia OR antihyperalgesic OR anti-hyperalgesic OR anti-allodynic OR antiallodynic OR anti-allodynia OR antiallodynia
AND Cannabinoids	Cannabinoids OR cannabis OR marijuana OR marihuana OR hemp OR hashish OR cannabinoid OR cannabinoids OR cannabidiol OR tetrahydrocannabinol OR "endocannabinoid modulator" OR FAAH OR MGL OR MAGL OR ABHD6 OR ABHD12 OR "fatty acid binding protein" OR NAAA OR endocannabinoid OR endocannabinoids OR endo-cannabinoid OR FAAH inhibitor OR FAAH inhibition OR MAGL inhibitor OR MAGL inhibition OR MGL inhibitor OR MGL inhibition OR

	anandamide transport inhibitor OR anandamide transport inhibition OR "ABHD6 inhibitor" OR "ABHD6 inhibition" OR "ABHD12 inhibitor" OR "ABHD12 inhibition" OR NAAA inhibitor OR NAAA inhibition OR "Fatty acid Binding Protein inhibitor" OR "fatty acid binding protein inhibition" OR FABP inhibition OR FABP inhibitor OR allosteric modulator OR "endocannabinoid modulators" OR "endo-cannabinoid modulators" OR "endo-cannabinoid modulator" OR FAAH inhibitors OR MAGL inhibitors OR MGL inhibitors OR anandamide transport inhibitors OR "ABHD6 inhibitors" OR "ABHD12 inhibitors" OR NAAA inhibitors OR "Fatty acid Binding Protein inhibitors" OR FABP inhibitors OR allosteric modulators OR PEA OR palmitoylethanolamide
AND	Animal search filters (de Vries et al., 2014, Hooijmans et al., 2010)

The search terms were kept broad to ensure that relevant literature would not be missed. Similarly, no restrictions were placed on date published or language used and pain-associated behavioural outcome measures were not included in the search strategy due to the likelihood of inconsistent reporting in the title and abstract (Leenaars et al., 2012). Searching multiple electronic databases is in line with common guidance for clinical systematic reviews (Lefebvre et al., 2008) and accounts for how databases differ in their coverage of journals and the indexing of studies. Searching multiple databases also mitigates the risk of selection bias thereby increasing robustness of the overall results.

4.2.4.2. Study Selection

Studies were screened against the inclusion and exclusion criteria based upon the title and abstract of the identified articles (Figure 4.4). Screening was conducted by the crowd of scientists. Each study was reviewed by two independent reviewers and discrepancies were resolved by a third independent reviewer. The inclusion/exclusion criteria were displayed on the user interface to assist reviewers, in addition to further information given in their training material (Tables 4.4 & 4.5).

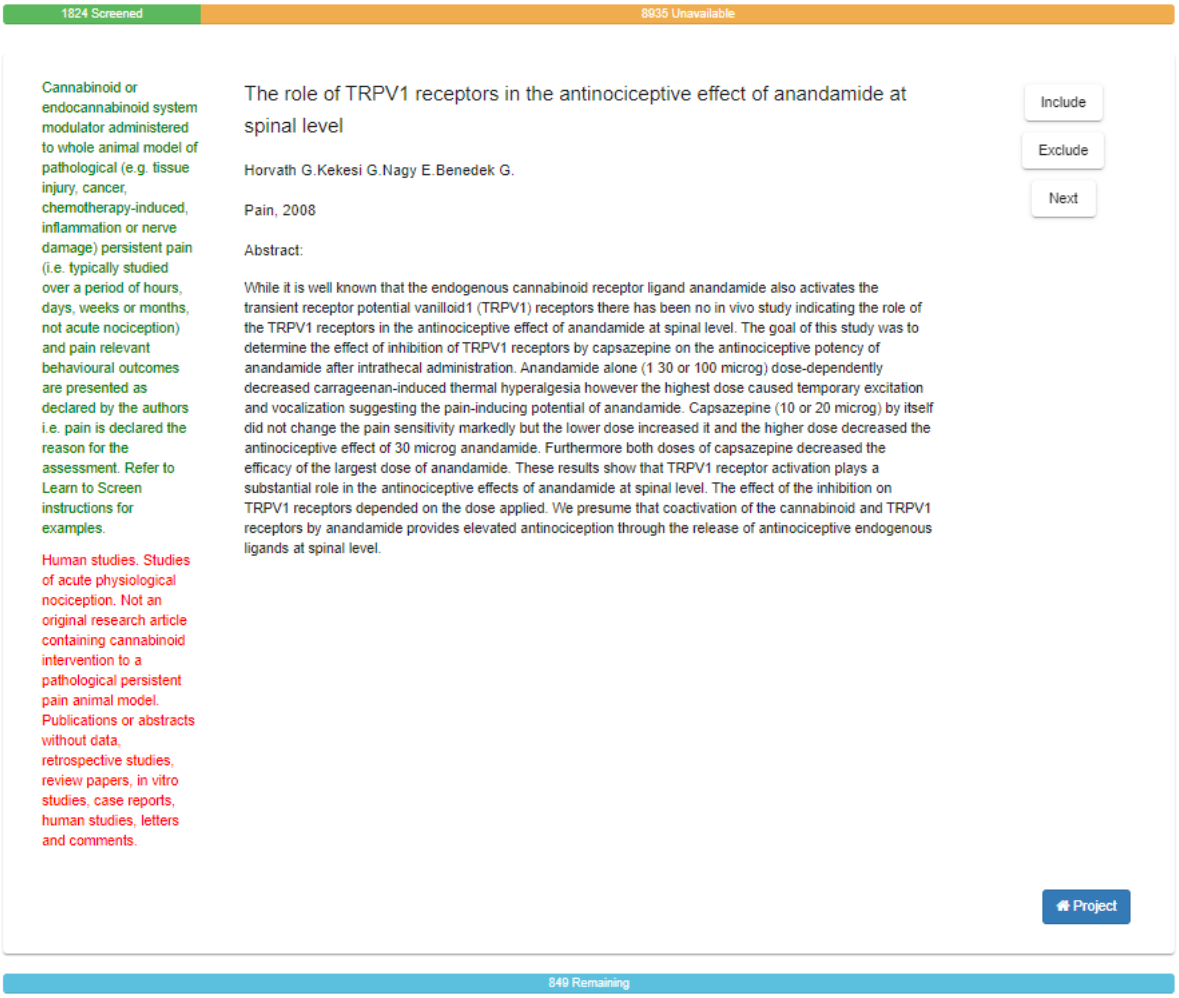


Figure 4.4. Screenshot of the SyRF user interface for study selection. On the left-hand side are the prompts for inclusion/exclusion criteria, the green writing denotes the inclusion criteria, the red writing the exclusion criteria. The title, authors, journal and date and abstract are presented in the middle with the decision buttons 'include', 'exclude', and 'next' on the right-hand side.

The inclusion criteria were any study which reported a cannabinoid, cannabis-based medicine or endocannabinoid system modulator (Table 4.4) administered to a whole animal model of pathological or injury related (e.g., tissue injury, cancer, chemotherapy-induced, inflammation or nerve damage; Table 4.5) persistent pain (i.e. typically studied over a period of hours, days, weeks or months) and presented pain relevant (as declared by the authors) behavioural outcome measures (Table 4.5). The exclusion criteria were studies of acute nociception or studies in which a pathological or injury-related model of persistent pain was not induced. Not an original or primary research article. Publications/abstracts without data, retrospective studies, reviews, in vitro studies, case reports, human studies, letters, and comments. Once title abstract screening was complete, the pdfs of the full texts were retrieved using Endnote and Imperial College London's subscription service.

Table 4.4. Terminology and Definitions (Soliman et al., 2019a).

Term	Definition	Examples/typical products
(Herbal) Cannabis	The whole plant or parts or material from the plant (e.g. flowers, buds, resin, leaves)	*Cannabis sativa*, hashish
Medical cannabis	The term 'medical cannabis' (or 'medical marijuana') is used for cannabis plants, plant material, or full plant extracts used for medical purposes.	**Bedrocan®, Bedrobinol®, Tilray 10THC/10CBD®**
Cannabinoids	Cannabinoids are biologically active constituents of cannabis, or synthetic compounds, usually having affinity for and activity at cannabinoid receptors.	THC, CBD, CP55,940, WIN55,212-2, HU210, nabilone, AM1241, AM1710, JWH133, LY2828360, MDA19
Phytocannabinoid	A cannabinoid found in the cannabis plant or purified/extracted from plant material	THC, CBD, CBDA
Endocannabinoid	An endogenous ligand found in the body of humans and other animals and which has affinity for, and activity at, cannabinoid receptors	Anandamide, 2-AG
Endocannabinoid system modulators	In addition to individual phytocannabinoids, cannabis-derived or cannabis-based medicines, and cannabis extracts, other pharmacological approaches under development for manipulation of the endocannabinoid system include selective synthetic cannabinoid receptor agonists or antagonists, inhibitors of the catabolism (e.g. fatty acid amide hydrolase [FAAH] inhibitors) transport (e.g. fatty acid binding protein [FABP] inhibitors) or reuptake of endocannabinoids or allosteric modulators of cannabinoid receptor signalling.	PF-04457845, URB597, URB937, AM404, VDM11, URB602, JZL184, ZCZ011, GAT211, rimonabant (SR141716A), AM251, SR144528, AM630
Cannabis-based (or cannabis-derived) medicines	Registered medicinal cannabis extracts with defined and standardized phytocannabinoid content, particularly THC and THC/CBD.	Nabiximols (Sativex®), dronabinol, marinol, Epidiolex®

NB: CBD: cannabidiol; FABP: fatty acid binding protein; THC: tetrahydrocannabinol; 2-AG: 2-arachidonoyl glycerol.

Table 4.5. Examples of animal models and pain-associated behavioural outcomes. The lists are not exhaustive.

Animal Model of Pathological Persistent Pain	Behavioural Outcomes
• Antiretroviral e.g. stavudine (d4T), 2,3-dideoxycytidine (ddc), zidovudine (AZT) • Carrageenan • Chemotherapy induced e.g. paclitaxel, oxaliplatin, cisplatin • Chronic constriction injury • Crush • Diabetes • Formalin • HIV protein • HIV+antiretroviral e.g. gp-120+ddc • Sciatic nerve ligation • Spinal cord injury • Spared nerve injury • Streptozoticin • Traumatic nerve injury • Varicella zoster virus	• Acetone test/ethylchoride • Burrowing • CatWalk • Cold plate • Conditioned or operant behaviour • Elevated plus maze • Grimace Scale • Hargreave's test • Hot plate • Open Field • Randall-Siletto • Rotarod • Tail immersion (cold/heat) • Tail flick (heat) • von-Frey electronic • von-Frey filament

It was an aspiration to use machine learning for the screening. However, following the manual screening of 4000 articles, the machine was not able to perform with high enough sensitivity and specificity, likely due to the low inclusion prevalence so all articles were manually screened. Following on, those human decisions were used to train the machine for error analysis (discussed in Chapter 7). The low prevalence of included studies can be an indication of an over inclusive search string. However, in this context it was particularly challenging to narrow the search terms further, not only due to the requirement to identify all relevant literature but because it was difficult to identify the population due to the challenge of differentiating studies in which an intervention(s) was tested for antinociception in healthy animals, animal models of acute pain or animal models of injury-related or pathological persistent pain (the desired population). Unlike the reporting of clinical randomised controlled trials there is large variation in the terms used and reporting of population, intervention, comparator, outcome (PICO) elements in the title and abstract of preclinical studies.

Additionally, the focus of preclinical studies and therefore what is reported in the abstract may not represent the full study and therefore relevant experimental data may still be reported in the study.

4.2.4.3. Annotation and Data Extraction

I built an annotation and data extraction form on SyRF to extract the following information from studies:

1. What the animal model of injury-related or pathological persistent pain was
2. What cannabinoids, cannabis-based medicine or endocannabinoid system modulator was tested for antinociceptive efficacy
3. What pain-associated behavioural assessment(s) was conducted and how did the treated animal perform to a suitable control

There are five main sections: study, disease model induction, treatment, outcome assessment, cohort, and experiment. I designed the questions required to extract the necessary data to answer the research questions and developed guidance for reviewers. It is a requirement that the questions be focused and unambiguous i.e., for which there can only be one answer if reported by the study authors. As much as possible I created the questions so that answers could be selected from drop down lists or yes/no buttons could be selected. E.g., Question: What was the reported sex of the animal? Select from list: Male, Female, Mixed, Not reported. Question: Do the authors report blinded assessment of outcome? Select Yes/No. This assisted in the uniformity of data extraction and the subsequent categorisation of the data for comparison throughout the analysis. It is also pertinent that only what is required and will be analysed is extracted to ensure efficiency. I provided a worked example of an included study to the reviewers, which included screenshots of the user interface and instructions of how to answer questions; presented in figures 4.5- 4.10. SyRF allows for the entry of data pertaining to multiple models (Figure 4.6), interventions (Figure 4.7), outcomes (Figure 4.8), cohorts (Figure 4.9), and experiments (Figure 4.10).

Annotation and data extraction of each study was performed by two independent reviewers and discrepancies reconciled by a third independent reviewer. Full text screening was incorporated into this stage. Reviewers were required to confirm whether the study met the inclusion criteria. For efficiency, they were then required to check that the study contained all the data pertinent to a meta-analysis e.g., the reporting of a pain-associated behavioural outcome measure, behavioural outcome data (average and variance), sample size and a suitable control. If the study was missing key information for meta-analysis, only the study level annotation questions were completed. Contextually important decisions had been made about what data would be extracted and how it would be analysed, as communicated in the protocol (Soliman et al., 2019a) and had to be incorporated into the annotation and data extraction form. These are explained in more detail here:

Firstly, at the study level, this review aimed to assess the efficacy of interventions but there were also important data relating to the pharmacokinetics and pharmacodynamics that should also be considered in

conjunction with anti-nociceptive effect. Although outside the scope of this review, it was important to gain an understanding of the prevalence of reporting of experiments conducted alongside behavioural outcome assessments including e.g., confirmation of the target, effects on motor activity, pharmacokinetic investigation including tissue concentrations and toxicity as well as assessment of side effects on anxiety, depression and dependency (Soliman et al., 2021).

As is the norm within systematic reviews, internal validity of studies was explored to assess the risk of bias and determine the validity of outcomes. Also at the study level, studies were scored against a 5 point methodological quality criteria, an adaptation of the CAMARADES checklist (Macleod et al., 2004) which included the following items: (1) Sample size calculation, (2) allocation concealment, (3) randomisation, (4) blinded of assessment of outcome, and (5) reporting of animal exclusions. This gave the prevalence of reporting of methodological quality criteria but is not in itself an assessment of risk of bias. To determine risk, it was necessary to collect the methods used to mitigate experimental bias. In accordance with the SYRCLE risk of bias tool (Hooijmans et al., 2014) based upon the reported method for each of the 5 methodological quality item risk was scored high, low, or unclear. A distinction between prevalence of reporting and risk of bias should be made.

With regards to experimental design, it was necessary to define the comparator. A suitable control was defined as an animal in which the model has been induced and administered an appropriate control treatment i.e., vehicle. An animal could not provide its own control e.g., contralateral for ipsilateral due to the possibility of contralateral sensory changes that could affect outcome measures. Similarly, sample sizes of animal cohorts are often reported as a range e.g., n=8-10 per group, as it is unclear the number of animals the more conservative estimate was used e.g., 8. Outcome data can be reported in text, in tables and in graphs. For data presented in graphs reviewers were required to use digital measuring tools e.g., WebPlotDigitizer to calibrate the axes and extract the outcome data values. In addition, for the extraction of outcome data presented as time series, the time point when the difference between treatment and control was largest was extracted. All doses of a drug tested were extracted regardless of whether they were reported as effective or not. Given the variation and propensity for studies to report different types of controls, sample size ranges, and different presentations of outcome data as well as other experimental design features like length of experimentation, these definitions ensured parity across the dataset and assisted in the accurate synthesis and interpretation of the data.



Figure 4.5. The study tab contains annotation questions that relate to the whole study e.g., the questions relating to the non-behavioural experimental procedures and risk of bias.

Disease Model

Information about the model can be found in the Methods section. *For this study information was extracted from the **Methods** section, paragraph **2.2. SCC paw model***

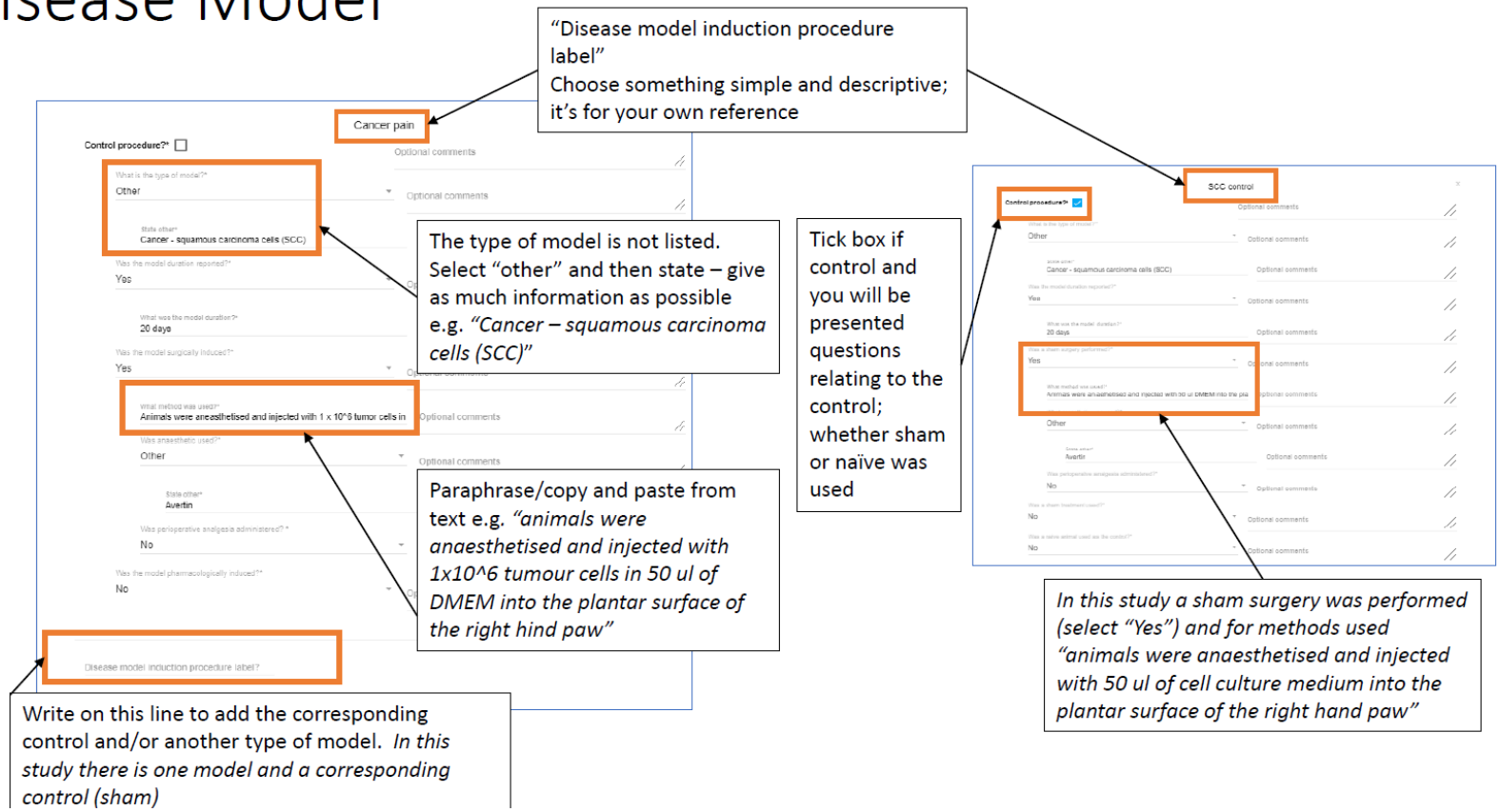


Figure 4.6. The disease model tab contained questions relating to the model induction including type of model and whether it was surgically or pharmacological induced.

Treatment

Information about the treatments can normally be found in the Methods section. *For this study information was extracted from the **Methods** section, paragraph **2.4. Win55,212-2 and AM1241 administration and pain behavioral testing.** 2 cannabinoids (WIN55,212-2 and AM1241) were tested against the same control (DMSO)*

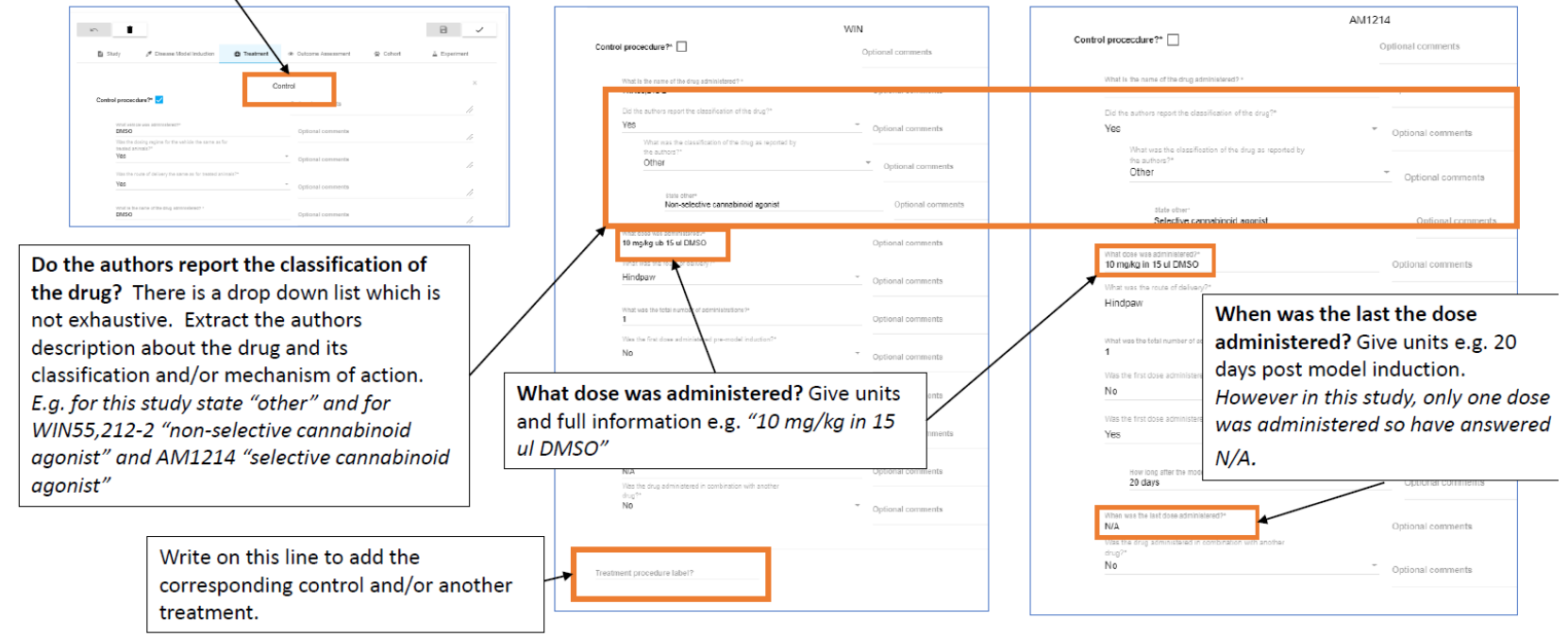


Figure 4.7. The treatment tab contained questions relating to the intervention and included intervention name, dose, route of administration and dosing regime.

Outcome Assessment

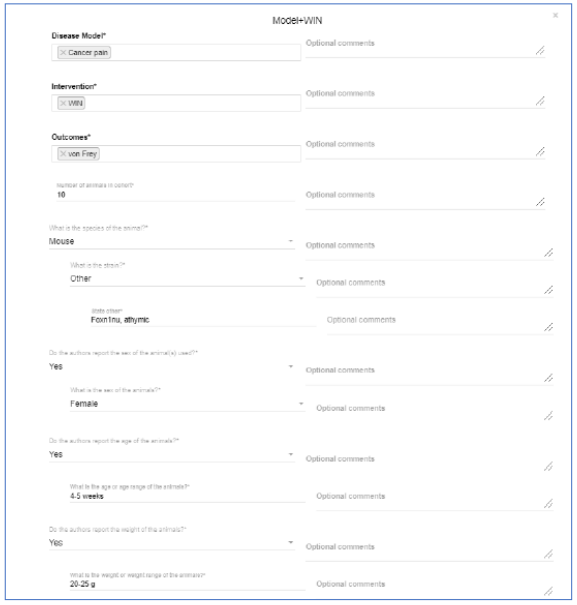
Information about the outcome assessments can normally be found in the Methods and Results sections. *For this study information was extracted from the **Methods** section, paragraph **2.3. Behavioural testing for the SCC paw model** and **Figure 2***



Figure 4.8. The outcome assessment tab contained questions relating to the pain-associated behavioural outcome assessment used including type of assessment, the structure of the data presented i.e., average and variance type, direction of effect and the method employed.

Cohort

Information about the cohorts can usually be found in the Methods section. *For this study information was extracted from the **Methods** section, paragraph **2.2. SCC paw model** and **Figure 2.***



Each cohort is a group of animals that have the same characteristics, are the same 'model' and have been given the same 'treatment'.
In this study, there are 4 cohorts;
sham animals,
model animals treated with WIN51,212-2,
model animals treated with AM1241 and
model animals treated with vehicle (DMSO)

Each must have their own form completed

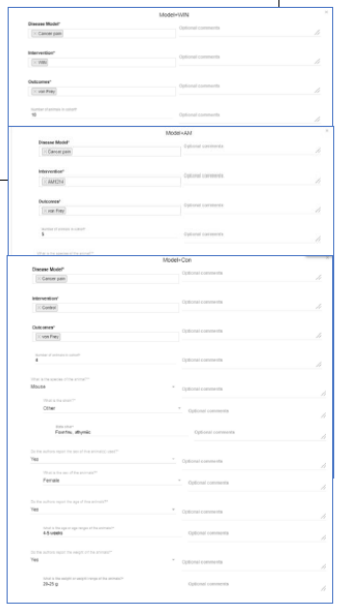


Figure 4.9. The cohort assessment tab contained questions relating to the animals' species, strain, sex, age, weight, and the sample size and combined the previously entered information; disease model, intervention, and outcome(s).

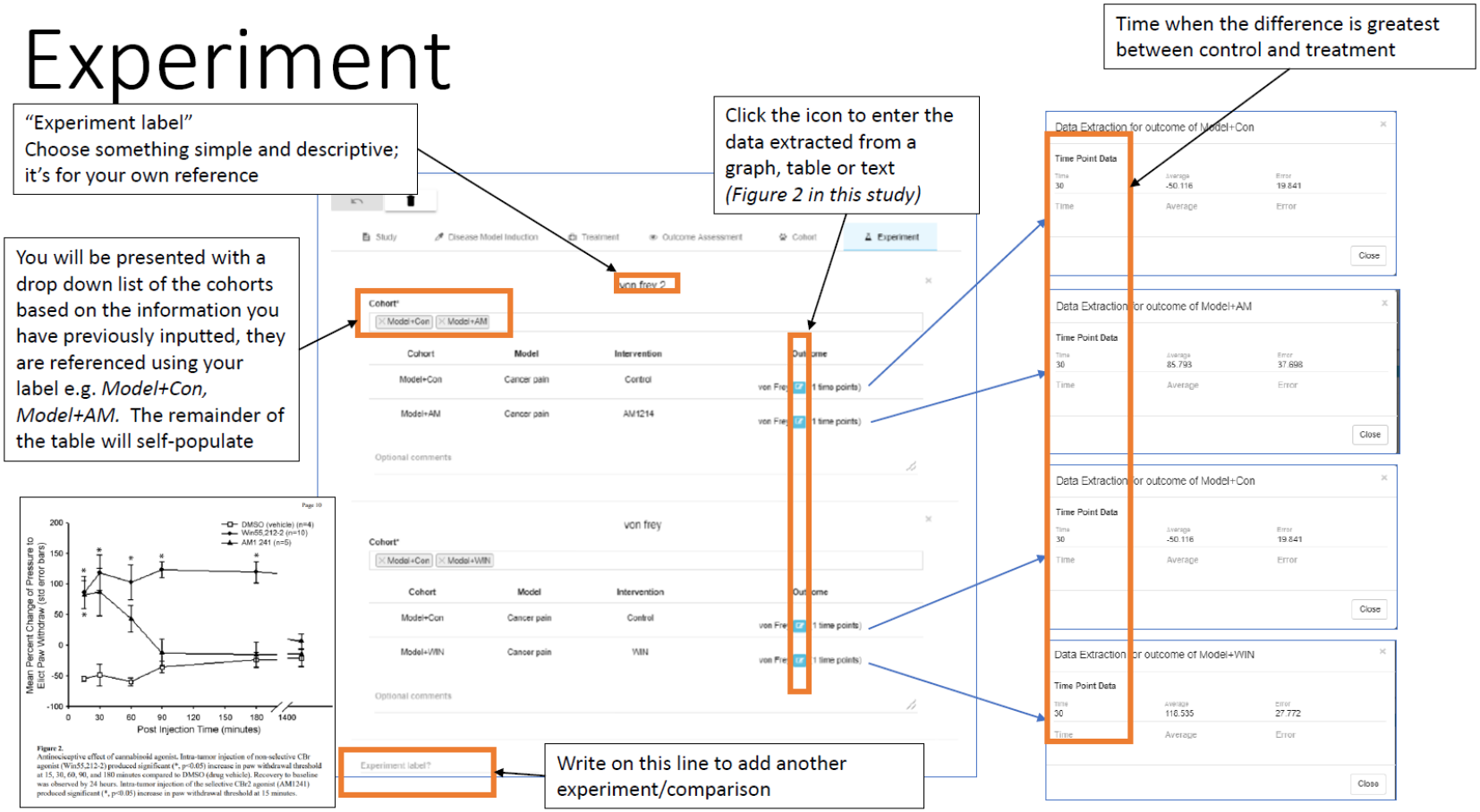


Figure 4.10. The Experiment tab allowed for the entry of outcome data extracted from a graph, table or text and included the timepoint, mean and variance.

4.2.4.4. Reconciliation

Data from each study were extracted by two independent reviewers therefore it was necessary to compare the data from each reviewer for agreement and resolve any disagreements. This process provides confidence that the information and data from the individual studies have been collated correctly. The way in which the SyRF data outputs were structured presented a complex challenge; it was a very large dataset comprised of hierarchical and nested data that required contextualised understanding to compare two reviewers. There were thousands of individual models, interventions, outcomes, cohorts, and experiments. In that format it was near impossible to perform reconciliation manually. I therefore set out with the CAMARADES lead developer, Chris Sena, to build a feature into the SyRF platform that could semi-automate the reconciliation process. This feature aimed to identify agreements between questions and answers. For studies in which there were multiple models, interventions, outcomes, cohorts, and experiments reviewers could input these sets of answers in any order. Therefore, we aimed to use machine learning to compare the sets, match and rank the sets that are most similar between the two reviewers and present the matched sets to the reconciler for them to confirm the match. To build such a feature takes considerable time, time that the project (nor CS) did not have so I was assisted by the creation of the first stage of the reconciliation platform; presenting two reviewers completed forms side by side (Figure 4.11) so that I and a small specialist team could manually perform reconciliation. Disagreements required the reconciler to return to the original study and determine which reviewer was right. The agreed data and reconciled data were imputed into a 'clean' annotation and data extraction form on the main SyRF platform. While this carried the risk of transcription errors it was the only feasible solution due to the complexity of the data structure.

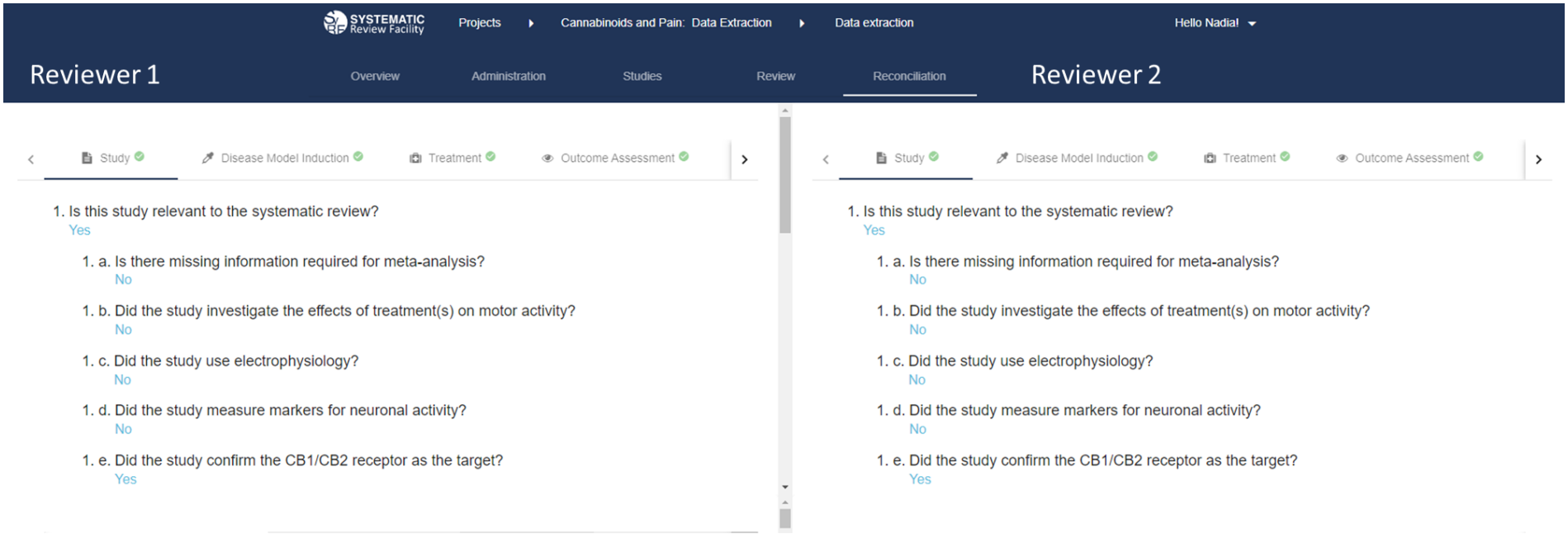


Figure 4.11. Screenshot of the reconciliation version of SyRF which allowed the presentation of two reviewers' completed annotation and data extraction forms side by side for the identification of agreements and disagreements.

4.2.5. Preparing Data for Synthesis

Once the data were reconciled and I had the data for each individual study, I needed to organise the data into a format that would allow for data cleaning and wrangling. The SyRF platform provided individual .csv file outputs for each of the tabs (study, model, intervention, outcome, cohort, and experiment). These needed to be joined by the appropriate matching identifier to create a complete dataset. Within each output the corresponding parent and children questions also had to be matched. I then transformed and mapped the collected data into the required format for the meta-analysis. This included categorising models by model type, behavioural assessments by assessment types, conversion of variance reported as standard error of the mean to standard deviation, and the calculation of the true number of control animals based upon the number of groups a control group served. Each row presented one outcome for a cohort of animals, so I next needed to identify the control groups for each of the treatment groups. Given that within an experiment there could be multiple treatment groups to a single control, or different control time point extracted per the different treatment groups in the experiment, it was not possible to automate this process. I manually matched over 2000 comparisons. In future projects, I will change the data structure and ensure that there is a treatment and corresponding comparator identifier to be able to automate this step. This will increase confidence in the accuracy and reproducibility of the process. Data cleaning typically involves removing inaccurate data from the dataset. I did not remove any data, but I checked the outliers were accurate and checked the magnitude and directions of the effects. It was in doing this that I identified a systematic error in the direction of effect assessed in the formalin test and the hypersensitivity to cold assessments. An error likely related to the non-compulsory nature of an annotation question that asks a reviewer/reconciler to confirm whether 'greater is worse' for the reported outcomes.

The time consuming, manual and error prone reconciliation and data preparation processes highlights the importance of considering data structure while building annotation and data extraction forms. It also highlights the current limitations of the SyRF platform in terms of how it provides the outputs. However, there are also opportunities for a platform like SyRF to be able to provide semi- or fully automated solutions to these steps which will ensure accuracy and complete, reproducible workflows that can be checked and built upon more readily in the future.

4.2.6. Meta-analysis

For the meta-analysis I followed the guidance as described by Vesterinen et al. (2014). In summary, for each comparison I calculated a Hedge's g standardised mean difference effect size. Effect sizes were weighted using the inverse variance method to reflect the contribution of each comparison to the total estimate of effect. Data reported from separate cohorts of animals were treated as independent comparisons and analysed separately within the meta-analysis. When more than one pain-associated behavioural outcome measure was reported for the same cohort of animals, the comparisons were combined to provide a single

nested comparison: a summary effect for that individual cohort. Cohort-level effect sizes were then pooled using the random-effects model (adjusted using the Hartung-Knapp-Sidik-Jonkman method). Last, heterogeneity was explored using stratified meta-analysis and the restricted maximum likelihood model (Soliman et al., 2021). Several approaches aim to give a more conservative or reliable estimate of effect: corrections for control group size, for small study bias (Hedge's g) and for the uncertainty of between-study heterogeneity (Knapp Hartung method). I also used the complete dataset; I did not remove outliers, nor did I set a threshold on the number of comparisons required for subgroup analyses, the rule of thumb that there should be at least 10 studies or comparisons (Borenstein et al., 2009). Therefore, the choice of meta-analytical methods and how the data is categorised will influence the overall results. The influence of these decisions will be explored in the next chapter (Cannabis & Pain Preclinical Systematic Review: Results and Further Analyses).

4.2.7. Discussion

Within this chapter I have described the complexity of the design and conduct of a preclinical systematic review. The delivery of a preclinical systematic review requires considerable project management. I have outlined the project management considerations, technological requirements and the unspecified tasks that facilitate the smooth and accurate design, conduct, analysis, and reporting of a systematic review that will ensure the aims of a review are met in a timely manner.

The success of a preclinical systematic review depends on the close collaboration and coordination of many researchers of differing expertise and priorities. It is necessary to engage both primary researchers and systematic review experts. Their engagement and contribution is facilitated by effective communication, willingness to listen, learn and compromise. The use of a platform like SyRF not only provides an opportunity to coordinate the contributions of many researchers but also the possibilities to use other technologies e.g., machine learning, to improve the feasibility of preclinical systematic reviews. The optimal use of a platform such as SyRF requires an understanding of the features, including how to build the project within the platform to fulfil the project requirements. It is also necessary to ensure that the limitations of the platform are understood to prevent impairing efficient contributions to the systematic review. The design for how the platform is used will impact the accurate and efficient annotation and data extraction and the data structure which will have downstream effects on the accuracy and efficiency of the subsequent reconciliation and data preparation and synthesis stages. In depth planning and iterative development and testing will improve the usability of the form and subsequent data structures thereby creating opportunities for the automation of the reconciliation and data preparation and synthesis stages. Together these steps will improve the accuracy and reproducibility of the data synthesis steps and it will provide efficiency savings throughout the systematic review process.

Despite the development of a protocol *a priori* which specifies the main tasks within the review there is still much freedom afforded to reviewers and at each stage of the review process and the decisions made are likely to influence the accuracy and reliability of the results. It is therefore imperative that reviewers understand the intricacies and nuances of the tasks implied tasks as summarised in Table 4.6.

The Cochrane Handbook comprised of 4 parts, 26 chapters and provides very comprehensive guidance (Higgins et al., Cochrane, 2021) for clinical systematic reviews and much can be followed or adapted for preclinical systematic reviews. In the absence of a fully comprehensive guide for preclinical systematic reviews particular attention should be paid to the principles and themes discussed in Part 2: Core Methods of the Cochrane Handbook. There are characteristics of preclinical systematic reviews e.g., the volume and broad variation in how studies are reported making identification, annotation, data extraction and analysis, in some instances, more complex than described in clinical systematic review guidance. Therefore, care and attention must be paid to the suitability of using methods created for clinical evidence synthesis, this is particularly pertinent to meta-analytical techniques given the underlying assumptions and the differences in data types. Guidance based upon adaptation of clinical systematic review methods for preclinical contexts are increasing (Table 4.6). At the start of restrictions relating to the COVID-19 pandemic, as wet labs were closing, I was approached by several groups for guidance on how to conduct preclinical systematic reviews. This coupled with my experiences of leading a preclinical systematic review gave impetus to write an up-to-date step-by-step guide. I published a topical review in the journal *PAIN* which articulates the potential and importance of preclinical systematic reviews as well as providing guidance and signposting to the resources for the conduct of a preclinical systematic review (Soliman et al., 2020). The topical review also specifically addresses the different method options available to researchers to improve feasibility of this important research tool (Soliman et al., 2020).

Table 4.6. The implied tasks for each stage of the systematic review that are not described in the protocol that need to be conducted to perform a preclinical systematic review and meta-analysis.

Stage of Systematic Review	Implied Tasks	Preclinical Specific Guidance
Search	• Identify appropriate electronic databases • Split research question into critical search components • Identify standardised and free-text terms • Test for specificity and sensitivity and adjust • Export and de-duplicate search results • Upload search results to review platform	Hooijmans et al. (2010) Leenaars et al. (2012) de Vries et al. (2014)
Study Selection	• Specify inclusion and exclusion criteria • Provide examples for each of the PICO elements • Provide guidance on the platform and with additional materials to aid reviewers • Communicate how to apply the criteria and how to navigate the platform	de Vries et al. (2015a)
Annotation and Data Extraction	• Build an annotation and data extraction form • Design focused, unambiguous questions to extract the data required to answer the research question(s) and so that data are categorised to allow for comparison • Ensure the questions and form design allow for the extraction of data as presented within studies e.g., numerical data from graphs, denote units e.g., for time and dose, ranges • Test • Be cognisant of final data structure; try to avoid hierarchical and nested data structures • Communicate rules for data extraction and how to use the data extraction form	syrf.org.uk
Reconciliation	• Identify agreement and disagreement for each annotation questioned answer and data extracted • Reconcile disagreements between reviewers by checking against the original study	syrf.org.uk
Analysis	• Prepare the data for analysis; combine the outputs, identify the comparisons, calculate the true number of control animals, categorise accordingly e.g., model type and outcome assessment type, for sub-group analyses, check data for accuracy (magnitude and direction of effect) • Summarise study characteristics • Consider the biological context and best grouping/sub-grouping • Consider whether the data (e.g., size of dataset and type of data) meet the assumptions of the proposed meta-analytical approaches	Vesterinen et al. (2014) Wang et al. (2018) Macleod et al. (2020) Tanriver-Ayder et al. (2021)

5. Cannabis & Pain Preclinical Systematic Review – Part 2: The Results and Further Analyses

5.1. Introduction

Cannabis is being legalised in countries around the world for both recreational and medicinal uses. There is a lot of conflicting evidence for the use of cannabis-based medicines for pain management and as presented in the previous chapter, the International Association for the Study of Pain (IASP) established a Presidential Task Force (TF) on Cannabis and Cannabinoid Analgesia which aimed to assess the comprehensively appraise the evidence. I led the preclinical systematic review which afforded the opportunity to assess crowd science and machine learning as methods for improving the feasibility, efficiency, and accuracy of preclinical systematic reviews.

Pain is an unpleasant sensory and emotional experience associated with, or resembling that associated with, actual or potential damage (Raja et al., 2020). Pain can serve an adaptive role, however for individuals living with chronic pain it can have a limiting and detrimental effect on function, social and psychological wellbeing. It is a leading cause of disease burden worldwide and diminishes quality of life (Rice et al., 2016). Even acute (e.g. inflammatory or postoperative) pain can be detrimental by causing impaired recovery and leading to chronic pain (Pogatzki-Zahn et al., 2017). Current therapies are of modest efficacy and have adverse effects (e.g. Finnerup et al. (2015) and Pogatzki-Zahn et al. (2017)) hence, there remains a considerable unmet medical need.

Cannabinoids, cannabis-based medicines, and endocannabinoid system modulators as potential therapeutics for pain management are of increasing research interest. The endocannabinoid system, comprised of the cannabinoid type 1 (CB_1) and type 2 (CB_2) receptors, their endogenous ligands, and the enzymes that metabolise the endogenous ligands, is implicated in pain modulation *in vivo*. Hence, cannabinoids, cannabis-based medicines and endocannabinoid system modulators as possible therapeutics for pain management have been studied extensively in animal models (reviewed most recently in our companion narrative review (Finn et al., 2021).

There are, however, several unanswered research questions remaining. For example, there is uncertainty regarding the clinical evidence for the analgesic efficacy of cannabinoids, cannabis-based medicines, and endocannabinoid system modulators, and it is not clear whether the current clinical evidence, based upon efficacy and safety considerations, justifies their use for pain management (Fisher et al., 2020, Sideli et al., 2020). These residual questions strengthen the rationale for assessing the full evidence base. Improving our understanding of the preclinical literature will better inform future clinical research. Animal models of injury-related and pathological persistent pain are used to investigate the underlying pathophysiology as well as to

assess the efficacy and adverse effect profile of potential analgesics. Such studies provide justification and indications for clinical trials. The failure to translate findings from preclinical research to clinical treatment has raised questions about the predictive validity and utility of animal models in drug development (Ioannidis, 2005). Limitations in experimental design (Andrews et al., 2016a, Mogil, 2009), conduct (Currie et al., 2019, Mogil, 2017), analysis and reporting (Clark, 2016, Rice et al., 2008) may be compounding the challenges of translational pain research and hindering the development of effective therapies.

This preclinical systematic review employed crowd science and machine learning to meet the aims of the review. Each method was also assessed on ability to improve the feasibility of systematic review processes (discussed in Chapters 6 & 7). The aims of the review were to estimate anti-nociceptive efficacy and assess the impact of methodological quality and study design characteristics on the overall estimate of effect. This chapter reports the synthesis of the qualitative and quantitative data. Using pain-related behavioural outcome data an estimate of the antinociceptive efficacy of cannabinoids, cannabis-based medicines, and endocannabinoid system modulators in animal models of pathological or injury-related persistent pain was calculated. The impact of methodological quality, study design characteristics and potential publication bias was explored (Soliman et al., 2021). This chapter also expands on the analyses that were conducted and explores how different decisions regarding meta-analytical approaches can impact the results. To reflect the intrinsic collaborative effort of the empirical work presented in this chapter, some of the work will be described in the first-person plural voice. All analyses were conducted independently.

5.2. Methods

The systematic review methods have been described in detail in the previous chapter (Chapter 4), the focus here is the qualitative and quantitative analysis. The methods were prespecified in the study protocol and are available on PROSPERO (CRD42019124804) and published in PAIN Reports (Soliman et al., 2019a). Briefly, in April 2019, I conducted the search of 3 online electronic databases and used crowd science and a machine learning algorithm to identify studies that met the inclusion criteria. Data pertaining to methodological quality criteria and information regarding the model type and method of induction, intervention, dosing regime, and pain-associated behavioural assessment and outcome data were collected. Each study was reviewed by two independent reviewers and reconciled by a third independent reviewer.

5.2.1. Risk of Bias

In accordance with the CAMARADES checklist (Macleod et al., 2004) and adaptation of the SYRCLE Risk of Bias tool (Hooijmans et al., 2014) the risk of bias of the included studies was assessed by recording the reporting of five methodological quality criteria: sample size calculation, randomisation, allocation concealment, blinded assessment of outcome, and reporting of animal exclusions. Criteria were extracted by two independent reviewers and discrepancies reconciled by a third. Risk was scored high, low, or unclear based upon the reporting of the method.

5.2.2. Meta-Analysis

A meta-analysis was conducted in accordance with the guidelines described by Vesterinen et al. (2014).

5.2.2.1. Definition of a Comparison

A comparison was defined as a measure of pain-associated behavioural outcome in a group of animals that had been administered a cannabinoid, cannabis-based medicine, or endocannabinoid system modulator, compared to a group of animals that had been administered a vehicle control. Often the same group of animals undergo multiple pain-associated behavioural outcome measurements. In those instances, the comparisons were combined to give one outcome statistic (a nested comparison) that represents the global measure of the outcome in that comparison. Where a control group served multiple treatment groups in an experiment, the number of control animals was corrected by dividing the number of animals reported in the control group by the number of treatment groups it served.

5.2.2.2. Calculating an Effect Size for each Comparison

The data extracted from the included studies were continuous data and for each group of animals the mean outcome and the variance was extracted. With this type of data there are three options for calculating an effect size; absolute mean difference, normalised mean difference or standardised mean difference.

An **absolute mean difference** calculates the effect size by calculating the difference between the reported means of the two groups and their corresponding variance. This method requires the data to have come from the same outcome measure and for all included studies to report this outcome measure on the same scale. Considering that there were many types of pain-associated behavioural outcome measures and reporting outcomes on different measurement scales across the different studies, this would not give an effect size that would allow reliable comparison between studies or overall summary of effect.

Normalised mean difference is appropriate for when one of the groups is comprised of healthy, naïve, or wild-type animals and can be used to calculate an effect size which is an absolute difference in means as a proportion of the mean in the control group. However, on the basis that a minority of studies would assess and report outcomes of the treatment in naïve animals it was decided not to collect naïve animal data from the studies.

Therefore, **standardised mean difference** effect sizes were calculated. This method calculates the difference between the means of each group divided by the pooled variance. This was the most appropriate method because naïve animal data was not available and because it accounts for the different pain-associated behavioural outcomes allowing for all outcome measures to be presented on a standardised scale. A correction for small sample size bias was also applied, Hedge's G. Small samples are considered groups of less than 10 animals (Rooke et al., 2011).

5.2.2.3. Weighting Effect Sizes

The included studies within the review have varied precision and therefore a weighted mean was calculated for each comparison. This ensured that studies or comparisons that were more precise carried more weight in their contribution to the overall summary effect, with those that are less precise thus carrying less weight. Weights were calculated using the inverse variance method; individual effect sizes were multiplied by the inverse of their squared standard error.

5.2.2.4. The Statistical Model of Analysis

The next step was to pool the effect sizes to give the distribution of effect sizes of the individual comparisons a weighted mean (the summary estimate), a weighted sum of the square of the deviations from that mean (the heterogeneity) and an estimate of the variance of the effect sizes beyond that expected by chance (tau-squared, τ^2). Using the standard error, the 95 % confidence intervals were then calculated. There are two methods for pooling effect sizes: the fixed effects model and the random effects model.

The fixed effects model assumes that there is one true effect size which underlies all the studies included in the analysis and that any differences in observed effects are due to sampling error. In contrast, the random effect model assumes that the true effect can vary from study to study, or that there is an array of true effects due to the varied study design characteristics; in this review the studies varied in the type of model, animal characteristics, intervention, and pain-associated behavioural outcome measures etc. Therefore, I used the random effects model of analysis because there was a considerable amount of diversity in the study design characteristics and thus the effect sizes are assumed to represent a random sample of these effect sizes. The overall summary effect represents the mean of the population of true effects and takes into consideration the within study variance (sampling error) and between study variance (differences in true effect sizes between studies). Therefore, studies are weighted using the inverse sum of the study variance. τ^2, a measure of between study variation, was estimated using the Restricted Maximum Likelihood (REML). I made this decision based upon simulations using preclinical data conducted and reported by Tanriver-Ayder et al. (2021). It was deemed to be the most appropriate estimator because the effect sizes are based on continuous outcome data; this is supported by clinical data and simulations (Veroniki et al., 2016, Viechtbauer, 2005, Tanriver-Ayder et al., 2021). In addition, I used Knapp-Hartung adjustments to calculate the confidence interval around the pooled effect, thereby reducing the chance of false positives (Knapp and Hartung, 2003, Langan et al., 2019, IntHout et al., 2014).

5.2.2.5. Assessing Heterogeneity

It was necessary to try to understand what factors may be influencing the between study differences of true effect sizes. However, the array of true effects is comprised of both true variance and sampling error. It is a requirement therefore to separate the true variance to identify real differences among studies. This portion of variance was then used to create several measures of heterogeneity:

The first step to partition variance was to calculate Cochran's Q; an estimate of between study heterogeneity. It is calculated as the weighted sum of differences between individual effect sizes and the pooled effect across studies under the fixed effects model. The values of Q follow a chi-square distribution, on the assumption that all studies share a common effect size and therefore when Q equals the degrees of freedom all variation is due to sampling error within studies. The difference between Q and degrees of freedom reflects the excess variation, the variation attributed to between-study differences, the differences in the true effects. However, Q is sensitive to the number of comparisons. It has low power when the number of comparisons is small and thus is susceptible to identifying heterogeneity that is not genuinely true.

The significance of the heterogeneity was also calculated. The null hypothesis being that all studies share a common effect size under which Q follows a central chi-squared distribution with degrees of freedom equal to K-1 therefore a p-value less than alpha (0.05) can lead us to reject the null hypothesis and conclude that the studies do not share a common effect size. However, a non-significant p-value should be taken as evidence that the effect sizes are consistent, it too is susceptible to low power. With a small number of studies and/or large within-study variation, even substantial between-studies dispersion might yield a non-significant p-value. Similarly, a significant p-value may reflect a large dispersion with imprecise studies or small dispersion with precise studies.

To remove the dependence on the number of comparisons I^2 was also calculated. I^2 is a descriptive statistic that addresses what proportion of the observed variance that reflects real differences in effect size. The scale of I^2 has a range of 0-100 % and allows us to discuss the amount of variance on a relative scale. It is calculated from Cochran's Q (I^2 = 100 % x (Q – *df*)/Q). A value of 0 % indicates no observed between-study heterogeneity and 0-25 % very low, 25-50 % low, 50-75 % moderate, 75-100 % (Higgins et al., 2003).

5.2.2.6. Exploring Sources of Heterogeneity

A stratified analysis builds upon the assumption that studies that share certain characteristics will be similar compared to those that do not share characteristics. A stratified meta-analysis was performed, partitioning heterogeneity according to study quality criteria, type of outcome measure, model type, intervention class and animal characteristics e.g., strain and sex. For each subgroup, I used the random effects model and calculated Cochran's Q (Cochran, 1954). These heterogeneity statistics were then summed and deducted from the total heterogeneity to give an estimate of the remining heterogeneity which accounts for the between study differences. The significance of the differences was determined using the chi squared distribution.

I also used stratified meta-analysis to assess the anti-nociceptive efficacy of the interventions in the different model types on the theoretical basis that different aetiology may lead to different sensory profiles and therefore drugs with differing mechanisms of action will differ in their efficacy. However, this did not consider the outcome measure used. A criticism of preclinical pain research is the overreliance on stimulus

evoked behavioural outcome measures and their face validity (Mogil, 2009, Rice et al., 2018). The measure of hypersensitivity while appropriate for some indications e.g., post-operative and musculoskeletal (Pogatzki-Zahn et al., 2018, Pogatzki-Zahn et al., 2017), it is not for others e.g., neuropathic pain which is often characterised by spontaneous pain and pain in the presence of sensory loss (Baron et al., 2017). Therefore, using stratified meta-analysis I have assessed the hypothesis that animal models representing different pain pathophysiological mechanisms will differ in their sensitivity i.e., neuropathic pain associated models will not demonstrate the ability to predict analgesic efficacy whereas inflammation, cancer, and post-operative models will be sensitive to mechanical limb withdrawal as a pain-associated behavioural outcome measure.

Recent studies have suggested that a meta-regression is more effective than stratified meta-analysis for exploring sources of heterogeneity and reporting differences between sub-groups (Wang et al., 2018, Tanriver-Ayder et al., 2021). This is because partitioning is less powerful since stratification of preclinical data will lead to smaller subset of unpowered studies with poorer estimates of amounts of heterogeneity (Tanriver-Ayder et al., 2021). Although the Cannabis & Pain dataset is large and, in many instances, does not follow the small dataset challenges as described in these studies, I have built upon the analysis presented in the review and performed a forced entry multiple meta-regression. This is where all predictors are forced into the regression model at the same time. I also performed a multi-model inference meta-regression. This method does not try to build a 'best' model to explain most the variance, but all possible predictor combinations are modelled. This results in the creation of several meta-regressions thereby all possible predictor combinations and how they perform are evaluated. For all meta-regressions I used the same predefined theoretical rationale as used for the stratified meta-analysis; rats and mice were analysed separately, and the predictors were drug class, model type, outcome assessment type, sex, and strain.

5.2.3. Publication Bias

Three different analyses were performed to estimate if the unavailability of research data could have skewed the overall estimate of effect, of particular concern is an overestimate of effect due to the absence of null or negative data (Sena et al., 2010). The risk of publication bias was assessed using funnel plot assessment for asymmetry (Zwetsloot et al., 2017) and Egger's regression (Egger et al., 1997). Trim and fill were then performed to impute theoretically missing studies (Duval and Tweedie, 2000). This evaluation was conducted at the study level (all comparisons within a study were nested) because a range of positive, null, and negative results can be presented within a single study. However, an alternative approach presented here is to perform publication bias assessment based upon outcomes thereby including all possible comparisons that contributed to the analysis of the effect.

5.2.3.1. Funnel Plots

Plotting effect sizes on the x axis against a study's precision on the y axis creates a scatter plot that allows for visual inspection for the potential presence of asymmetry that may be attributed to publication bias. I initially

plotted precision calculated as the inverse of the variance of a study (Vesterinen et al., 2014) however it was brought to my attention that this can lead to distortion and an overestimation of potential publication bias (Zwetsloot et al., 2017). I therefore updated the analysis in accordance with Zwetsloot et al. (2017) who suggest that for studies with small sample sizes, a sample size based precision estimate is more appropriate, calculating precision as 1/√n.

The underlying assumption is that the precision of a study increases with increased sample size, therefore smaller, less precise studies are scattered widely at the bottom of the graph and larger, more precise studies are less widely scattered at the top of the graph. When the data points are symmetrically, inverted funnel shape and centred on the overall summary of effect, then there is no suggestion of publication bias. Asymmetry suggests possible publication bias (Higgins et al., Cochrane, 2021).

5.2.3.2. Egger's Regression

Egger's regression allows for statistical assessment of the potential presence of publication bias from the funnel plot (Egger et al., 1997). I plotted a linear regression of standardised effect sizes (effect sizes divided by their standard error) on y axis against precision (1/√n) on the x axis. It determines whether the regression line and its 95% confidence interval, for precision vs standardised effect size intersect at the origin of the graph. If the line does not, then the presence of publication bias is concluded.

5.2.3.3. Trim and Fill

Trim and fill analysis nonparametrically attempts to correct for funnel plot asymmetry by identifying and imputing theoretically missing studies (Duval and Tweedie, 2000). This enables a quantification of missing studies and recalculation of the overall effect size in the absence of publication bias. Firstly, small studies responsible for the asymmetry are 'trimmed' from the funnel plot. Secondly, using the remaining studies within the funnel plot and estimate of the true centre of the funnel is determined. Lastly, the 'trimmed' studies are replaced in areas around the centre where potential studies may be missing.

5.3. Results

There were 473 studies that met the inclusion criteria; however, 99 studies were missing key information so could not be included in the meta-analysis (Figure 5.1).

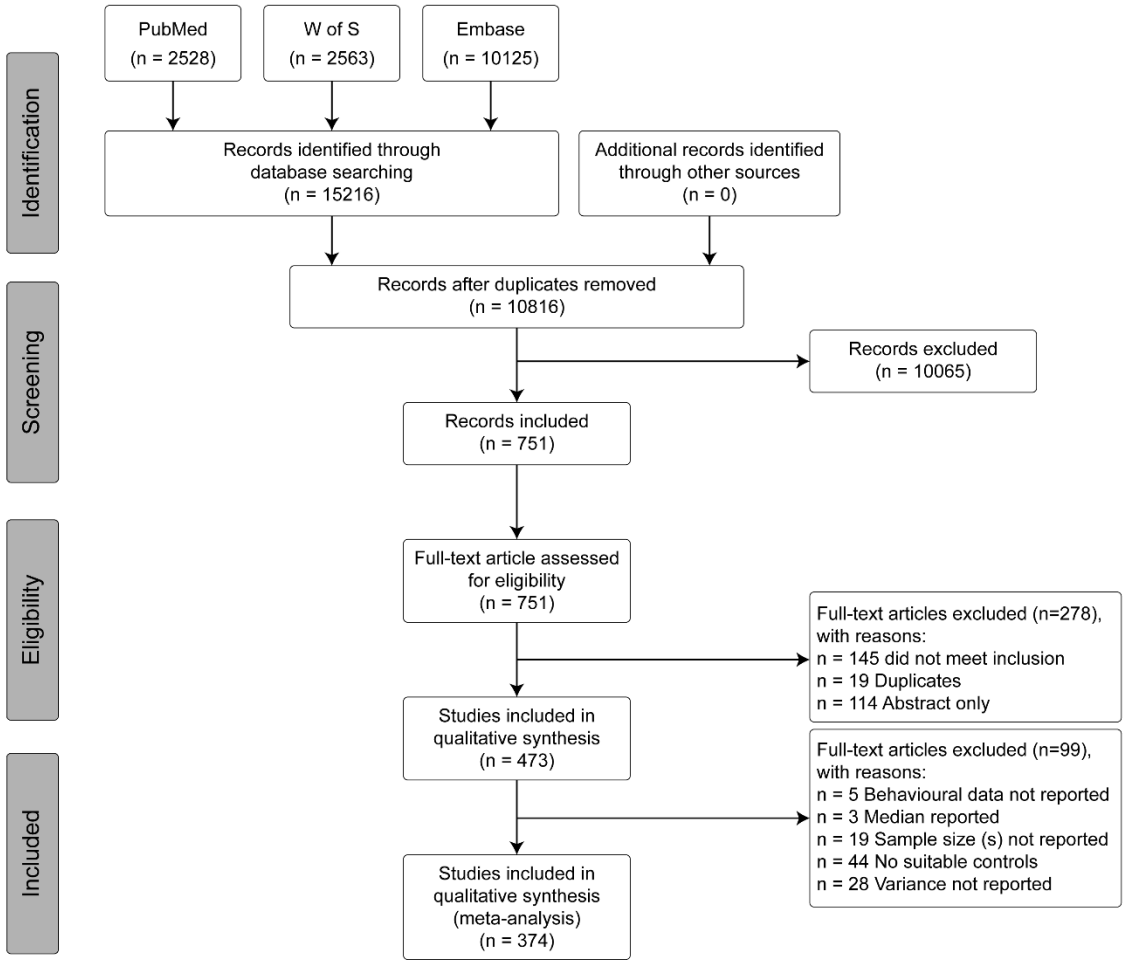


Figure 5.1. A flow diagram of articles identified from the bibliographic search of 3 electronic databases, PubMed, Web of Science and Ovid Embase, conducted on 9 April 2019. The diagram provides the breakdown of records through de-duplication, screening, eligibility until final inclusion in both qualitative and quantitative analysis. Reported in accordance with the PRISMA guidelines. W of S, Web of Science.

5.3.1. Study Characteristics

In the 374 studies included in the meta-analysis, 171 interventions were assessed for antinociceptive effect in models of pathological or injury related persistent pain (see Supplemental Digital Content 1, for included study list and study characteristics). The drugs are listed by mechanism of action (23 drug classes) in Table 5.1 (see Supplemental Digital Content 2, for full list). There were a total of 1544 individual treatment versus control comparisons. CB_1 and CB_2 receptor agonists were tested most frequently, 281 (19 %) and 299 (20 %) comparisons, respectively. The most frequently tested drug was the CB_1 receptor agonist, WIN55,212-2 (n=194 comparisons, 13 %). 20 model types were reported (Table 5.2); inflammation, nerve injury, and formalin were used most frequently: 30 % (n=467 comparisons), 27 % (n=413 comparisons), 15 % (n=235 comparisons) respectively. Over 16,000 animals were included in the analysis with a median of 24 animal's pain-associated behavioural outcome data extracted from each included study (a range of 4 to 162).

Table 5.1. Summary of the drug classes assessed for anti-nociceptive effect in animal models of injury-related or pathological persistent pain. FAAH, fatty acid amide hydrolase, PPAR, peroxisome proliferator-activated receptor, THC, delta-9 tetrahydrocannabinol, CBD, cannabidiol, FABP, fatty acid-binding protein, NAAA, N-acylethanolamine-hydrolysing acid amidase, MAGL, monoacylglycerol lipase, PAM, positive allosteric modulator, TRPV1, transient receptor potential vanilloid receptor 1, TRPA1, transient receptor potential ankyrin 1, ABHD6, abhydrolase domain containing 6.

Drug Class	**Number of Studies**	**Number of Nested Comparisons**
CB_2 receptor agonist	75	299
CB_1 receptor agonist	88	281
Non-selective cannabinoid receptor agonist	71	230
FAAH inhibitor	57	217
PPAR-alpha agonist	40	121
THC	16	69
Anandamide transport inhibitor	18	64
CBD	17	63
Monoacylglycerol lipase inhibitor	23	58
FABP inhibitor	3	31
Unknown mechanism of action	6	25
NAAA inhibitor	4	20
CB_1 receptor inverse agonist	7	19
Diacylglycerol lipase inhibitor	3	14
Dual FAAH/MAGL inhibitor	4	10
CB_1 receptor PAM	1	5
FAAH inhibitor/TRPV1 agonist	1	5
CB_2 receptor inverse agonist	2	4
ABHD6 inhibitor	1	3
FAAH inhibitor/TRPA1 agonist	1	2
PPAR-gamma antagonist	1	2
GPR55 agonist	1	1
Hemp oil	1	1

Table 5.2. Summary of the model types used to assess the anti-nociceptive effect of cannabinoids, cannabis-based medicines, and endocannabinoid system modulators.

Model Type	Number of Studies	Number of Nested Comparisons
Inflammation	434	467
Nerve injury	348	413
Formalin	223	235
Chemotherapy	112	128
Diabetes	63	74
Cancer	57	65
Post-operative	27	52
Visceral inflammation	20	31
Chemical cauterization	1	16
Migraine	9	13
HIV	4	11
Capsaicin	5	9
Heat injury	2	7
Multiple sclerosis	6	7
Musculoskeletal	2	4
Antiretroviral	1	3
Burn injury	1	3
Mustard-oil	3	3
Sickle cell disease	2	2
Mild traumatic brain injury	1	1

All experiments were conducted in rodents. Fifty-six % (n=865 comparisons) were conducted in rats and 44 % (n=678 comparisons) were conducted in mice. Male animals were used in 86 % (n=1334 comparisons) and female animals were used in 7 % (n=110 comparisons). Two % (n=28) used mixed sex groups and 5 % (n=74 comparisons) did not report the sex of the animals used. Evoked limb withdrawal to mechanical and thermal stimuli were the most frequently used pain-associated behavioural outcome measures; 51 % (n=791 comparisons) and 22 % (n=343 comparisons) respectively.

To inform a narrative review (Finn et al., 2021) and future research, the following was annotated for each study and is summarised in Table 5.3: Whether the study investigated the effects of drug on non-pain related motor activity, the pharmacokinetic properties, tissue concentrations, and where applicable confirmed the cannabinoid receptors as the target. Additionally, we assessed whether the study investigated potential toxic effects, effects on dependency and on aspects of animal behaviour potentially reflecting anxiety- and depression-related behaviour. We also assessed whether electrophysiology (e.g., wide dynamic range and nociceptive specific cells) or markers for neuronal activity (e.g., c-fos, Fos, ERK, p38 MAPK) were measured in studies of antinociceptive efficacy.

Table 5.3. The number of studies that conducted further experimentation to gain further understanding of the cannabinoids, cannabis-based medicines, and endocannabinoid system modulators in conjunction with anti-nociceptive effect.

	Number of studies	**%**
Confirmation (where applicable) of the CB_1/CB_2 target	207	69
Effects on motor activity	124	33
Investigate pharmacokinetics	26	7
Tissue concentrations	25	7
Electrophysiology	20	5
Potential toxic effects	19	5
Measure markers of neuronal activity	19	5
Effect on anxiety/depression	11	3
Effects on dependency	10	3

5.3.2. Risk of Bias

The overall risk of bias of the 374 included studies is unclear. The reporting of methodological quality criteria was low: 47% (175) reported blinded assessment of outcome, 32% (118) reported randomisation (to treatment or control group), 14% (54) reported animal exclusions, 13% (49) reported predetermined animal exclusion criteria, 4% (15) reported allocation concealment, and 3% (12) reported a sample size calculation (Figure 5.2A). This contrasts with the reporting of conflict of interest, 54% (203), and high reporting of compliance with animal welfare regulations, 94% (353). The highest number of criteria reported was 4 out of 6 in 10 studies. The methods for how bias was mitigated were rarely reported; therefore, the studies are at an unclear risk of bias (Figure 5.2B; see Table, supplemental digital content 6, a traffic light plot presenting the risk of bias judgement for each study).

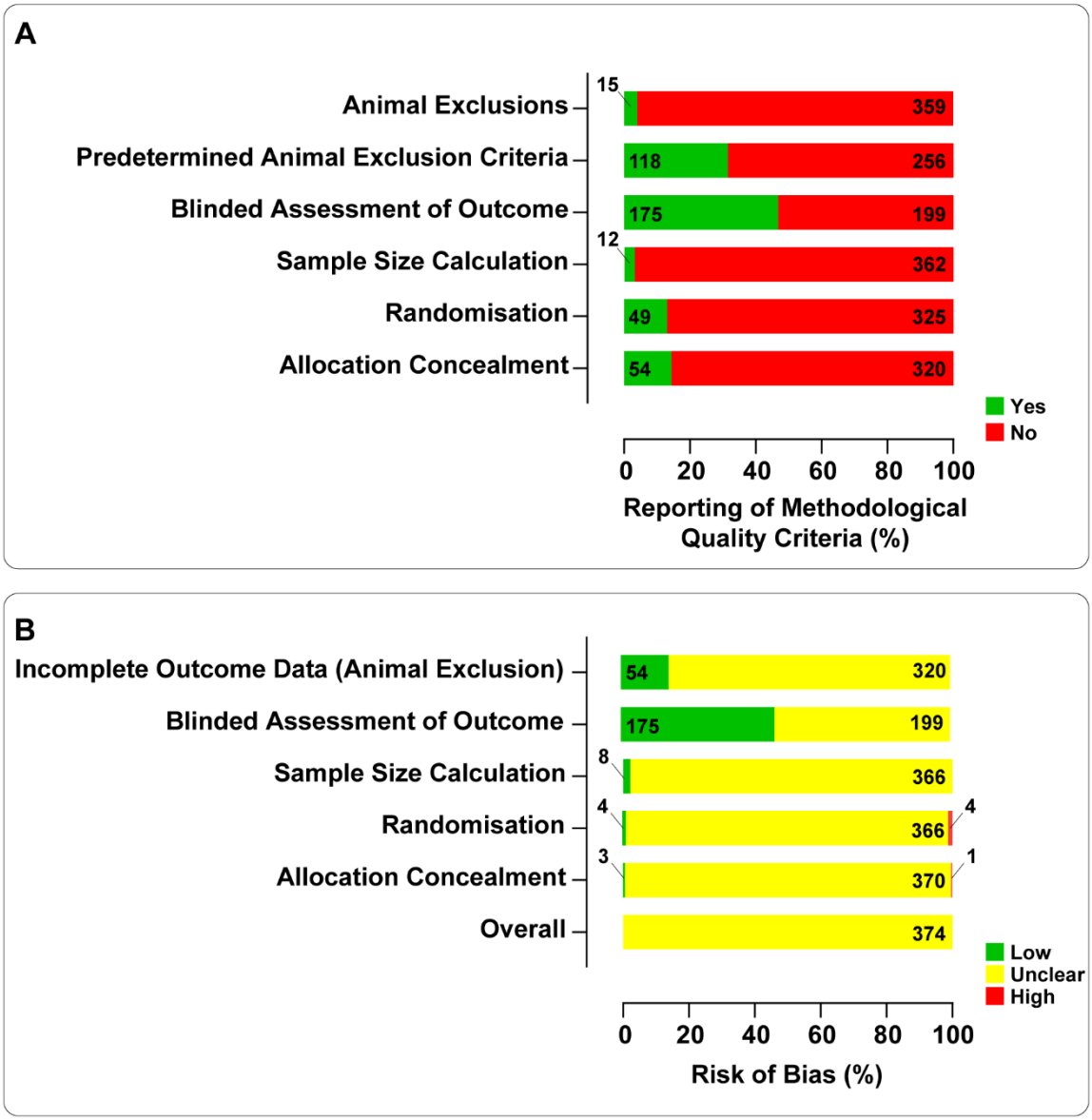


Figure 5.2. The reporting of methodological quality criteria (A) and a summary bar plot showing the proportion of studies with a given risk of bias for each methodological quality criteria (B) for the 374 included studies.

5.3.3. Meta-analysis of the antinociceptive efficacy of treatment with a cannabinoid, cannabis-based medicine, or endocannabinoid system modulator

A total of 374 studies, comprising 1544 comparisons, investigated the effects of cannabinoids, cannabis-based medicines, and endocannabinoid system modulators in models of pathological or injury-related persistent pain. Prophylactic and/or therapeutic administration of the drugs led to a significant attenuation of pain-associated behaviour compared to control (SMD=1.321 [95 % CI 1.232 – 1.411]). Heterogeneity was moderate (I^2 = 61.58 %) (Figure 5.3).

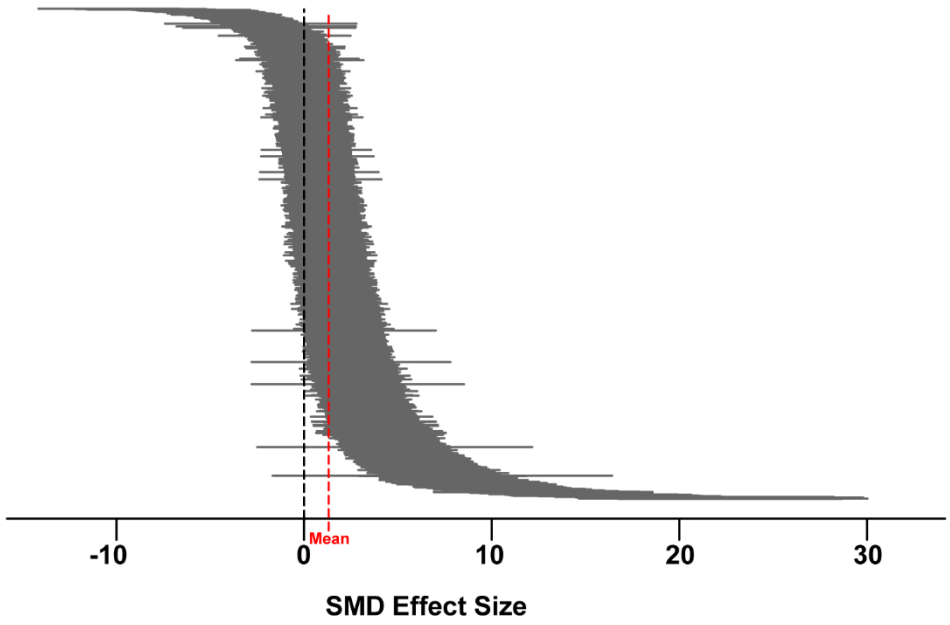


Figure 5.3. A caterpillar plot of the 1,544 nested comparisons extracted from the 374 studies included in the meta-analysis. Hedge's g standardised mean differences (SMD) were calculated for each comparison. Effect sizes were pooled using the random effects model and heterogeneity estimated with the restricted maximum likelihood model. Overall effect size=1.321. Q=4101.26, d.f. 1543, p <0.0001. I^2 = 61.58 %.

5.3.4. Effects of Study Characteristics on Antinociceptive Efficacy

Subgroup analyses demonstrated that species accounted for a significant proportion of heterogeneity (Q=17, d.f. 2, p<0.005). Therefore, rats and mice have been analysed separately. Further subgroup analyses were conducted to determine if the antinociceptive effect varies due to study design characteristics.

5.3.4.1. Rats

Rats were used in 276 studies (n = 6479, 864 comparisons) to assess the potential antinociceptive effect of the treatments. 95 interventions were tested in 11 model types. The treatments led to a significant attenuation of pain-associated behaviour compared to control (SMD=1.306 [95 % CI 1.199 – 1.412]). Heterogeneity was moderate (I^2 = 57.8 %, Q=2044.69, d.f. 863, p<0.0001).

The drug and drug class accounted for a significant proportion of heterogeneity (Q=1338.17, d.f. 94, p<0.0001 and Q=36.37 d.f. 20 p<0.05). Novel compounds not included in the classification were classified based upon the mechanism of action reported by the study authors. CB_1 receptor agonists and CB_2 receptor agonists were most frequently assessed (194 and 159 comparisons respectively). Most drug classes resulted in a significant antinociceptive effect; NAAA inhibitors produced the largest significant attenuation of pain-associated behaviour compared to control (SMD=1.59 [95% CI 1.17 – 2.01]), whereas PPAR-gamma antagonist, GPR55 agonist, hemp oil, FABP inhibitors and CB_2 receptor inverse agonist did not have significant effect, however these were of single studies or single comparisons. Although their effect is smaller, the CB_1 and CB_2 receptor agonists are comprised from data from a high number of animals and comparisons

therefore can have greater confidence in the stability of the effect. The smallest significant effect was elicited by CBD (SMD=1.12 [95% CI 0.84 – 1.40]). Most drugs were assessed post model induction (702 comparisons). Whether the drug was administered prophylactically, or therapeutically, did not account for a significant amount of heterogeneity (Q=0.30, d.f. 2, p=0.9) (Figure 5.4A).

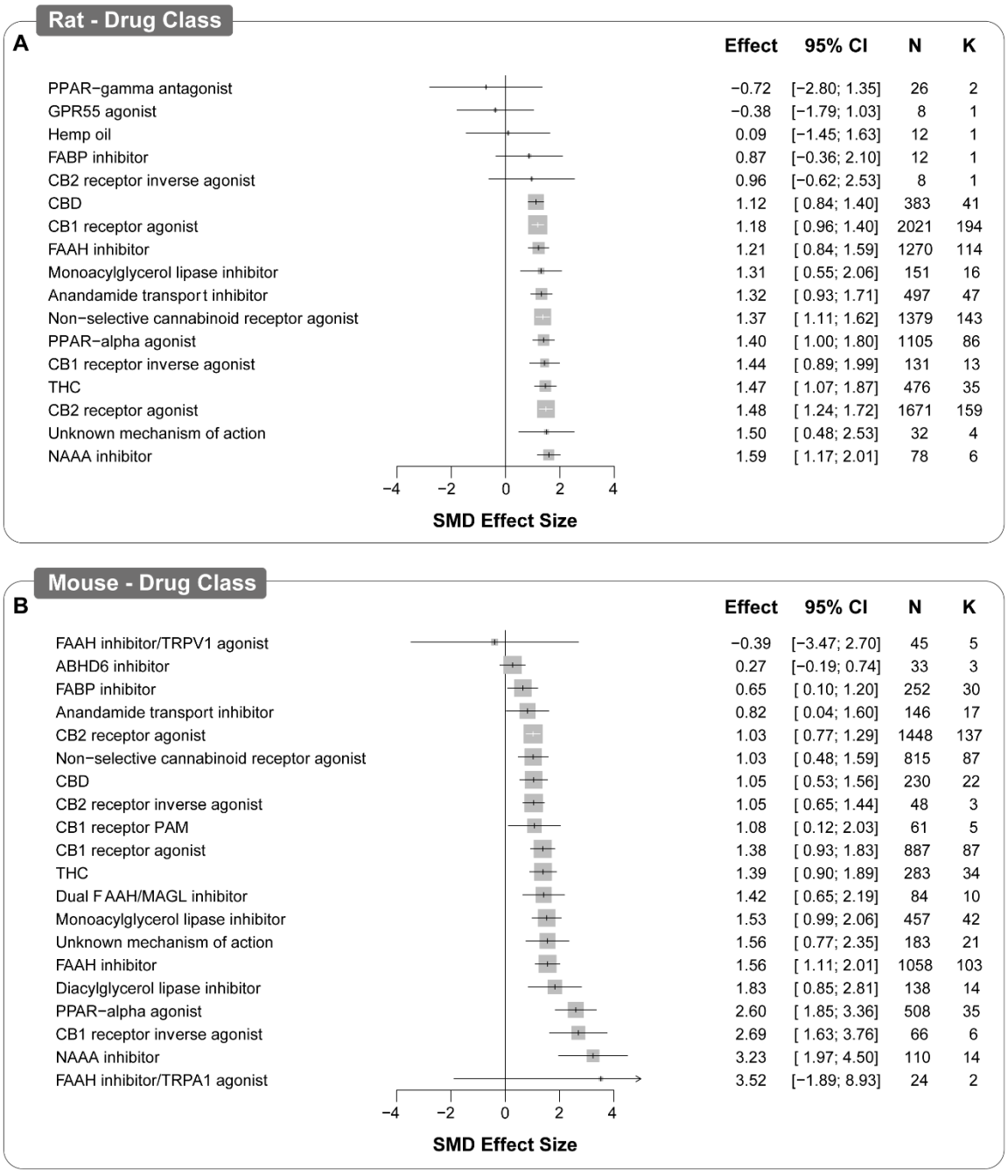


Figure 5.4. Forest plot of drug classes assessed in rat (A) and mouse (B) models of injury related or pathological persistent pain. The size of the square represents the weight (%). The weight is the influence that individual subgroup has on the pooled result. N denotes the number of animals that contribute to that subgroup. K denotes the number of comparisons that comprise each subgroup.

5.3.4.2. *Mice*

Mice were used in 153 studies (n = 6876, 677 comparisons) to assess the potential antinociceptive effect of the treatments. 110 drugs were tested in 15 model types. Overall, the treatments led to a statistically significant attenuation of pain-associated behaviour compared to control (SMD = 1.353[95 % CI 1.199 – 1.506], p<0.0001). Heterogeneity was moderate (I^2=66.7 %, Q=2027.72, d.f. 676, p<0.0001).

The drug and drug class accounted for a significant proportion of heterogeneity (Q=5332.76, d.f. 109, p<0.0001 and Q=82.26 d.f. 23 p<0.0001). CB_2 receptor agonists and FAAH inhibitors were assessed most (137 and 103 comparisons respectively). As in rats, most drug classes produced a significant antinociceptive effect. The largest significant attenuation of pain-associated behaviour compared to control was reported for NAAA inhibitors (SMD = 3.23 [95% CI 1.97-4.50]) although, we can have more confidence in the smaller effect sizes of the CB_2 receptor agonists and FAAH inhibitors. Additionally, a FAAH inhibitor/TRPV1 agonist and ABHD6 inhibitors did not significantly attenuate pain-associated behaviours.

Most drugs were assessed post model induction (475 comparisons) and whether they were administered pre- or post-model induction accounted for a significant amount of heterogeneity (SMD=1.173 [95% CI 0.921-1.434] and SMD=1.415 [95 % CI 1.227-1.602] Q=15.07, d.f. 2, p=0.005) (Figure 5.4B).

5.3.4.3. *Animal model characteristics*

Stratified meta-analyses were conducted to determine the influence of animal model characteristics on the observed effect sizes; The results are presented thematically below and with reference to both figures 5.5 (rats) and 5.6 (mice).

5.3.4.3.1. Model Type

In both rats and mice, the model type accounted for a significant proportion of heterogeneity (Q=36.56, d.f. 13, p<0.0025 and Q=173.03, d.f. 14, p<0.0001 respectively). Of the 15 model types, inflammation followed by nerve injury were modelled most frequently in both rats (269 and 259 comparisons) and mice (195 and 154 comparisons). In rats, the largest attenuation of pain-associated behaviour compared to control was reported in models of burn injury (SMD=2.23 [95% CI 0.33 – 4.14]) and the smallest significant attenuation was reported in models of inflammation (SMD=1.16 [95% CI 0.93 - 1.38]. The overall estimate of effect was significant for most rat model types except heat injury, migraine, and capsaicin models (Figure 5.5A). In mice, the largest significant attenuation of pain-associated behaviour compared to control was reported in the mouse model of HIV protein associated neuropathy (SMD=7.932 [95% CI 5.115 – 10.748], whereas the smallest significant attenuation was reported in nerve injury models (SMD=1.04 [95% CI 0.71 – 1.36]). The estimate of effect was significant in most mouse models, except sickle cell disease, visceral inflammation, multiple sclerosis, and mustard oil (Figure 5.6A).

5.3.4.3.2. Strain

Strain accounted for a significant proportion of heterogeneity; 12 different strains of rats (Q=39.41, d.f. 11, p<0.0001) and 29 different strains of mice were reported (Q=281.08, d.f. 28, p<0.0001). In rats, Sprague Dawley and Wistar strain were reported most (n=5148, 458 comparisons and n=3655, 361 comparisons respectively). All report significant effects except Wistar-Kyoto, Fischer 344, Wistar Han and Wistar albino, and the obese diabetic ZDF/crl-lepr/fa, as well as those strains that were not reported (Figure 5.5b). In mice, the C57BL/6J strain was most frequently reported (n=1794, 182 comparisons). The effects of the drugs were significant in over half of the strains but insignificant in 13 strains (Figure 5.6b).

5.3.4.3.3. Sex

In rats, sex accounted for a significant proportion of heterogeneity (Q=21.65, d.f. 3, p<0.001) (Figure 5c) whilst mouse sex did not (Q=7.39, d.f. 3. P=0.06) (Figure 5.6c). Most of the data are from male animals. Female only animal groups were used in 23 studies (6 %, n=1307), mixed sex groups in 8 studies (2 %, n=86) compared to 302 studies using male groups. In rats, 91 % of the experiments (n=8337, 786 comparisons) and 3 % (n = 650, 48 comparisons) of the experiments were conducted using male and female rats, respectively. To add, the differences between the groups is more likely due to the larger effect sizes reported by groups that did not state the sex of the animals or used mixed sex groups: the sex was not reported for 28 comparisons (n=213) and 4 comparisons (n=60) used mixed sex groups. In mice, 81 % of the experiments used male animals (n = 5574, 678 comparisons) and 9 % used female animals (n = 657, 66 comparisons). The sex was not reported for 43 comparisons (n=438) and 24 comparisons (n=207) used mixed sex groups.

5.3.4.4. Pain-Associated Behavioural Outcome Measures

In both rats (Figure 5.5d) and mice (Figure 5.6d), the type of pain-associated outcome measure accounted for a significant proportion of heterogeneity (rat; Q=160.28, d.f. 5, p<0.0001 and mouse; Q=20.96, d.f. 5 p<0.001). Evoked limb withdrawal to mechanical stimulation was most frequently reported (451 and 449 comparisons rats and mice respectively). Significant effects were observed for all outcome assessment types for rats, however, in mice, effect sizes for complex and non-evoked assessment types were not significant.

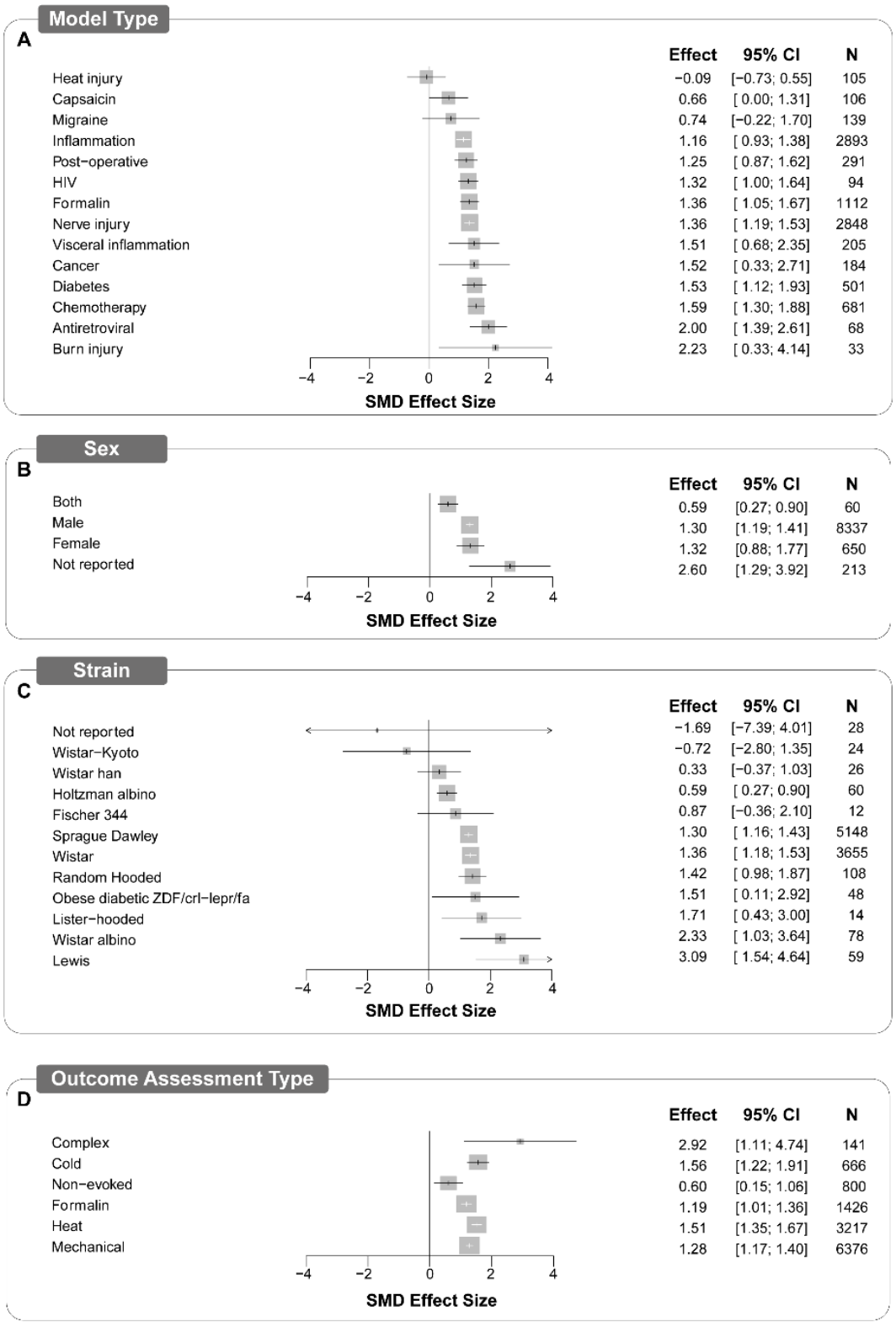


Figure 5.5. Forest plots of study design characteristics of experiments in which treatments were assessed for antinociceptive efficacy in rat models of persistent or injury-related persistent pain. Model type (A), strain (B), and sex (C) and outcome assessment type (D) account for a significant proportion of heterogeneity. The size of the square represents the weight (%). The weight is the influence that individual subgroup has on the pooled result. N denotes the number of animals that contribute to that subgroup.

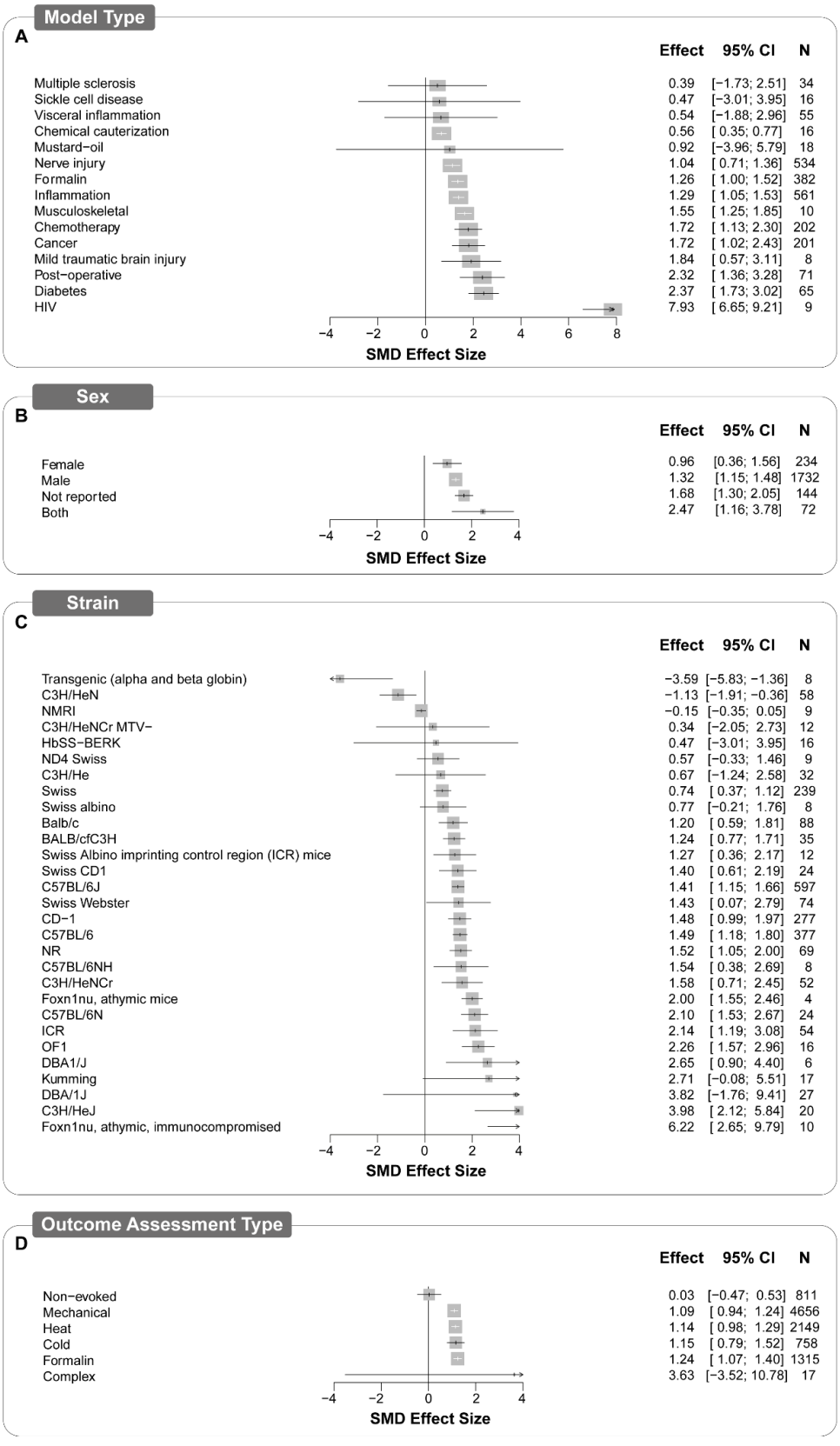


Figure 5.6. Forest plots of study design characteristics of experiments in which treatments were assessed for antinociceptive efficacy in mouse models of persistent or injury-related persistent pain. Model type (A), strain (B), sex (C), outcome assessment type (D) all account for a significant proportion of heterogeneity. The size of the square represents the weight (%). The weight is the influence that individual subgroup has on the pooled result. N denotes the number of animals that contribute to that subgroup.

5.3.5. Antinociceptive Efficacy of Drug Classes in Different Model Types

In rats, CB_1 and CB_2 receptor agonists were assessed most, with 10 and 9 model types, respectively. The antinociceptive effect of the two drug classes was significant in most model types (see Table 5.4 for drug class effects in individual model types); nerve injury, chemotherapy induced peripheral neuropathy (CIPN) and cancer models. However, CB_1 receptor agonists did not significantly attenuate pain-associated behaviour in inflammation and heat injury models (31 and 7 comparisons respectively). Within rat inflammation models, the results are mixed; the CB_1 receptor agonists did not significantly attenuate pain-associated behaviours in models of carrageenan (10 comparisons) and osteoarthritis (1 comparison). CB_2 receptor agonists did not significantly attenuate pain-associated behaviour in the rat formalin and migraine models (4 comparisons in each). FAAH inhibitors and PPAR-alpha agonists (exclusively PEA, OEA and analogues of PEA) were also assessed in a broad range of model types, 6 and 10 model types respectively. PPAR-alpha agonists significantly attenuated pain-associated behaviours across all model types whereas FAAH inhibitors demonstrated antinociceptive effects in neuropathic pain-associated models (e.g., nerve injury, CIPN and diabetes) but a mixed effect was observed in inflammation-associated models. FAAH inhibitors (comprising 26 individual drugs) significantly attenuated pain-associated behaviours in the Complete Freund's Adjuvant and formalin models but not in carrageenan (11 comparisons) or osteoarthritis (12 comparisons) models (see Table 5.5 for drug class effects in rat inflammation models).

Table 5.4. Drug class effects in rat neuropathic pain associated models. Rows highlighted blue show effect, highlighted orange no effect. 'x' denotes was not tested. N comparisons; number of comparisons, SMD; standardised mean difference, SE; standard error, CI; confidence interval.

	Nerve injury					Inflammation					Formalin					Chemotherapy					Diabetes				
Drug Class	**N comparisons**	**SMD**	**SE**	**Lower CI**	**Upper CI**	**N comparisons**	**SMD**	**SE**	**Lower CI**	**Upper CI**	**N comparisons**	**SMD**	**SE**	**Lower CI**	**Upper CI**	**N comparisons**	**SMD**	**SE**	**Lower CI**	**Upper CI**	**N comparisons**	**SMD**	**SE**	**Lower CI**	**Upper CI**
Anandamide transport inhibitor	14	1.57104	0.41456	0.759	2.384	5	0.67201	0.20454	0.271	1.073	18	1.15697	0.34871	0.474	1.84	x	x	x	x	x	4	1.49296	0.48232	0.548	2.438
CB_1 receptor agonist	69	1.33929	0.17816	0.99	1.688	31	0.37222	0.28617	-0.189	0.933	29	1.54535	0.38946	0.782	2.309	25	1.5451	0.28012	0.996	2.094	22	1.64705	0.28369	1.091	2.203
CB_1 receptor inverse agonist	4	1.70929	0.27469	1.171	2.248	5	0.43869	0.19581	0.055	0.822	1	1.71445	0.65712	0.427	3.002	x	x	x	x	x	x	x	x	x	x
CB_2 receptor agonist	54	1.69281	0.20208	1.297	2.089	47	1.41355	0.26799	0.888	1.939	4	0.3689	0.9812	-1.554	2.292	25	1.4446	0.21032	1.032	1.857	3	0.71656	0.2288	0.268	1.165
CB_2 receptor inverse agonist	1	0.95896	0.80411	-0.617	2.535	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CBD	6	1.74962	0.43383	0.899	2.6	19	0.89797	0.17818	0.549	1.247	x	x	x	x	x	x	x	x	x	x	7	1.0765	0.41823	0.257	1.896
FAAH inhibitor	30	1.70937	0.35433	1.015	2.404	35	0.25476	0.28316	-0.3	0.81	8	1.4183	0.17616	1.073	1.764	11	2.64754	0.3943	1.875	3.42	14	3.24878	0.89013	1.504	4.993
FABP inhibitor	1	0.86684	0.62706	-0.362	2.096	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
GPR55 agonist	1	-0.3794	0.71953	-1.79	1.031	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Hemp oil	1	0.09013	0.7874	-1.453	1.633	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Monoacylglycerol lipase inhibitor	3	1.28821	0.43878	0.428	2.148	3	0.03673	0.03932	-0.04	0.114	4	2.79987	0.78568	1.26	4.34	3	2.65747	0.60522	1.471	3.844	x	x	x	x	x
NAAA inhibitor	x	x	x	x	x	6	1.58857	0.21591	1.165	2.012	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Non-selective cannabinoid receptor agonist	52	1.4003	0.17505	1.057	1.743	57	1.17661	0.29422	0.6	1.753	23	1.82661	0.33175	1.176	2.477	3	1.46793	0.23028	1.017	1.919	4	0.83073	0.54581	-0.239	1.9
PPAR-alpha agonist	17	0.45291	0.13119	0.196	0.71	44	2.09248	0.40317	1.302	2.883	12	1.18422	0.6478	-0.085	2.454	4	1.4313	0.71822	0.024	2.839	x	x	x	x	x
PPAR-gamma antagonist	x	x	x	x	x	x	x	x	x	x	2	-0.7247	1.05762	-2.798	1.348	x	x	x	x	x	x	x	x	x	x
THC	6	1.188	0.31109	0.578	1.798	17	2.2766	0.31439	1.66	2.893	7	0.74858	0.18459	0.387	1.11	x	x	x	x	x	x	x	x	x	x
Unknown mechanism of action	x	x	x	x	x	x	x	x	x	x	4	1.50467	0.52115	0.483	2.526	x	x	x	x	x	x	x	x	x	x

	Cancer					Post-operative					Visceral Inflammation					Migraine					HIV				
Drug Class	**N comparisons**	**SMD**	**SE**	**Lower CI**	**Upper CI**	**N comparisons**	**SMD**	**SE**	**Lower CI**	**Upper CI**	**N comparisons**	**SMD**	**SE**	**Lower CI**	**Upper CI**	**N comparisons**	**SMD**	**SE**	**Lower CI**	**Upper CI**	**N comparisons**	**SMD**	**SE**	**Lower CI**	**Upper CI**
Anandamide transport inhibitor	x	x	x	x	x	x	x	x	x	x	5	2.44669	0.54502	1.378	3.515	x	x	x	x	x	x	x	x	x	x
CB_1 receptor agonist	3	0.66953	0.26998	0.14	1.199	7	0.60535	0.54576	-0.464	1.675	x	x	x	x	x	x	x	x	x	x	2	1.50484	0.09566	1.317	1.692
CB_1 receptor inverse agonist	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CB_2 receptor agonist	5	2.23617	0.94198	0.39	4.082	14	1.30058	0.29192	0.728	1.873	x	x	x	x	x	4	1.09718	1.19826	-1.251	3.446	x	x	x	x	x
CB_2 receptor inverse agonist	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CBD	x	x	x	x	x	9	1.82857	0.27916	1.281	2.376	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
FAAH inhibitor	x	x	x	x	x	x	x	x	x	x	4	0.07168	1.33994	-2.555	2.698	3	0.18476	1.80767	-3.358	3.728	6	1.2208	0.2267	0.776	1.665
FABP inhibitor	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
GPR55 agonist	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Hemp oil	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Monoacylglycerol lipase inhibitor	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
NAAA inhibitor	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Non-selective cannabinoid receptor agonist	x	x	x	x	x	3	1.40464	0.25995	0.895	1.914	x	x	x	x	x	1	2.56499	1.14466	0.322	4.808	x	x	x	x	x
PPAR-alpha agonist	x	x	x	x	x	x	x	x	x	x	6	1.49029	0.27447	0.952	2.028	x	x	x	x	x	x	x	x	x	x
PPAR-gamma antagonist	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
THC	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	5	0.51298	0.23485	0.053	0.973	x	x	x	x	x
Unknown mechanism of action	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

	Capsaicin					Heat injury					Antiretroviral					Brun injury				
Drug Class	**N comparisons**	**SMD**	**SE**	**Lower CI**	**Upper CI**	**N comparisons**	**SMD**	**SE**	**Lower CI**	**Upper CI**	**N comparisons**	**SMD**	**SE**	**Lower CI**	**Upper CI**	**N comparisons**	**SMD**	**SE**	**Lower CI**	**Upper CI**
Anandamide transport inhibitor	1	-0.2978	0.78378	-1.834	1.238	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CB_1 receptor agonist	x	x	x	x	x	6	-0.1353	0.10079	-0.333	0.062	x	x	x	x	x	x	x	x	x	x
CB_1 receptor inverse agonist	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	3	2.23199	0.97181	0.327	4.137
CB_2 receptor agonist	3	1.49242	0.66581	0.187	2.797	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CB_2 receptor inverse agonist	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CBD	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
FAAH inhibitor	2	0.56669	0.42829	-0.273	1.406	1	12.1013	3.86917	4.518	19.685	x	x	x	x	x	x	x	x	x	x
FABP inhibitor	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
GPR55 agonist	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Hemp oil	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Monoacylglycerol lipase inhibitor	3	-0.0174	0.294	-0.594	0.559	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
NAAA inhibitor	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Non-selective cannabinoid receptor agonist	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
PPAR-alpha agonist	x	x	x	x	x	x	x	x	x	x	3	2.00046	0.31348	1.386	2.615	x	x	x	x	x
PPAR-gamma antagonist	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
THC	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Unknown mechanism of action	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Table 5.5. Drug class effects in rat inflammation models. Rows highlighted blue show effect, highlighted orange no effect. 'x' denotes was not tested. N comparisons; number of comparisons, SMD; standardised mean difference, SE; standard error, CI; confidence interval.

	Carrageenan					Complete Freund's Adjuvant					Osteoarthritis					Prostaglandin E2					Nerve Growth Factor				
Drug Class	**K**	**SMD**	**SE**	**Lower CI**	**Upper CI**	**K**	**SMD**	**SE**	**Lower CI**	**Upper CI**	**K**	**SMD**	**SE**	**Lower CI**	**Upper CI**	**K**	**SMD**	**SE**	**Lower CI**	**Upper CI**	**K**	**SMD**	**SE**	**Lower CI**	**Upper CI**
Anandamide transport inhibitor	x	x	x	x	x	5	0.67201	0.20454	0.271	1.073	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CB_1 receptor agonist	10	-1.46447	1.13712	-3.693	0.764	12	0.53719	0.2586	0.03	1.044	1	1.03897	0.89421	-0.714	2.792	8	1.64723	0.42752	0.809	2.485	x	x	x	x	x
CB_1 receptor inverse agonist	5	0.43869	0.19581	0.055	0.822	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CB_2 receptor agonist	10	1.39073	0.25786	0.885	1.896	18	1.02865	0.21647	0.604	1.453	19	2.11323	0.69738	0.746	3.48	x	x	x	x	x	x	x	x	x	x
CB_2 receptor inverse agonist	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CBD	10	0.71715	0.10873	0.504	0.93	5	1.58956	0.44889	0.71	2.469	4	0.81394	0.59003	-0.342	1.97	x	x	x	x	x	x	x	x	x	x
FAAH inhibitor	11	-0.3767	0.74444	-1.836	1.082	11	1.10186	0.26285	0.587	1.617	12	0.23059	0.29161	-0.341	0.802	x	x	x	x	x	x	x	x	x	x
FABP inhibitor	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
GPR55 agonist	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Hemp oil	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Monoacylglycerol lipase inhibitor	x	x	x	x	x	x	x	x	x	x	3	0.03673	0.03932	-0.04	0.114	x	x	x	x	x	x	x	x	x	x
NAAA inhibitor	3	1.55426	0.43615	0.699	2.409	3	1.63016	0.20147	1.235	2.025	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Non-selective cannabinoid receptor agonist	39	0.79319	0.34236	0.122	1.464	7	1.88537	0.42908	1.044	2.726	1	1.01718	0.89112	-0.729	2.764	6	11.1581	1.99419	7.25	15.067	2	1.61085	1.1826	-0.707	3.929
PPAR-alpha agonist	20	4.3084	1.27432	1.811	6.806	2	2.52808	0.85582	0.851	4.205	9	0.98105	0.13293	0.721	1.242	9	6.54709	2.17691	2.28	10.814	2	2.00702	0.72122	0.593	3.421
PPAR-gamma antagonist	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
THC	1	1.28127	1.13456	-0.942	3.505	16	2.33634	0.33133	1.687	2.986	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Unknown mechanism of action	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

In mice, CB_2 receptor agonists and FAAH inhibitors were assessed most across 10 and 8 model types respectively (see Tables 5.6 and 5.7 for full drug class effects in mouse models). CB_2 receptor agonists significantly attenuated pain-associated behaviours in mouse cancer and visceral inflammation models but not in nerve injury (33 comparisons), multiple sclerosis model (6 comparisons) and CIPN (2 comparisons) models. FAAH inhibitors significantly attenuated pain-associated behaviours in inflammation (38 comparisons), nerve injury (29 comparisons), formalin (15 comparisons), post-operative (4 comparisons), and diabetes (3 comparisons) models but not in CIPN (8 comparisons), visceral inflammation (5 comparisons) and mustard oil (1 comparison) models. CB_1 and CB_2 receptor agonists were both assessed in 6 model types and significantly attenuated pain-associated behaviours in all model types except nerve injury. Like rats, PPAR-alpha agonists significantly attenuated pain-associated behaviours in the 4 model types in which they were assessed; nerve injury, inflammation, formalin, and CIPN models.

Table 5.6. Drug class effects in mouse neuropathic pain-associated models. Rows highlighted blue show effect, highlighted orange no effect. 'x' denotes was not tested. N comparisons; number of comparisons, SMD; standardised mean difference, SE; standard error, CI; confidence interval.

	Nerve injury					Inflammation					Formalin					Chemotherapy					Diabetes				
Drug Class	**N comparisons**	**SMD**	**SE**	**Lower CI**	**Upper CI**	**N comparisons**	**SMD**	**SE**	**Lower CI**	**Upper CI**	**N comparisons**	**SMD**	**SE**	**Lower CI**	**Upper CI**	**N comparisons**	**SMD**	**SE**	**Lower CI**	**Upper CI**	**N comparisons**	**SMD**	**SE**	**Lower CI**	**Upper CI**
ABHD6 inhibitor	3	0.27282	0.23671	-0.191	0.737	5	1.61548	0.89731	-0.143	3.374	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Anandamide transport inhibitor	x	x	x	x	x	x	x	x	x	x	8	1.05967	0.32848	0.416	1.703	x	x	x	x	x	x	x	x	x	x
CB_1 receptor agonist	20	0.14361	0.49392	-0.824	1.112	16	1.34337	0.22289	0.907	1.78	14	0.71127	0.3287	0.067	1.356	6	1.10081	0.42269	0.272	1.929	4	3.55789	1.37566	0.862	6.254
CB_1 receptor inverse agonist	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	2	4.52933	1.32528	1.932	7.127
CB_1 receptor PAM	4	0.70301	0.43186	-0.143	1.549	1	2.45202	0.70659	1.067	3.837	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CB_2 receptor agonist	33	0.08534	0.21511	-0.336	0.507	28	1.18214	0.16506	0.859	1.506	29	0.89633	0.22406	0.457	1.335	2	4.05731	2.14087	-0.139	8.253	7	1.55942	0.33375	0.905	2.214
CB_2 receptor inverse agonist	x	x	x	x	x	x	x	x	x	x	3	1.0487	0.20191	0.653	1.444	x	x	x	x	x	x	x	x	x	x
CBD	3	2.00786	0.27015	1.478	2.537	3	-0.68146	0.34962	-1.367	0.004	2	1.8584	0.15837	1.548	2.169	10	1.30772	0.36393	0.594	2.021	x	x	x	x	x
Diacylglycerol lipase inhibitor	1	3.69134	1.25331	1.235	6.148	10	1.45344	0.51812	0.438	2.469	x	x	x	x	x	3	2.2707	1.4209	-0.514	5.056	x	x	x	x	x
Dual FAAH/MAGL inhibitor	2	2.07785	0.37257	1.348	2.808	6	1.69614	0.67643	0.37	3.022	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Endocannabinoid uptake inhibitor	x	x	x	x	x	1	-0.71883	0.66531	-2.023	0.585	3	2.4799	0.16946	2.148	2.812	x	x	x	x	x	x	x	x	x	x
FAAH inhibitor	29	1.66713	0.39759	0.888	2.446	38	1.385	0.31048	0.776	1.994	15	1.77527	0.34659	1.096	2.455	8	1.09442	0.74607	-0.368	2.557	3	2.59614	0.34913	1.912	3.28
FAAH inhibitor/TRPA1 agonist	x	x	x	x	x	3	-2.19377	0.87249	-3.904	-0.484	2	3.52467	2.76009	-1.885	8.934	x	x	x	x	x	x	x	x	x	x
FAAH inhibitor/TRPV1 agonist	x	x	x	x	x	x	x	x	x	x	2	2.67491	2.09755	-1.436	6.786	x	x	x	x	x	x	x	x	x	x
FABP inhibitor	x	x	x	x	x	21	0.96079	0.23045	0.509	1.412	5	0.73533	0.64755	-0.534	2.005	x	x	x	x	x	x	x	x	x	x
Monoacylglycerol lipase inhibitor	13	1.34953	0.33303	0.697	2.002	9	2.50333	0.76369	1.007	4	4	1.8711	0.51571	0.86	2.882	9	1.04584	0.75935	-0.442	2.534	x	x	x	x	x
NAAA inhibitor	5	4.54488	0.99525	2.594	6.496	9	2.51948	0.77667	0.997	4.042	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Non-selective cannabinoid receptor agonist	23	0.45613	0.41607	-0.359	1.272	22	0.1488	0.39907	-0.633	0.931	15	1.16682	0.48072	0.225	2.109	6	1.9835	1.60727	-1.167	5.134	4	3.13626	1.09413	0.992	5.281
PPAR-alpha agonist	9	2.88627	1.05724	0.814	4.958	8	2.14529	0.73526	0.704	3.586	9	2.3539	0.82839	0.73	3.978	7	3.24916	0.70591	1.866	4.633	x	x	x	x	x
THC	3	1.48173	0.12516	1.236	1.727	4	2.40301	1.28708	-0.12	4.926	13	1.74776	0.41988	0.925	2.571	6	1.78732	0.96445	-0.103	3.678	x	x	x	x	x
Unknown mechanism of action	6	0.953	1.80187	-2.579	4.485	12	1.37145	0.31073	0.762	1.98	2	9.6161	10.563	-11.087	30.319	x	x	x	x	x	x	x	x	x	x

	Cancer					Post-operative					Visceral inflammation					Chemical cauterisation					HIV				
Drug Class	**N comparisons**	**SMD**	**SE**	**Lower CI**	**Upper CI**	**N comparisons**	**SMD**	**SE**	**Lower CI**	**Upper CI**	**N comparisons**	**SMD**	**SE**	**Lower CI**	**Upper CI**	**N comparisons**	**SMD**	**SE**	**Lower CI**	**Upper CI**	**N comparisons**	**SMD**	**SE**	**Lower CI**	**Upper CI**
ABHD6 inhibitor	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Anandamide transport inhibitor	4	-1.1781	0.56643	-2.288	-0.068	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CB_1 receptor agonist	20	1.61608	0.53104	0.575	2.657	7	4.68375	0.8464	3.025	6.343	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CB_1 receptor inverse agonist	x	x	x	x	x	x	x	x	x	x	2	1.90166	0.30167	1.31	2.493	x	x	x	x	x	x	x	x	x	x
CB_1 receptor PAM	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CB_2 receptor agonist	23	2.06862	0.38533	1.313	2.824	x	x	x	x	x	1	1.53783	0.58996	0.382	2.694	4	0.53037	0.23194	0.076	0.985	x	x	x	x	x
CB_2 receptor inverse agonist	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CBD	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	4	0.74344	0.28885	0.177	1.31	x	x	x	x	x
Diacylglycerol lipase inhibitor	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Dual FAAH/MAGL inhibitor	x	x	x	x	x	x	x	x	x	x	2	0.5671	0.01871	0.53	0.604	x	x	x	x	x	x	x	x	x	x
Endocannabinoid uptake inhibitor	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
FAAH inhibitor	x	x	x	x	x	4	0.66081	0.10853	0.448	0.874	5	4.22914	4.52423	-4.638	13.096	x	x	x	x	x	x	x	x	x	x
FAAH inhibitor/TRPA1 agonist	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
FAAH inhibitor/TRPV1 agonist	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
FABP inhibitor	x	x	x	x	x	x	x	x	x	x	4	-1.88512	0.60896	-3.079	-0.692	x	x	x	x	x	x	x	x	x	x
Monoacylglycerol lipase inhibitor	3	-0.69635	0.39802	-1.476	0.084	4	1.24201	0.75896	-0.246	2.73	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
NAAA inhibitor	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Non-selective cannabinoid receptor agonist	7	1.26762	1.45504	-1.584	4.119	4	2.09672	0.53822	1.042	3.152	1	0.03505	0.79063	-1.515	1.585	x	x	x	x	x	3	7.9316	0.65454	6.649	9.214
PPAR-alpha agonist	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
THC	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Unknown mechanism of action	x	x	x	x	x	x	x	x	x	x	1	2.11847	0.66035	0.824	3.413	8	0.48764	0.12787	0.237	0.738	x	x	x	x	x

	Multiple sclerosis					Musculoskeletal					Mustard Oil					Sickle cell disease					Mild traumatic brain injury				
Drug Class	**N comparisons**	**SMD**	**SE**	**Lower CI**	**Upper CI**	**N comparisons**	**SMD**	**SE**	**Lower CI**	**Upper CI**	**N comparisons**	**SMD**	**SE**	**Lower CI**	**Upper CI**	**N comparisons**	**SMD**	**SE**	**Lower CI**	**Upper CI**	**N comparisons**	**SMD**	**SE**	**Lower CI**	**Upper CI**
ABHD6 inhibitor	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Anandamide transport inhibitor	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CB_1 receptor agonist	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CB_1 receptor inverse agonist	x	x	x	x	x	x	x	x	x	x	2	3.13423	1.02032	1.134	5.134	x	x	x	x	x	x	x	x	x	x
CB_1 receptor PAM	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CB_2 receptor agonist	6	-0.18487	1.10219	-2.345	1.975	4	1.54966	0.15461	1.247	1.853	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CB_2 receptor inverse agonist	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CBD	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Diacylglycerol lipase inhibitor	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Dual FAAH/MAGL inhibitor	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Endocannabinoid uptake inhibitor	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
FAAH inhibitor	x	x	x	x	x	x	x	x	x	x	1	-3.93714	1.13795	-6.167	-1.707	x	x	x	x	x	x	x	x	x	x
FAAH inhibitor/TRPA1 agonist	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
FAAH inhibitor/TRPV1 agonist	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
FABP inhibitor	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Monoacylglycerol lipase inhibitor	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
NAAA inhibitor	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Non-selective cannabinoid receptor agonist	1	3.50612	0.52005	2.487	4.525	x	x	x	x	x	x	x	x	x	x	2	0.47169	1.77531	-3.008	3.951	1	1.83875	0.64907	0.567	3.111
PPAR-alpha agonist	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
THC	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Unknown mechanism of action	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Table 5.7. Drug class effects in mouse inflammation models. Rows highlighted blue show effect, highlighted orange no effect. 'x' denotes was not tested. N comparisons; number of comparisons, SMD; standardised mean difference, SE; standard error, CI; confidence interval.

	Carrageenan					Complete Freund's Adjuvant					Osteoarthritis					Lipopolysaccharide					Collagen-induced arthritis				
Drug Class	**K**	**SMD**	**SE**	**Lower CI**	**Upper CI**	**K**	**SMD**	**SE**	**Lower CI**	**Upper CI**	**K**	**SMD**	**SE**	**Lower CI**	**Upper CI**	**K**	**SMD**	**SE**	**Lower CI**	**Upper CI**	**K**	**SMD**	**SE**	**Lower CI**	**Upper CI**
Anandamide transport inhibitor	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	1	-0.7188	0.66531	-2.023	0.585	x	x	x	x	x
CB_1 receptor agonist	5	0.80951	0.30804	0.206	1.413	4	2.05959	0.45117	1.175	2.944	2	1.25897	0.00267	1.254	1.264	1	2.24939	0.77442	0.732	3.767	x	x	x	x	x
CB_1 receptor inverse agonist	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CB_1 receptor PAM	1	2.45202	0.70659	1.067	3.837	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CB_2 receptor agonist	x	x	x	x	x	26	1.17637	0.18341	0.817	1.536	2	1.33653	0.32937	0.691	1.982	x	x	x	x	x	x	x	x	x	x
CB_2 receptor inverse agonist	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CBD	3	-0.68146	0.34962	-1.367	0.004	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Diacylglycerol lipase inhibitor	4	1.42487	0.59961	0.25	2.6	2	3.86029	2.93149	-1.885	9.606	x	x	x	x	x	10	1.45344	0.51812	0.438	2.469	x	x	x	x	x
Dual FAAH/MAGL inhibitor	19	1.49145	0.51735	0.477	2.505	13	1.74794	0.64436	0.485	3.011	1	1.58541	0.59515	0.419	2.752	x	x	x	x	x	x	x	x	x	x
FAAH inhibitor	3	-2.19377	0.87249	-3.904	-0.484	x	x	x	x	x	x	x	x	x	x	3	0.76689	0.34649	0.088	1.446	1	2.64744	0.89285	0.897	4.397
FAAH inhibitor/TRPV1 agonist	6	1.42062	0.24187	0.947	1.895	15	0.78331	0.30083	0.194	1.373	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
FABP inhibitor	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
GPR55 agonist	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Hemp oil	8	2.67287	0.87019	0.967	4.378	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Monoacylglycerol lipase inhibitor	9	2.51948	0.77667	0.997	4.042	1	1.65792	0.87841	-0.064	3.38	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
NAAA inhibitor	6	0.25991	0.23776	-0.206	0.726						x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Non-selective cannabinoid receptor agonist	6	1.31701	0.40151	0.53	2.104	16	0.03876	0.57983	-1.098	1.175	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
PPAR-alpha agonist	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	2	3.82494	2.84928	-1.76	9.409
PPAR-gamma antagonist	4	2.40301	1.28708	-0.12	4.926	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
THC	12	1.37145	0.31073	0.762	1.98	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Unknown mechanism of action	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

5.3.6. Antinociceptive Efficacy of Drug Classes in Different Model Types Assessed using Mechanical Limb Withdrawal

In rat mechanical hypersensitivity outcomes, comparisons of the drug classes and model types did not reveal a consistent difference between the most frequently reported neuropathic pain associated models and other model types (Table 5.8). However, tCB_1 receptor agonists and non-selective cannabinoid receptor agonists did not attenuate mechanical hypersensitivity in inflammation models yet did in the three neuropathic pain associated models (Table 5.9). In mice, there was one difference of interest: CB_1 receptor agonists did not attenuate mechanical hypersensitivity in nerve injury models though did in both chemotherapy and diabetes models as well as inflammation and cancer models.

Table 5.8. Drug class effects assessed by mechanical limb withdrawal in rats. (A) most frequently assessed other model types (B) most frequently assessed neuropathic pain associated models. Rows highlighted blue show effect, highlighted orange no effect. 'x' denotes was not tested. N comparisons; number of comparisons, SMD; standardised mean difference, SE; standard error, CI; confidence interval.

A

	Inflammation (SMD = 0.862 [0.59-1.59])					Post-operative (SMD = 1.141 [1.01 - 1.27])					Cancer				
Drug Class	K	SMD	SE	Lower CI	Upper CI	K	SMD	SE	Lower CI	Upper CI	K	SMD	SE	Lower CI	Upper CI
Anandamide transport inhibitor	2	2.620093	0	0	0	x	x	x	x	x	x	x	x	x	x
CB_1 receptor agonist	38	0.17858	0.279519	-0.369	0.726	7	0.926075	0.061951	0.805	1.047	3	4.237708	1.510734	1.277	7.199
CB_1 receptor inverse agonist	5	0.392303	0.151583	0.095	0.689	x	x	x	x	x	x	x	x	x	x
CB_2 receptor agonist	19	1.1317	0.21898	0.703	1.561	14	1.322681	0.135136	1.058	1.588	3	3.047206	0.424958	2.214	3.88
CBD	5	1.750348	0.410302	0.946	2.555	9	1.206834	0.081201	1.048	1.366	x	x	x	x	x
FAAH inhibitor	18	1.249208	0.094797	1.063	1.435	x	x	x	x	x	x	x	x	x	x
MAGL inhibitor	2	3.701183	1.0417	1.659	5.743	x	x	x	x	x	x	x	x	x	x
Non-selective cannabinoid receptor agonist	30	0.121397	0.239781	-0.349	0.591	3	2.045714	0.15751	1.737	2.354	x	x	x	x	x
PPAR-alpha agonist	27	0.884705	0.365216	0.169	1.601	x	x	x	x	x	x	x	x	x	x
THC	12	3.250993	0.127486	3.001	3.501	x	x	x	x	x	x	x	x	x	x

B

	Nerve injury (SMD = 1.439 [1.31 - 1.57])					Chemotherapy (SMD = 1.576 [1.271 - 1.88])					Diabetes				
Drug Class	K	SMD	SE	Lower CI	Upper CI	K	SMD	SE	Lower CI	Upper CI	K	SMD	SE	Lower CI	Upper CI
Anandamide transport inhibitor	8	1.566173	0.254599	1.067	2.065	x	x	x	x	x	x	x	x	x	x
CB_1 receptor agonist	71	1.132553	0.110027	0.917	1.348	24	0.06097	0.029466	0.003	0.119	16	2.044389	0.341951	1.374	2.715
CB_1 receptor inverse agonist	3	2.471623	0.28108	1.921	3.023	x	x	x	x	x	x	x	x	x	x
CB_2 receptor agonist	59	1.578806	0.151738	1.281	1.876	28	1.215273	0.181782	0.859	1.572	3	1.077705	0.272209	0.544	1.611
CB_2 receptor inverse agonist	1	1.483119	1.546585	-1.548	4.514	x	x	x	x	x	x	x	x	x	x
CBD	7	1.724561	0.776195	0.203	3.246	x	x	x	x	x	x	x	x	x	x
FAAH inhibitor	35	1.758066	0.132106	1.499	2.017	5	1.867804	0.183829	1.508	2.228	7	1.667965	0.21931	1.238	2.098
FABP inhibitor	2	1.33586	0	0	0	x	x	x	x	x	x	x	x	x	x
GPR55 agonist	1	-0.73283	1.931518	-4.519	3.053	x	x	x	x	x	x	x	x	x	x
Hemp oil	1	-0.75726	1.6303	-3.953	2.438	x	x	x	x	x	x	x	x	x	x
MAGL inhibitor	3	1.366042	0.169642	1.034	1.699	3	1.556084	0.087773	1.384	1.728	20	0.806706	0.077022	0.656	0.958
Non-selective cannabinoid receptor agonist	39	1.720876	0.094356	1.536	1.906	3	1.401503	0.071958	1.26	1.543	x	x	x	x	x
PPAR-alpha agonist	12	0.451797	0.02397	0.405	0.499	2	5.772671	0.587699	4.621	6.925	x	x	x	x	x
THC	3	2.952317	1.240966	0.52	5.385	x	x	x	x	x	x	x	x	x	x

Table 5.9. Drug class effects assessed by mechanical limb withdrawal in mice. (A) most frequently assessed other model types (B) most frequently assessed neuropathic pain associated models. Rows highlighted blue show effect, highlighted orange no effect. 'x' denotes was not tested. N comparisons; number of comparisons, SMD; standardised mean difference, SE; standard error, CI; confidence interval.

A

	Inflammation SMD = 1.0420 [0.8082; 1.2757]					Cancer SMD = 1.2376 [0.7063; 1.7688]				
Drug Class	**K**	**SMD**	**SE**	**Lower CI**	**Upper CI**	**K**	**SMD**	**SE**	**Lower CI**	**Upper CI**
Anandamide uptake inhibitor	x	x	x	x	x	3	-0.45709	0.186538	-0.823	-0.091
CB_1 receptor agonist	16	1.725784	0.118575	1.493	1.958	11	1.039174	0.226399	0.595	1.483
CB_1 receptor PAM	1	4.911206	2.002926	0.986	8.837	x	x	x	x	x
CB_2 receptor agonist	25	1.324226	0.120433	1.088	1.56	15	2.004417	0.414825	1.191	2.817
CBD	3	-2.91948	0.515207	-3.929	-1.91	x	x	x	x	x
Diacylglyerol lipase inhibitor	11	1.149421	0.192837	0.771	1.527	x	x	x	x	x
FAAH inhibitor	23	0.439557	0.129181	0.186	0.693	x	x	x	x	x
FAAH/MAGL inhibitor	6	1.33261	0.082773	1.17	1.495	x	x	x	x	x
FABP inhibitor	5	1.90392	0.19408	1.524	2.284	x	x	x	x	x
MAGL inhibitor	5	1.634692	0.190933	1.26	2.009	3	-0.05419	0.005313	-0.065	-0.044
Non-selective cannabinoid receptor agnoist	19	-0.04129	0.392374	-0.81	0.728	10	1.223989	0.813988	-0.371	2.819
Paracetamol	4	1.334029	0.098166	1.142	1.526	x	x	x	x	x
PPAR-alpha agonist	7	1.968789	0.419151	1.147	2.79	x	x	x	x	x
Unknown mechanism of action	9	1.554302	0.17743	1.207	1.902	x	x	x	x	x

B

	Nerve injury SMD = 0.7656 [0.4669; 1.0644]					Chemotherapy SMD = 1.0892 [0.6624; 1.5160]					Diabetes SMD = 1.2120 [0.9016; 1.5225]				
Drug Class	K	SMD	SE	Lower CI	Upper CI	K	SMD	SE	Lower CI	Upper CI	K	SMD	SE	Lower CI	Upper CI
ABHD6 inhibitor	3	0.398624	0.454907	-0.493	1.29	x	x	x	x	x	x	x	x	x	x
CB_1 receptor agonist	20	0.046897	0.315491	-0.571	0.665	6	1.141136	0.18104	0.786	1.496	4	1.031099	0.133757	0.769	1.293
CB_1 receptor inverse agonist	x	x	x	x	x	x	x	x	x	x	2	1.495237	0.222968	1.058	1.932
CB_1 receptor PAM	4	1.055714	0.732396	-0.38	2.491	x	x	x	x	x	x	x	x	x	x
CB_2 receptor agonist	29	0.59089	0.289166	0.024	1.158	2	1.632361	0.326045	0.993	2.271	7	1.148679	0.113661	0.926	1.371
CBD	3	1.139723	0.637436	-0.11	2.389	7	-0.05831	0.036414	-0.13	0.013	x	x	x	x	x
Diacylglyerol lipase inhibitor	1	1.371786	0.20747	0.965	1.778	3	2.541011	0.472759	1.614	3.468	x	x	x	x	x
FAAH inhibitor	30	1.085367	0.360943	0.378	1.793	x	x	x	x	x	3	6.612814	0.13983	6.339	6.887
FAAH/MAGL inhibitor	2	1.703099	0.011552	1.68	1.726	8	0.878303	0.572599	-0.244	2.001	x	x	x	x	x
MAGL inhibitor	13	1.194413	0.072379	1.053	1.336	5	0.228179	1.488326	-2.689	3.145	x	x	x	x	x
NAAA inhibitor	5	1.275476	0.151964	0.978	1.573	x	x	x	x	x	x	x	x	x	x
Non-selective cannabinoid receptor agnoist	17	-0.17463	0.386672	-0.932	0.583	6	0.788617	0.863045	-0.903	2.48	x	x	x	x	x
PPAR-alpha agonist	10	4.133137	0.79323	2.578	5.688	7	1.701888	0.278768	1.156	2.248	x	x	x	x	x
THC	3	2.232748	0.209593	1.822	2.644	6	0.720228	0.277784	0.176	1.265	x	x	x	x	x
Unknown mechanism of action	3	0.370788	1.646431	-2.856	3.598	x	x	x	x	x	x	x	x	x	x

5.3.7. Exploring Sources of Heterogeneity with Meta-Regression

A meta-regression was performed to further investigate further or differences between the stratified analysis, the study design characteristics that may influence the size of the intervention effect. In rats, the forced entry meta-regression model (mixed effects with maximum likelihood estimator), the variables sex, strain, model type, drug class and outcome assessment type accounted for (R^2) 36 % of heterogeneity (I^2 (unaccounted variability) 41 % and $\tau^2 = 0.531$). For each predictor, the coefficients estimates are listed in Table 5.10. It explains which variables are likely associated with higher effect sizes (positive coefficient estimates) and those associated with lower effect sizes (negative coefficient estimates)(Table 5.10).

Table 5.10. The rat coefficient estimates from multivariable meta-regression of the variables drug class, model type, outcome assessment, animal strain and sex. The redundant predictors were removed from the model. Significance levels; *** P< 0.001, ** P< 0.01, * P< 0.05.

Predictor levels	Coefficient Estimate	SE	t-value	p-value	95 % CI Lower	95 % CI Upper	Significance Level
CB_1 receptor agonist	-0.030	0.2609	-0.1144	0.909	-0.542	0.482	
CB_1 receptor inverse agonist	-0.009	0.5395	-0.0171	0.9863	-1.068	1.050	
CB_2 receptor agonist	0.351	0.2747	1.2792	0.2012	-0.188	0.891	
CB_2 receptor inverse agonist	-0.177	1.3559	-0.1302	0.8965	-2.838	2.485	
CBD	0.020	0.3355	0.0605	0.9518	-0.638	0.679	
FAAH inhibitor	0.179	0.2818	0.6335	0.5266	-0.375	0.732	
FABP inhibitor	0.911	1.3149	0.6926	0.4887	-1.670	3.492	
GPR55 agonist	-1.535	1.2789	-1.2006	0.2303	-4.046	0.975	
Hemp oil	-1.351	1.3372	-1.0106	0.3125	-3.976	1.273	
Monoacylglycerol lipase inhibitor	0.613	0.4658	1.3168	0.1883	-0.301	1.528	
NAAA inhibitor	0.544	0.5904	0.9221	0.3567	-0.614	1.703	
Non-selective cannabinoid receptor agonist	0.251	0.2618	0.9593	0.3377	-0.263	0.765	
PPAR-alpha agonist	0.401	0.2889	1.3861	0.1661	-0.167	0.968	
PPAR-gamma antagonist	-0.375	1.0138	-0.3694	0.7119	-2.365	1.616	
THC	0.682	0.3769	1.8102	0.0706	-0.058	1.422	
Unknown mechanism of action	0.284	0.8668	0.3273	0.7435	-1.418	1.985	
Burn Injury	0.207	1.1869	0.1741	0.8618	-2.123	2.536	
Cancer	-0.559	0.841	-0.6643	0.5067	-2.210	1.092	
Capsaicin	-1.293	0.81	-1.5966	0.1107	-2.883	0.297	
Chemotherapy	-0.092	0.6996	-0.132	0.895	-1.466	1.281	
Diabetes	0.150	0.7234	0.2077	0.8355	-1.270	1.570	
Formalin	0.072	0.7936	0.0904	0.928	-1.486	1.630	
Heat Injury	-1.887	0.8471	-2.228	0.0262	-3.550	-0.225	*
HIV	-0.423	0.8187	-0.5171	0.6053	-2.030	1.184	

Inflammation	-0.806	0.6755	-1.1938	0.2329	-2.132	0.520	
Migraine	-0.632	0.8297	-0.7615	0.4466	-2.260	0.997	
Nerve Injury	-0.462	0.6765	-0.6824	0.4952	-1.790	0.866	
Post-operative	-0.273	0.7316	-0.3728	0.7094	-1.709	1.163	
Visceral Inflammation	0.359	0.7902	0.4548	0.6493	-1.192	1.911	
Complex	-0.198	0.7847	-0.2529	0.8004	-1.739	1.342	
Formalin	0.288	0.4931	0.5831	0.56	-0.680	1.256	
Mechanical	0.759	0.2704	2.8086	0.0051	0.229	1.290	**
Non-evoked	-0.510	0.3971	-1.2846	0.1993	-1.290	0.269	
Thermal	1.045	0.2885	3.6217	0.0003	0.479	1.611	***
Holtzman albino	-0.801	0.9455	-0.8471	0.3972	-2.657	1.055	
Lewis	4.181	0.924	4.5249	<.0001	2.368	5.995	***
Lister-hooded	1.992	1.4044	1.4181	0.1565	-0.765	4.748	
Not reported	-1.173	1.1882	-0.9871	0.3239	-3.505	1.159	
Obese Diabetic	1.133	0.8142	1.3919	0.1643	-0.465	2.732	
Random hooded	3.130	0.728	4.2991	<.0001	1.701	4.559	***
Sprague Dawley	1.465	0.5245	2.7926	0.0054	0.435	2.495	**
Wistar	1.485	0.5244	2.8323	0.0047	0.456	2.515	**
Wistar Albino	1.892	1.1392	1.6611	0.0971	-0.344	4.129	
Female	-0.915	0.5215	-1.7552	0.0796	-1.939	0.108	
Male	-0.976	0.4457	-2.1894	0.0288	-1.851	-0.101	*

Despite the model having a good fit (determined by the Akaike information criterion; a mathematical method for evaluating how well a model fits the data it was generated from (Bozdogan, 1987)) it was necessary to explore the relationship between the variables; the regression coefficient can be inflated because of multicollinearity in the model. The variance inflation factor (VIF) was used to detect multi-collinearity. It measured the correlation and strength of correlation between the predictor variables in the regression model (Table 5.11). The general rules for the interpretation of VIF can be debated but a value below 1 indicates there is no correlation between predictor variables in the model. A value between 1 and 5 indicates moderate correlation and a value greater that 5 indicates severe correlation between a predictor and other predictor variables in the model. The VIF values indicate moderate and severe collinearity across all predictors and therefore the collinearity is cause for concern as this will undermine the reliability of the regression model results.

Table 5.11. Multicollinearity in the regression model. Rows highlighted amber represent VIF values between 1 and 5 and indicate moderate correlation. Rows highlighted red represent VIF values above 5 and indicate severe correlation. The redundant predictors were removed from the model.

Predictor	Predictor Levels	VIF value
Drug	CB_1 receptor agonist	4.7267
	CB_1 receptor inverse agonist	1.835
	CB_2 receptor agonist	4.5711
	CB_2 receptor inverse agonist	1.0404
	CBD	1.9726
	FAAH inhibitor	3.6665
	FABP inhibitor	1.2462
	GPR55 agonist	1.039
	Hemp oil	1.0351
	Monoacylglycerol lipase inhibitor	1.516
	NAAA inhibitor	1.233
	Non-selective cannabinoid receptor agonist	3.6283
	PPAR-alpha agonist	3.1296
	PPAR-gamma antagonist	1.4686
	THC	2.779
	Unknown mechanism of action	1.2573
Model	Burn Injury	2.0825
	Cancer	3.6625

	Capsaicin	3.4681
	Chemotherapy	13.4725
	Diabetes	11.0955
	Formalin	27.6486
	Heat Injury	3.2142
	HIV	3.1847
	Inflammation	39.1076
	Migraine	4.8608
	Nerve Injury	39.6344
	Post-operative	7.1149
	Visceral Inflammation	4.6979
Outcome Assessment	Complex	1.2382
	Formalin	11.8894
	Mechanical	7.384
	Non-evoked	4.0243
	Thermal	6.1013
Strain	Holtzman albino	2.7811
	Lewis	1.665
	Lister-hooded	1.3646
	Not reported	1.5487
	Obese Diabetic	1.5675
	Random hooded	2.8795
	Sprague Dawley	27.3356
	Wistar	26.1987
	Wistar Albino	1.2515
Sex	Female	6.3509
	Male	6.2858

Rather than rely on one model, I next used multi-model inference. In this technique all possible predictor combinations are modelled. The multi-model inference meta-regression yielded similar results to the forced entry model (Table 5.12), and in addition provided the averaged importance of each predictor across all models showing that all 5 variables are of importance (drug class, model type, outcome assessment, strain, and sex) (Figure 5.7). Although, multicollinearity between predictors will also threaten the multi-model inference meta-regression.

Table 5.12. The rat coefficient estimates from the multi-model inference meta-regression of the variables drug class, model type, outcome assessment, animal strain and sex. Significance levels; *** P< 0.001, ** P< 0.01, * P< 0.05.

Predictor Level	Predictor	Coefficient Estimate	SE	z value	Pr(>|z|)	Significance Level
Drug	CB_1 receptor agonist	-3.20E-02	0.266	0.120	0.904	
	CB_1 receptor inverse agonist	-3.26E-02	0.548	0.060	0.953	
	CB_2 receptor agonist	3.59E-01	0.280	1.281	0.200	
	CB_2 receptor inverse agonist	1.94E-01	1.385	0.140	0.888	
	CBD	1.58E-02	0.342	0.046	0.963	
	FAAH inhibitor	1.84E-01	0.287	0.639	0.523	
	FABP inhibitor	9.04E-01	1.357	0.666	0.505	
	GPR55 agonist	-1.56E+00	1.313	1.184	0.236	
	Hemp oil	-1.38E+00	1.367	1.008	0.313	
	Monoacylglycerol lipase inhibitor	6.15E-01	0.473	1.301	0.193	
	NAAA inhibitor	5.33E-01	0.604	0.882	0.378	
	Non-selective cannabinoid receptor agonist	2.44E-01	0.267	0.916	0.360	
	PPAR-alpha agonist	4.22E-01	0.295	1.430	0.153	
	PPAR-gamma antagonist	-3.94E-01	1.045	0.377	0.706	
	THC	6.86E-01	0.385	1.781	0.075	
	Unknown mechanism of action	2.82E-01	0.876	0.322	0.747	
Model	Burn Injury	2.69E-01	1.216	0.222	0.825	
	Cancer	-5.08E-01	0.872	0.583	0.560	
	Capsaicin	-1.28E+00	0.839	1.524	0.128	
	Chemotherapy	-5.69E-02	0.727	0.078	0.938	
	Diabetes	1.97E-01	0.750	0.263	0.792	
	Formalin	1.19E-01	0.820	0.145	0.885	
	Heat Injury	-1.86E+00	0.879	2.111	0.035	*
	HIV	-4.03E-01	0.848	0.476	0.634	

	Inflammation	-7.77E-01	0.703	1.105	0.269	
	Migraine	-5.97E-01	0.858	0.696	0.486	
	Nerve Injury	-4.22E-01	0.704	0.599	0.549	
	Post-operative	-2.29E-01	0.758	0.303	0.762	
	Visceral Inflammation	3.75E-01	0.818	0.458	0.647	
Outcome Assessment	Complex	-1.93E-01	1.066	0.490	0.811	
	Formalin	3.04E-01	0.808	0.239	0.545	
	Mechanical	7.80E-01	0.277	2.812	0.005	**
	Non-evoked	-5.04E-01	0.407	1.238	0.216	
	Thermal	1.07E+00	0.295	3.629	0.000	***
Strain	Lewis	4.25E+00	0.934	4.552	0.000	***
	Lister-hooded	2.01E+00	1.445	1.392	0.164	
	Not reported	-1.27E+00	1.222	1.035	0.301	
	Obese Diabetic	1.16E+00	0.834	1.386	0.166	
	Random hooded	3.16E+00	0.742	4.263	0.000	***
	Sprague Dawley	1.48E+00	0.536	2.767	0.006	***
	Wistar	1.51E+00	0.536	2.810	0.005	***
	Wistar Albino	1.90E+00	1.146	1.654	0.098	
	Hotzman albino	2.47E-03	0.114	0.022	0.983	
	Wistar-Kyoto	-2.42E-05	0.009	0.003	0.998	
	Wistar Han	-1.16E-05	0.006	0.002	0.999	
Sex	Female	-1.12E-01	0.868	0.129	0.897	
	Male	-1.65E-01	0.854	0.194	0.846	
	Not reported	8.31E-01	0.967	0.859	0.390	

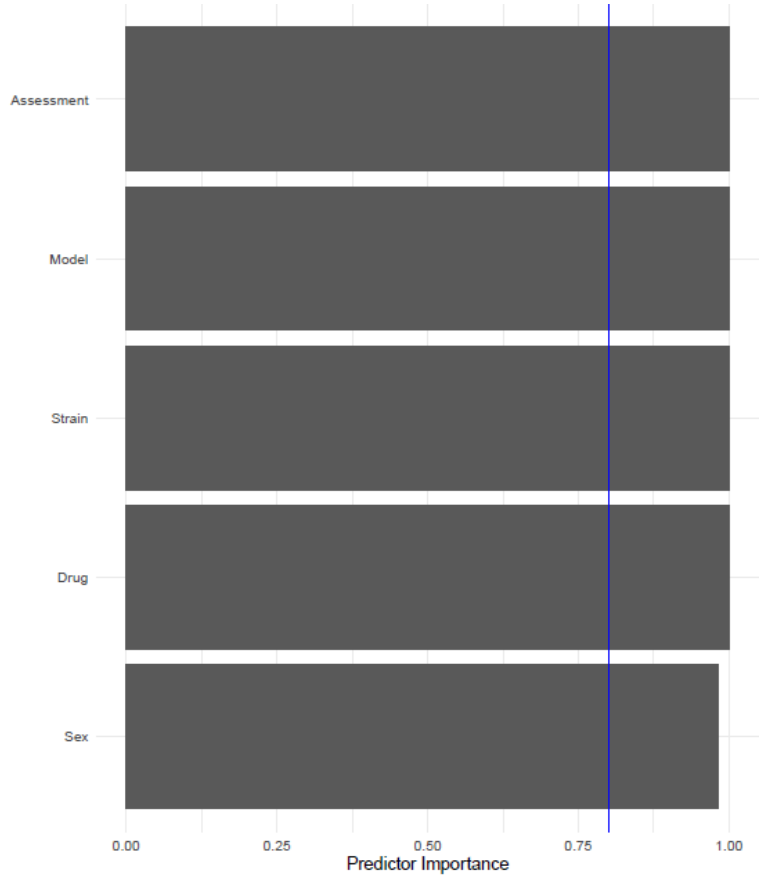


Figure 5.7. Model-averaged predictor importance plot (without consideration for interactions between predictors) for rats in which cannabinoids, cannabis-based medicines and endocannabinoid system modulators assessed for antinociceptive efficacy in animal models of injury-related of pathological persistent pain. Outcome assessment type; 1.00; Model type; 1.00; Strain; 1.00; Drug class; 0.99; Strain; 0.98.

In mice, the forced entry meta-regression model (mixed effects with maximum likelihood estimator), the variables sex, strain, model type, drug class and outcome assessment type accounted for (R^2) 56 % of heterogeneity (I^2 (unaccounted variability) 48 % and τ^2 = 0.821). For each predictor, the coefficients estimates are listed in Table 5.13. It explains which variables are likely associated with higher effect sizes (positive coefficient estimates) and those associated with lower effect sizes (negative coefficient estimates).

Table 5.13. The mouse coefficient estimates from multivariable meta-regression of the variables drug class, model type, outcome assessment, animal strain and sex. The redundant predictors were removed from the model.

Predictor levels	Coefficient Estimate	SE	t-value	p-value	95 % CI Lower	95 % CI Upper	Significance Level
Anandamide transport inhibitor	-0.2024	0.9321	-0.2172	0.8281	-2.0329	1.628	
CB1 receptor agonist	0.8976	0.8374	1.0719	0.2842	-0.747	2.5423	
CB1 receptor inverse agonist	3.7882	1.189	3.1861	0.0015	1.4532	6.1233	**
CB1 receptor PAM	0.3388	1.0553	0.3211	0.7483	-1.7336	2.4112	
CB2 receptor agonist	0.2559	0.8208	0.3118	0.7553	-1.356	1.8679	
CB2 receptor inverse agonist	-0.3646	1.2038	-0.3028	0.7621	-2.7288	1.9996	
CBD	0.69	0.9053	0.7622	0.4462	-1.0878	2.4679	
Diacylglycerol lipase inhibitor	0.8653	0.924	0.9365	0.3494	-0.9492	2.6799	
Dual FAAH/MAGL inhibitor	1.0769	0.9894	1.0884	0.2768	-0.8662	3.0199	
FAAH inhibitor	0.7199	0.8299	0.8675	0.386	-0.9099	2.3498	
FAAH inhibitor/TRPA1 agonist	1.2705	1.4503	0.876	0.3814	-1.5778	4.1187	
FAAH inhibitor/TRPV1 agonist	-1.058	1.1427	-0.9259	0.3549	-3.3021	1.1861	
FABP inhibitor	-0.2048	0.8896	-0.2302	0.818	-1.9518	1.5422	
Monoacylglycerol lipase inhibitor	0.7876	0.8481	0.9287	0.3534	-0.8779	2.4532	
NAAA inhibitor	2.1109	1.0233	2.0627	0.0396	0.1012	4.1205	*
Non-selective cannabinoid receptor agonist	0.285	0.8351	0.3412	0.733	-1.355	1.9249	
PPAR-alpha agonist	2.28	0.87	2.6207	0.009	0.5714	3.9886	**
THC	0.8589	0.8853	0.9701	0.3324	-0.8798	2.5975	
Unknown mechanism of action	0.9599	0.9794	0.9801	0.3274	-0.9635	2.8832	
Chemical cauterisation	-1.2241	0.8417	-1.4543	0.1464	-2.8771	0.4289	
Chemotherapy	-0.0962	0.4811	-0.2	0.8415	-1.041	0.8485	
Diabetes	0.5893	0.5631	1.0466	0.2957	-0.5165	1.6951	
Formalin	-0.6766	0.5681	-1.191	0.2341	-1.7922	0.439	
HIV	7.1354	1.792	3.9817	<.0001	3.616	10.6547	***

Inflammation	-0.5691	0.4231	-1.3452	0.1791	-1.3999	0.2617	
Mild traumatic brain injury	-1.9847	1.449	-1.3697	0.1713	-4.8304	0.861	
Multiple sclerosis	-0.2096	0.7991	-0.2623	0.7932	-1.7789	1.3597	
Musckuloskeletal	0.8063	0.9107	0.8853	0.3763	-0.9822	2.5948	
Mustard oil	-2.4361	1.2508	-1.9477	0.0519	-4.8924	0.0203	
Nerve injury	-0.9514	0.4287	-2.2193	0.0268	-1.7933	-0.1095	*
Post-operative	0.9846	0.5799	1.6977	0.0901	-0.1544	2.1235	
Sickle cell disease	-2.0145	1.1679	-1.7249	0.085	-4.3081	0.2791	
Visceral inflammation	-1.9348	0.706	-2.7404	0.0063	-3.3213	-0.5483	**
Formalin	0.2003	0.5832	0.3434	0.7314	-0.945	1.3455	
Mechanical	-0.4777	0.3543	-1.348	0.1782	-1.1735	0.2182	
Non evoked	-0.9995	0.5682	-1.7589	0.0791	-2.1154	0.1164	
Thermal	-0.245	0.3998	-0.6129	0.5401	-1.0301	0.5401	
BALkB/cfC3H	1.0495	0.9442	1.1115	0.2668	-0.8048	2.9037	
C3H/He	-1.1088	0.8033	-1.3804	0.168	-2.6863	0.4687	
C3H/HeJ	1.8475	1.0145	1.8211	0.0691	-0.1448	3.8398	
C3H/HeN	-3.1225	0.6962	-4.4851	<.0001	-4.4897	-1.7553	***
C3H/HeNCr	-1.3781	0.7567	-1.8213	0.0691	-2.8641	0.1079	
C3H/HeNCr MTV-	-1.8323	1.1149	-1.6435	0.1008	-4.0218	0.3571	
C57BL/6	-0.1053	0.5698	-0.1848	0.8534	-1.2244	1.0138	
C57BL/6J	-0.4206	0.5577	-0.7541	0.4511	-1.5158	0.6747	
C57BL/6N	0.8287	0.7888	1.0506	0.2938	-0.7204	2.3778	
C57BL/6NH	2.9116	1.5317	1.9008	0.0578	-0.0965	5.9198	
CD-1	-0.5665	0.5584	-1.0146	0.3107	-1.663	0.53	
DBA/1J	-0.4335	1.1614	-0.3733	0.7091	-2.7143	1.8473	
DBAk1/J	0.6152	1.6341	0.3765	0.7067	-2.594	3.8245	
Foxn1nu, athymic mice	1.0927	1.3135	0.8319	0.4058	-1.4868	3.6722	
Foxn1nu, athymic, immunocompromised	3.3186	1.3384	2.4795	0.0134	0.6901	5.947	*
ICR	-1.225	0.7533	-1.6262	0.1044	-2.7044	0.2543	
Kumming	-1.2615	0.9287	-1.3584	0.1749	-3.0853	0.5623	

ND4 Swiss	-1.0066	0.9936	-1.0131	0.3114	-2.9578	0.9446	
NMRI	-2.0473	0.8471	-2.417	0.0159	-3.7108	-0.3838	*
Not reported	-0.6083	0.9059	-0.6715	0.5021	-2.3873	1.1707	
OF1	-0.1413	1.1472	-0.1231	0.902	-2.3941	2.1116	
Swiss	-1.3646	0.5688	-2.399	0.0167	-2.4816	-0.2475	*
Swiss albino	-1.6795	1.015	-1.6546	0.0985	-3.6729	0.3139	
Swiss Albino imprinting control region (ICR)	-0.0286	0.8123	-0.0352	0.9719	-1.6239	1.5667	
Swiss CD1	-0.5732	0.9952	-0.5759	0.5649	-2.5276	1.3813	
Swiss Webster	-1.1128	0.6861	-1.622	0.1053	-2.4602	0.2346	
Transgenic (alpha and beta globin)	-6.3632	1.693	-3.7586	0.0002	-9.688	-3.0384	***
Female	-2.6858	0.5674	-4.7337	<.0001	-3.8001	-1.5716	***
Male	-1.3422	0.481	-2.7906	0.0054	-2.2868	-0.3976	**
Not reported	-0.9543	0.5966	-1.5996	0.1102	-2.126	0.2173	

As for rats, it was necessary to explore multicollinearity. Similarly, most predictors have a moderate or severe correlation (Table 5.14).

Table 5.14. Multicollinearity in the mouse multivariable regression model. Rows highlighted amber represent VIF values between 1 and 5 and indicate moderate correlation. Rows highlighted red represent VIF values above 5 and indicate severe correlation. The redundant predictors were removed from the model.

Predictor Group	Predictor	VIF
Drug	Anandamide transport inhibitor	5.4645
	CB_1 receptor agonist	19.3312
	CB_1 receptor inverse agonist	3.0214
	CB_1 receptor PAM	2.7069
	CB_2 receptor agonist	28.1063
	CB_2 receptor inverse agonist	2.6697
	CBD	6.9636
	Diacylglycerol lipase inhibitor	4.3012
	Dual FAAH/MAGL inhibitor	3.2273
	FAAH inhibitor	20.2268
	FAAH inhibitor/TRPA1 agonist	1.537
	FAAH inhibitor/TRPV1 agonist	2.1
	FABP inhibitor	8.0904
	Monoacylglycerol lipase inhibitor	10.9569
	NAAA inhibitor	3.0766
	Non-selective cannabinoid receptor agonist	18.3926
	PPAR-alpha agonist	10.0715
	THC	8.548
	Unknown mechanism of action	6.1847
Model	Chemical cauterisation	3.7659
	Chemotherapy	4.0659
	Diabetes	1.9958
	Formalin	12.473
	HIV	1.1459
	Inflammation	8.9551
	Mild traumatic brain injury	1.17
	Multiple sclerosis	1.8053

	Musculoskeletal	1.2911
	Mustard oil	1.7003
	Nerve injury	8.0673
	Post-operative	2.3496
	Sickle cell disease	1.5357
	Visceral inflammation	2.5953
Outcome assessment	Formalin	12.6058
	Mechanical	7.703
	Non evoked	5.4487
	Thermal	5.9086
Strain	BALkB/cfC3H	1.8053
	C3H/He	1.9
	C3H/HeJ	1.4909
	C3H/HeN	2.4661
	C3H/HeNCr	2.9565
	C3H/HeNCr MTV-	1.3901
	C57BL/6	12.029
	C57BL/6J	14.9871
	C57kBL/6N	2.0707
	C57BL/6NH	1.3891
	CD-1	8.8112
	DBA/1J	1.4247
	DBAk1/J	1.143
	Foxn1nu, athymic mice	1.2281
	Foxn1nu, athymic, immunocompromised	1.2125
	ICR	3.1508
	Kumming	1.5993
	ND4 Swiss	1.5537
	NMRI	1.9158
	Not reported	4.8672
	OF1	1.404
	Swiss	8.3571
	Swiss albino	1.7639

	Swiss Albino imprinting control region (ICR)	1.934
	Swiss CD1	2.5416
	Swiss Webster	3.4282
	Transgenic (alpha and beta globin	1.2664
Sex	Female	6.7914
	Male	8.6325
	Not reported	5.4107

The multi-model inference meta-regression yielded similar results to the forced entry model (Table 5.15), and in addition provided the averaged importance of each predictor across all models showing that all 5 variables are of importance (drug class, model type, outcome assessment, strain, and sex) (Figure 5.8). Although, multicollinearity between predictors will also threaten the multi-model inference meta-regression.

Table 5.15. The mouse coefficient estimates from the multi-model inference meta-regression of the variables drug class, model type, outcome assessment, animal strain and sex. Significance levels; *** P< 0.001, ** P< 0.01, * P< 0.05.

Predictor	Predictor levels	Coefficient Estimate	SE	z value	Pr(>|z|)	Significance Level
Drug	Anandamide transport inhibitor	-2.54E-01	9.75E-01	2.60E-01	0.795	
	CB1 receptor agonist	9.06E-01	8.79E-01	1.03E+00	0.303	
	CB1 receptor inverse agonist	3.84E+00	1.23E+00	3.12E+00	0.002	**
	CB1 receptor PAM	3.39E-01	1.11E+00	3.07E-01	0.759	
	CB2 receptor agonist	2.92E-01	8.63E-01	3.38E-01	0.735	
	CB2 receptor inverse agonist	-3.70E-01	1.27E+00	2.91E-01	0.771	
	CBD	7.11E-01	9.48E-01	7.50E-01	0.453	
	Diacylglycerol lipase inhibitor	9.23E-01	9.66E-01	9.55E-01	0.339	
	Dual FAAH/MAGL inhibitor	1.09E+00	1.03E+00	1.06E+00	0.290	
	FAAH inhibitor	7.50E-01	8.71E-01	8.61E-01	0.389	
	FAAH inhibitor/TRPA1 agonist	1.38E+00	1.51E+00	9.16E-01	0.360	
	FAAH inhibitor/TRPV1 agonist	-1.13E+00	1.18E+00	9.51E-01	0.342	
	FABP inhibitor	-1.99E-01	9.31E-01	2.13E-01	0.831	
	Monoacylglycerol lipase inhibitor	8.20E-01	8.89E-01	9.22E-01	0.357	
	NAAA inhibitor	2.19E+00	1.06E+00	2.07E+00	0.038	*
	Non-selective cannabinoid receptor agonist	2.93E-01	8.76E-01	3.34E-01	0.738	
	PPAR-alpha agonist	2.36E+00	9.14E-01	2.58E+00	0.010	**
	THC	8.78E-01	9.27E-01	9.47E-01	0.343	
	Unknown mechanism of action	9.03E-01	1.02E+00	8.87E-01	0.375	
Model	Chemical cauterisation	-1.34E+00	8.71E-01	1.54E+00	0.124	
	Chemotherapy	-1.04E-01	4.95E-01	2.10E-01	0.833	
	Diabetes	5.41E-01	5.80E-01	9.33E-01	0.351	
	Formalin	-6.52E-01	5.91E-01	1.10E+00	0.270	
	HIV	7.03E+00	1.75E+00	4.01E+00	0.000	***
	Inflammation	-5.78E-01	4.37E-01	1.32E+00	0.186	

	Mild traumatic brain injury	-2.08E+00	1.53E+00	1.36E+00	0.172	
	Multiple sclerosis	-2.74E-01	8.24E-01	3.33E-01	0.739	
	Musckuloskeletal	7.70E-01	9.39E-01	8.20E-01	0.412	
	Mustard oil	-2.48E+00	1.28E+00	1.93E+00	0.054	
	Nerve injury	-9.66E-01	4.42E-01	2.18E+00	0.029	*
	Post-operative	1.05E+00	5.98E-01	1.75E+00	0.080	
	Sickle cell disease	-2.16E+00	1.22E+00	1.76E+00	0.078	
	Visceral inflammation	-1.96E+00	7.32E-01	2.68E+00	0.007	**
Outcome assessment	Formalin	2.01E-01	6.02E-01	3.33E-01	0.739	
	Mechanical	-4.68E-01	3.68E-01	1.27E+00	0.203	
	Non evoked	-1.02E+00	6.01E-01	1.70E+00	0.090	
	Thermal	-2.12E-01	4.10E-01	5.18E-01	0.604	
Strain	BALB/cfC3H	9.50E-01	1.00E+00	9.46E-01	0.344	
	C3H/He	-1.19E+00	8.32E-01	1.43E+00	0.153	
	C3H/HeJ	-1.19E+00	8.32E-01	1.43E+00	0.153	
	C3H/HeN	1.76E+00	1.04E+00	1.69E+00	0.091	
	C3H/HeNCr	-3.29E+00	7.23E-01	4.55E+00	0.000	***
	C3H/HeNCr MTV-	-1.51E+00	7.79E-01	1.94E+00	0.052	
	C57BL/6	-1.96E+00	1.12E+00	1.76E+00	0.079	
	C57BL/6J	-2.12E-01	5.86E-01	3.62E-01	0.717	
	C57BL/6N	-5.30E-01	5.74E-01	9.23E-01	0.356	
	C57BL/6NH	7.03E-01	8.07E-01	8.71E-01	0.384	
	CD-1	-6.66E-01	5.75E-01	1.16E+00	0.247	
	DBA/1J	-5.20E-01	1.22E+00	4.27E-01	0.669	
	DBAk1/J	4.47E-01	1.68E+00	2.65E-01	0.791	
	Foxn1nu, athymic mice	1.00E+00	1.34E+00	7.47E-01	0.455	
	Foxn1nu, athymic, immunocompromised	3.38E+00	1.34E+00	2.52E+00	0.012	*
	ICR	-1.36E+00	7.78E-01	1.75E+00	0.081	
	Kumming	-1.29E+00	9.62E-01	1.34E+00	0.179	
	ND4 Swiss	-1.15E+00	1.04E+00	1.11E+00	0.268	

	NMRI	-2.22E+00	8.70E-01	2.56E+00	0.011	*
	Not reported	-7.77E-01	9.34E-01	8.32E-01	0.406	
	OF1	-3.10E-01	1.20E+00	2.59E-01	0.796	
	Swiss	-1.51E+00	5.86E-01	2.58E+00	0.010	*
	Swiss albino	-1.82E+00	1.04E+00	1.75E+00	0.081	
	Swiss Albino imprinting control region (ICR)	-1.36E-01	8.25E-01	1.64E-01	0.869	
	Swiss CD1	-6.14E-01	1.02E+00	6.01E-01	0.548	
	Swiss Webster	-1.23E+00	7.05E-01	1.75E+00	0.081	
	Transgenic (alpha and beta globin)	-6.63E+00	1.70E+00	3.91E+00	0.000	***
Sex	Female	-2.84E+00	5.77E-01	4.92E+00	0.000	***
	Male	-1.45E+00	4.87E-01	2.98E+00	0.003	**
	Not reported	-1.03E+00	6.07E-01	1.70E+00	0.090	

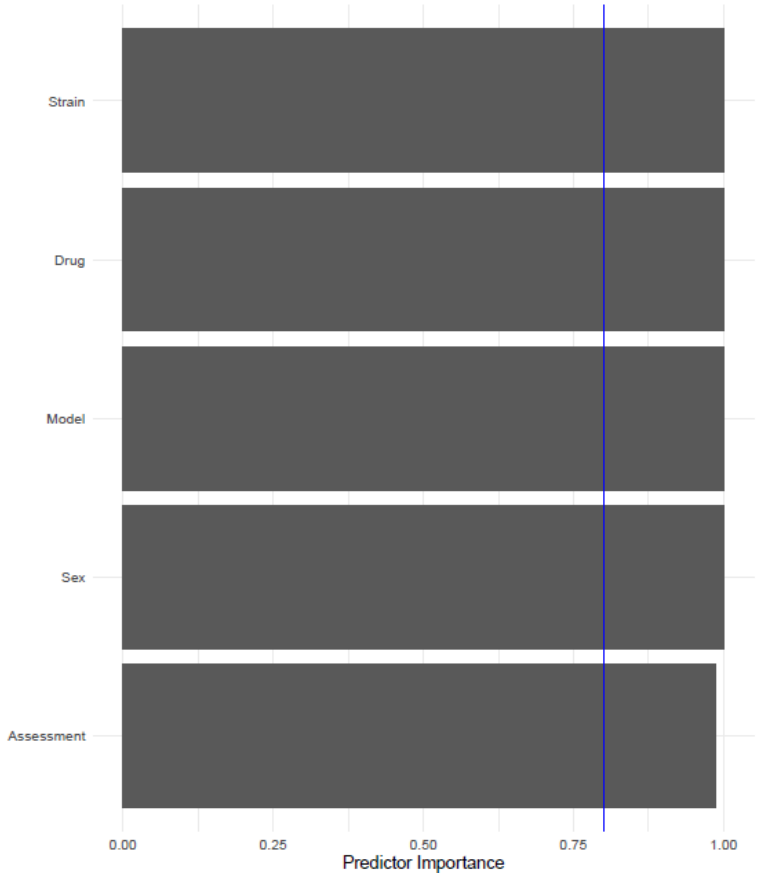


Figure 5.8. Model-averaged predictor importance plot (without consideration for interactions between predictors) for mice in which cannabinoids, cannabis-based medicines and endocannabinoid system modulators assessed for antinociceptive efficacy in animal models of injury-related of pathological persistent pain. Drug Class; 1.00; Strain; 1.00; Model type; 1.00; Sex; 1.00; Outcome assessment type; 0.99.

5.3.8. Impact of Methodological Quality

Reporting of blinded assessment of outcome, pre-determined animal inclusion criteria and animal exclusions did not account for a significant proportion of heterogeneity. Allocation concealment, randomisation and sample size calculation accounted for a significant proportion of heterogeneity. In the case of sample size calculation and allocation concealment the low prevalence of reporting may limit our ability to accurately determine their influence on the reported outcomes. However, larger effect sizes were reported in the studies that did not report allocation concealment and sample size calculations, SMD=1.345 vs SMD=1.055 ($Q=5.299$, d.f.1, $p=0.021$) and SMD=0.221 vs 1.349 ($Q=18.104$, d.f.1, $p<0.0001$) respectively. It was the converse for randomisation, where larger effect sizes were reported in studies that did report randomisation (SMD=1.471 vs SMD=1.245, $Q=5.792$, d.f.1, $p=0.016$)(Figure 5.9).

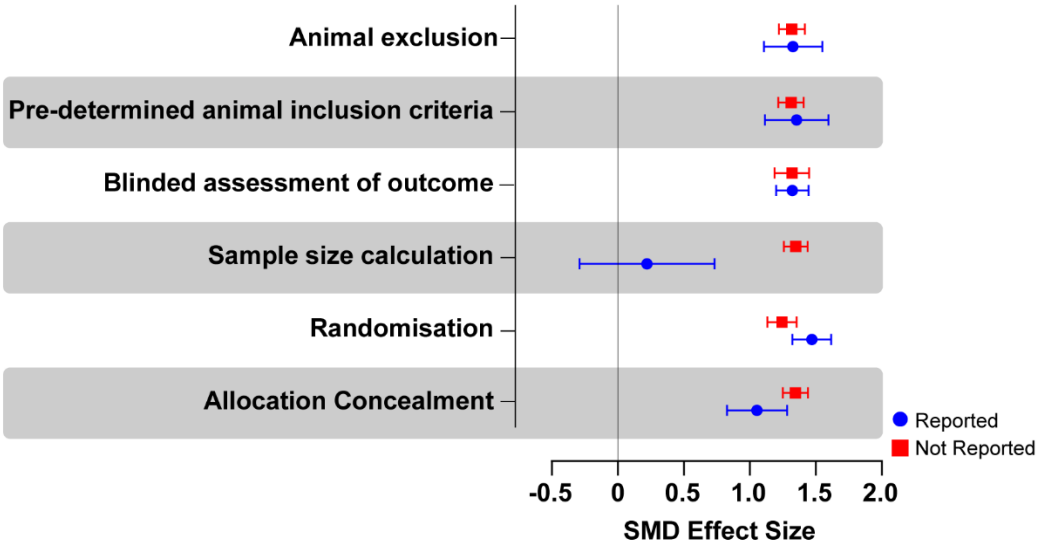


Figure 5.9. Effect sizes (and variance) associated with the reporting of methodological quality criteria.

5.3.9. Prevalence of Publication Bias at the Study Level

5.3.9.1. Initial (incorrect) and revised analysis of publication bias

Not following the most up to date methodology of Zwetsloot et al. (2017) a funnel plot was generated by plotting each study's effect size on the x axis against its standard error on the y axis. Analysis of the data from the 374 included studies have an overall effect size (SMD) of 1.321 [95 % CI 1.232 – 1.411]. Visual inspection of the funnel plot shows that normal distribution is off set to the right-hand side. This asymmetry might suggest that the analysis has been affected by publication bias and studies are missing. The coloured contours show that many studies do not report significant effects (plotted in the central white triangle) (Figure 5.10A). Trim and fill analysis imputed 93 theoretically missing studies to the left-hand side of the funnel plot (Figure 5.10B). Inclusion of the theoretically missing studies decreased the overall estimate of intervention effects by 38 % to a SMD=0.817 [95 % CI 0.644 – 0.991]. Egger's regression was consistent with effects of small studies ($p < 0.005$) suggesting there is substantial asymmetry which may be attributed to publication bias, within-study bias in small studies, or to other unknown factors.

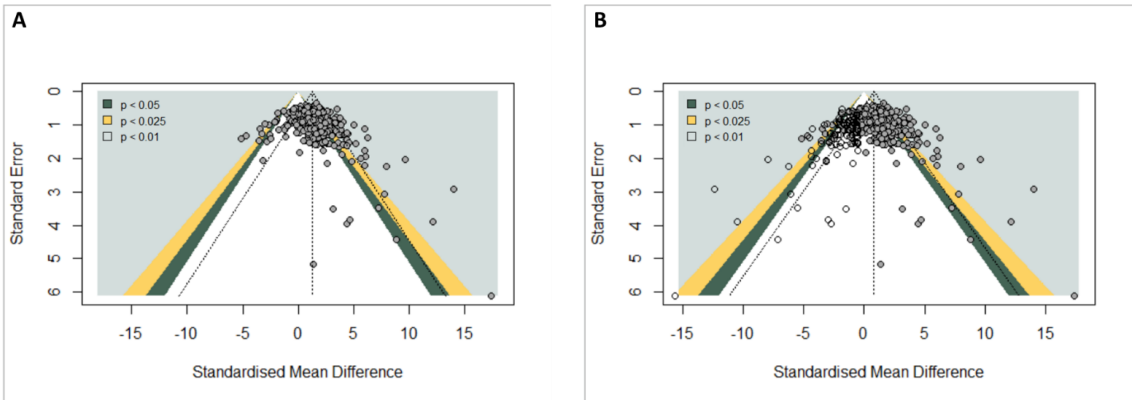


Figure 5.10. (Incorrect) Assessment of publication bias. (A) visual inspection of the funnel plot suggests asymmetry. Filled circles denote the published studies and the vertical dotted line the overall effect size. (B) Trim and fill analysis imputed 93 missing studies (open circles). The contours denote the significance of individual study effects.

The Zwetsloot et al. (2017) suggested that this method can lead to funnel plot distortion and is associated with an over-estimation of asymmetry which artificially lowers the threshold for publication bias. This distortion is presented in Fig 5.11A and shows that the above method should not be used for this type of data (small sample sizes that characterise preclinical studies).

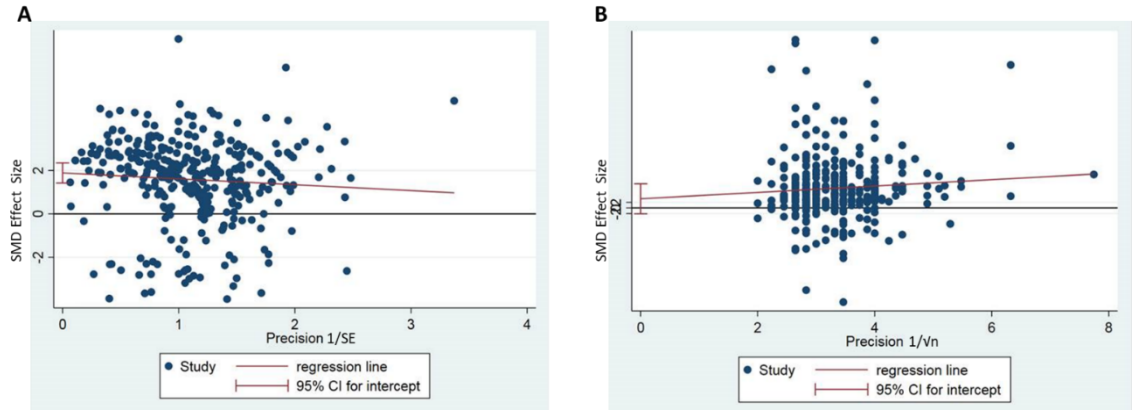


Figure 5.11. SMD vs SE funnel plots vs sample-size based precision plots (A) illustration of the distortion of SMD vs SE funnel plot. (B) plotting precision as 1/√n.

Analysis of the data from the 374 included studies has an overall effect size (SMD) of 1.321 (95% CI 1.232-1.411). Visual inspection of the funnel plot shows normal distribution. Egger's regression was not consistent with effects of small studies (p = 0.112) and did not indicate the presence of funnel

plot asymmetry. Trim and fill analysis did not impute any theoretically missing studies (Fig. 5.12).

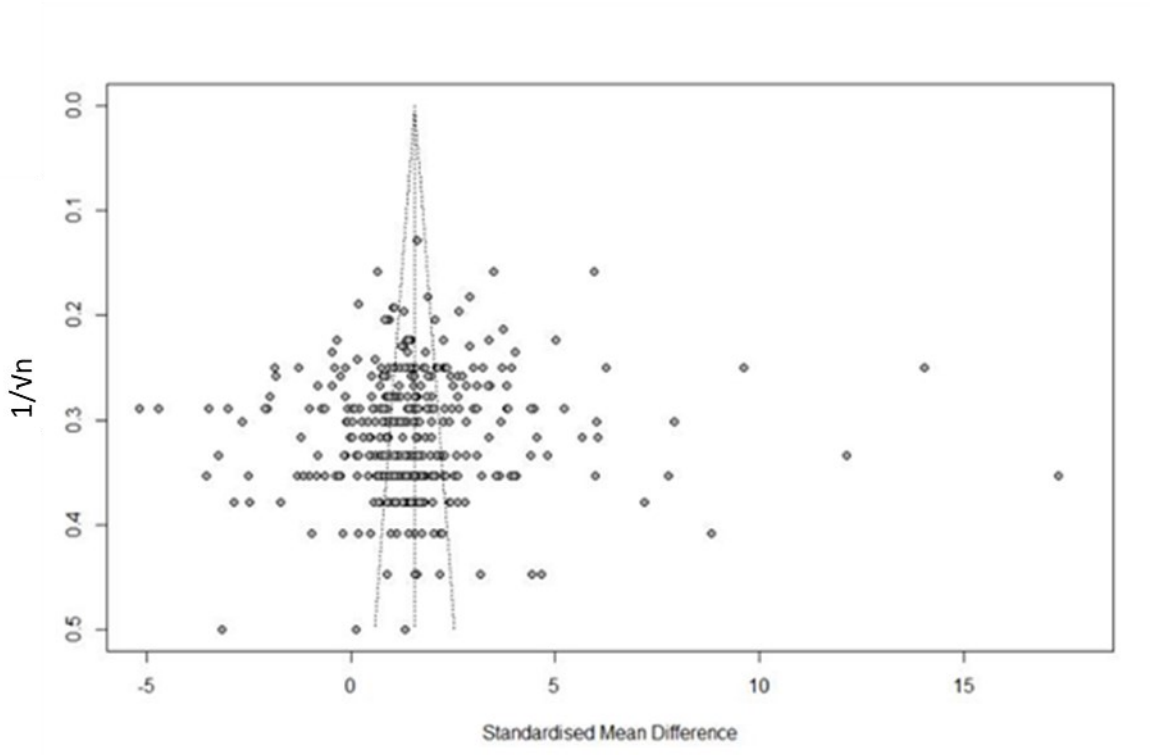


Figure 5.12. Assessment of publication bias (study level). Visual inspection of the funnel plot does not suggest asymmetry.

Publication bias assessment was performed based upon individual outcome measures. The visual inspection of the funnel plot does not show normal distribution (Fig 5.13A). Egger's regression indicated the presence of funnel plot asymmetry (p >0.001). Trim and fill analysis imputed 910 theoretically missing outcomes however this was on the right-hand side and not on the left-hand side which would be expected for publication bias (Fig.5.13B).

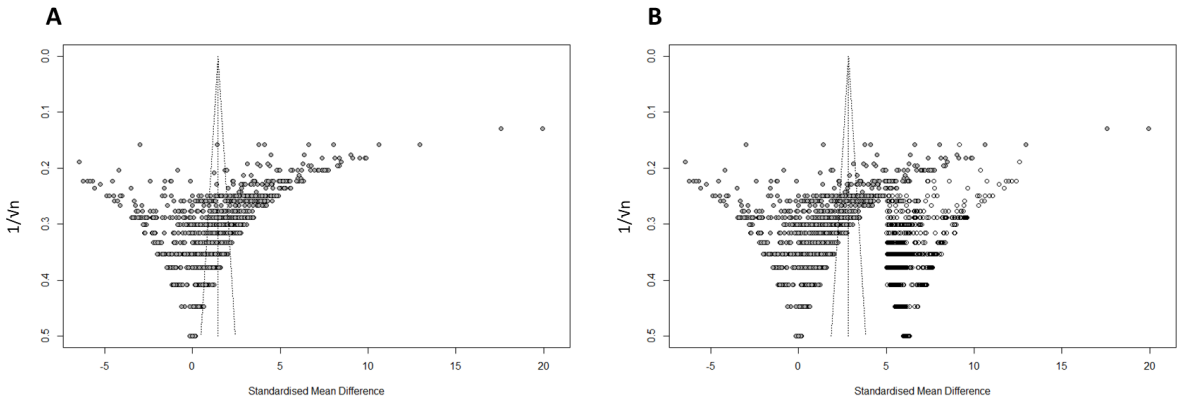


Figure 5.13. Assessment of publication bias based upon individual outcome comparisons. Visual inspection of the funnel plot suggests asymmetry. Filled circles denote the published studies and the vertical dotted line the overall effect size. (B) Trim and fill analysis imputed 910 theoretically missing outcomes (open circles).

5.4. Discussion

We reported a systematic review of preclinical studies in which cannabinoids, cannabis-based medicines and endocannabinoid system modulators were assessed for behavioural signs of antinociceptive efficacy in animal models of injury-related or pathological persistent pain (Soliman et al., 2021).

We identified 374 studies including over 16,000 rodents in which 171 different cannabis-based medicines, cannabinoids or endocannabinoid system modulators were assessed for antinociceptive efficacy in 20 different animal model types of injury-related or pathological persistent pain. All the included studies investigated effects in rodents only, suggesting a scarcity of studies investigating antinociceptive effects in larger animals e.g., canines and primates.

Most experiments were conducted in male Sprague Dawley rats, with inflammation and nerve injury most frequently modelled. Antinociceptive efficacy was measured predominantly by attenuation of hypersensitivity in evoked limb withdrawal assessments. The interventions led to a statistically significant attenuation of pain-associated behaviour compared to control, the overall SMD was 1.321 [95 % CI 1.232 – 1.411]. However, this groups together a very broad range of drugs, drug classes (some with opposing mechanisms of action), models and pain-associated behavioural outcome measures.

Thus, more useful insight of the antinociceptive efficacy of these drugs has been gained from the subgroup analyses.

5.4.1. Antinociceptive Efficacy of Cannabinoids, Cannabis-based Medicines, and Endocannabinoid System Modulators

Selective CB_1, CB_2, non-selective cannabinoid receptor agonists (including delta-9-tetrahydrocannabinol; THC), and PPAR-alpha agonists (predominantly palmitoylethanolamide; PEA) significantly attenuated pain-associated behaviours in a broad range of inflammatory and neuropathic pain models. Fatty acid amide hydrolase (FAAH) inhibitors, monoacylglycerol lipase (MGL) inhibitors and cannabidiol (CBD) significantly attenuated pain-associated behaviours in neuropathic pain models but yielded mixed results in inflammatory pain models. The differences of antinociceptive efficacy may be inherent to the interventions but are also likely to be influenced by other study design characteristics. Careful consideration should be given to the choice of species, strain, and sex in relation to the clinical condition being modelled, coupled with the need to assess efficacy using multiple species, strains, models, and pain-associated behavioural outcomes; the conclusions will be dependent on these variables (Mogil, 2017). Increasing the biological variation will improve the generalisability of the results.

There are many different strains of the cannabis plant each containing different amounts of phytocannabinoids (Berman et al., 2018). Of note, is the psychoactive component, THC, whose pharmacological effects are attributed to activity at both the cannabinoid receptors. THC significantly attenuated pain-associated behaviours in a broad range of models however, we have not been able to analyse the broader effects of the drug. Similarly, CBD significantly attenuated pain-associated behaviours in a broad range of models. Most drugs assessed were small molecule CB_1 and CB_2 receptor agonists and non-selective agonists. The CB_1 and CB_2 receptor agonists consistently significantly attenuated pain-associated behaviours in a broad range of rat models however their antinociceptive effect was less consistent in the mouse models. Unlike in rats, in mouse nerve injury models, CB_1 and CB_2 receptor agonists did not significantly attenuate pain related behaviour. We did not assess side effects (e.g., motor impairment, hypothermia or anxiolysis) that could influence the interpretation of pain-associated behaviour, although these assessments were not commonly reported.

Endocannabinoid system modulators e.g., FAAH and MGL inhibitors, are considered promising targets for analgesic drug development, thereby preventing the harms associated with cannabis and orthosteric cannabinoid agonists (reviewed by Guindon and Hohmann (2009). The evidence for efficacy of FAAH inhibitors is mixed. In rats and mice, FAAH inhibitors significantly attenuated pain-associated behaviours in nerve injury, formalin, and diabetes models. However, they did not

significantly attenuate pain-associated behaviours of rat models of inflammation and both rat and mouse models of visceral inflammation. The FAAH inhibitors significantly attenuated pain-associated behaviours in CFA but not in osteoarthritis models. The evidence indicates that FAAH inhibitors are least effective for osteoarthritis but may be a viable candidate for treatment of neuropathic pain conditions. Concomitantly, the FAAH inhibitor, PF-04457845, failed to demonstrate analgesic efficacy in a randomised placebo-controlled clinical trial of osteoarthritis patients (Huggins et al., 2012). Our findings support the potential utility of a prospective preclinical animal systematic review and meta-analysis to review the animal efficacy data prior to clinical trial.

The drugs were grouped by class, however, the drugs themselves can activate different signalling pathways and the signalling pathways necessary for therapeutic effect are not fully understood. This is particularly pertinent to the antinociceptive potential of inverse agonists; the data do not represent all inverse agonist studies merely only studies in which they were assessed for antinociceptive efficacy. The authors of the included studies have postulated that inverse agonists may yield an antinociceptive effect due to the subsequent reduction of pro-inflammatory and pro-nociceptive mediators.

5.4.2. External Validity

5.4.2.1. Misalignment between animal models and the clinical population

The models used in preclinical pain research generally are not often well matched to the clinical population (du Sert and Rice, 2014, Rice et al., 2018). Sex only accounted for a significant proportion of heterogeneity in rats although this is likely attributed to groups that did not state the sex of the animals or used mixed sex groups. The data from unreported and mixed groups of sexes also has greater variance, similar to an observation made in a preclinical systematic review of chemotherapy induced peripheral neuropathy (Currie et al., 2019). The studies have been conducted predominantly using male animals. It is likely that the paucity of female animals limits our ability to determine the influence of sex in the reported outcomes and hence may reduce the generalisability of the findings. Fisher et al. (2020) systematic review of RCTs reported that female patients (n=3691) equalled male patients (n=3613) in the 34 RCTs that reported sex; drugs were assessed in patients with neuropathic pain (n=13), cancer (n=6), acute pain after surgery (n=4) and multiple sclerosis (n=10) and chronic prostatitis, carpal tunnel syndrome and back pain in one trial each. Furthermore, most rodent studies relied upon stimulus-evoked behavioural outcome measures, which is a measure of hypersensitivity. In conditions where evoked pain is a clinically relevant aspect (e.g., postsurgical and musculoskeletal conditions), hypersensitivity measures may be relevant (Pogatzki-Zahn et al., 2017) and within these models the treatment effects are significant indicating potential benefits for these conditions. However, in neuropathic pain conditions, spontaneous pain is more often assessed (Dworkin et al.,

2005), and in many common neuropathic pain conditions, such as diabetic neuropathy, sensory loss is more common than evoked hypersensitivity (Maier et al., 2010, Baron et al., 2010a). There was not scope within the original study, but I have since performed a post-hoc analysis to investigate differences between mechanical hypersensitivity across the different model types. The data, for which there are other influencing factors, including sedative effects of drugs, methodological bias and interaction between experimenter and the animal, suggest that all the pain-associated models were sensitive to mechanical evoked limb withdrawal. So, although treatment effects are also significant in the neuropathic pain associated models translation to the clinic might be limited. These findings add weight to the use of a broader range of outcome measures and the need for animal models to better reflect the clinical condition and underly pathophysiology so that model induction results in sensory profiles that reflect patients' sensory profiles.

5.4.2.2. Misalignment between the cannabinoids, cannabis-based medicines and endocannabinoid system modulators assessed in preclinical trials and clinical trials

A substantially more diverse range of potential therapeutics have been assessed for antinociceptive efficacy in animal models (171) than reported in the recent review of randomised controlled trials in which 11 interventions were assessed in patients (Fisher et al., 2020). In the clinical systematic review, evidence of benefit was found for cannabis <7 days and nabiximols >7 days. However, the studies had an unclear or high risk of bias with the evidence scored as low or very-low quality and the authors conclude that "the evidence neither supports nor refutes claims of efficacy". The following interventions were assessed in clinical trials in the clinical meta-analysis, but do not feature in the preclinical meta-analysis; nabiximols (n=17), dronabinol (n=2), cannabinoid receptor agonist (n=2; AZD1940, GW842166), and THC congener (n=1; benzopyran peridine) (Fisher et al., 2020). It is possible that these drugs have not been tested for efficacy in animal models and/or the studies in which their effects are described are not reported with sufficient detail to be included in the meta-analysis.

The observed misalignment between preclinical and clinical trials suggests that the animal studies are not being optimally used to inform or predict direction for clinical trials or efficacy in the clinic. The justifications for clinical trials were likely borne from patients and their use of cannabis to alleviate pain. Animal studies are being conducted concurrently to clinical studies providing opportunity for both translation and back translation. Animal to human translation will always be unpredictable, especially considering the challenges of pain research. However, animal studies in the field of cannabinoid research may have greater utility than is currently being recognised. They can be especially useful for providing mechanistic insights into the pharmacology of cannabinoids, cannabis

based-medicines, and endocannabinoid system modulators, to facilitate development of human therapies.

5.4.3. Internal Validity

All included studies had an unclear risk of bias. The propensity to report methodological quality measures to reduce the risk of bias was low, and this finding is commensurate with other pain preclinical systematic reviews (e.g. Currie et al. (2019)). It likely reflects the fact that the reporting of many of these quality measures has not been required by journals nor convention in the preclinical field until recently. The included studies rarely reported the performance of power calculations to determine sample size (3 %) or animal exclusions (14 %), however, randomisation and blinding were more frequently reported (32 % and 47 % respectively). Our analyses did not show a consistent relationship between the reporting of methodological quality and smaller effect sizes (akin to Federico et al. (2020)), however, larger effect sizes were reported in studies that did not report allocation concealment and sample size calculations, both accounting for a significant proportion of heterogeneity. The methods used to mitigate bias were rarely reported and it was therefore not possible to accurately assess the risk of bias, leading to uncertainty in the validity of the outcomes.

Sample size should be determined using a power analysis and experiments are required to use sufficient animals to be adequately powered. If too few animals are used, the scientific validity and reliability of the results will be questioned. In relation, many experiments compared multiple treatment groups with one control group. This reduced the sensitivity because the control group was divided across the multiple treatment groups (the mean number of animals in the control group was reduced from n=8 to n=3). Bate and Karp (2014) assert that many sample size calculations are performed on the assumption that the experimental design consists of only two groups, a balanced design, however experimental design is often more complex with the control group often serving multiple treatment groups. They provide a strategy for reducing the risk of false positives by increasing the number of animals in the control group, although this is also contingent on the other factors that comprise a sample size calculation (e.g. effect, variability, significance level, analysis and experimental constraints). Using more animals is commensurate with the reduction, refinement, and replacement (3R) principles because it ensures that the animal sacrifice is weighted against the highest possible gain of knowledge, although existing animal care and use committees may view this as a conflict.

5.4.3.1. Meta-analytical Approaches & Further Exploration of Heterogeneity

For the original manuscript, subgroup analyses (stratified meta-analyses for categorical variables) were performed to investigate how study characteristics influence the overall estimates of effects. The subgroup analyses provided empirical evidence to inform experimental design and refine

modelling of injury-related or pathological persistent pain and the extent to which predefined study design and study risk of bias characteristics differ in their overall estimates of effect. The study design factors found to be significant sources of heterogeneity using stratified meta-analysis were animal species, strain and sex, model type, outcome measure type, and drug class.

Multiple meta-regression was used to explain the true effect size differences in the dataset. Model type, drug class, outcome assessment type, and sex and strain accounted for 36 % (R^2) of the heterogeneity in rats, and 56 % for mice. However, the moderate and severe multi-collinearity of the factors pose a threat to the model. It should also be noted that the p-values were not corrected for multiple comparisons which increases the probability of false statistical inferences (Chen et al., 2017). Multi-model inference was used to explore the importance of each of the predictors. The aim of this method is to identify the important predictors for hypotheses generation. The results show that all the factors (model type, drug class, outcome assessment type, sex, and strain) are likely of high importance.

A meta-regression has been described to be a more reliable method for small datasets and therefore should be used (Tanriver-Ayder et al., 2021). However, systematic reviews are observational, and any findings should be confirmed experimentally. Here, I have presented both stratified analyses, and despite the added granularity of the meta-regression, the conclusions remain the same. The experimental design characteristics such as model type, drug class, species, strain, sex, and outcome measure influence the observed effect. Meta-regression methods may provide more useful information when exploring sources of heterogeneity for the effect of an individual drug and/or for an individual model type and particularly should be used for small datasets.

5.4.4. Publication Bias

Unlike a recent preclinical systematic review of pregabalin in which a 27 % overestimate of effect was theorised (Federico et al., 2020), our analysis does not suggest presence of publication bias, the phenomenon wherein neutral or negative studies are not published. The overestimate of effect theorised by Federico et al. (2020) may be an overestimate of the existence of publication bias due to the method that was used. It is important to consider the limitations, firstly, that publication bias is not the only cause of asymmetry in funnel plots. It is also referred to as a method to detect small study effects i.e., the smaller therefore less precise studies have a different true effect from larger studies because the intensity of the intervention was higher in the small studies (Borenstein et al., 2009). Secondly, artefacts may be responsible: Zwetsloot et al. (2017) have demonstrated that funnel plots of standardised mean difference (SMD) plotted against standard error (SE) are susceptible to distortion, leading to an overestimation of the existence of publication bias.

Simulations by Zwetsloot et al. (2017) indicated that distortion was more severe when the primary studies had a small sample size and when an intervention effect was present, the characteristics of the Cannabis & Pain review data. The visible distortion is a result of SE increasing as the size of the SMD increases when there is an intervention effect. Their simulations also indicate that using SMD in combination with SE is not suitable for publication bias assessments and can lead to false-positive results. Therefore, I followed the guidance by Zwetsloot et al. (2017) and plotted the SMD against a sample size-based precision estimate (1√n) and did not detect publication bias when using nested comparisons per study nor individual outcome comparisons. The funnel plot for individual comparisons shows asymmetry and trim and fill imputed 910 theoretically missing comparisons to the right-hand side, suggesting positive studies are missing. This is at odds with the phenomena that is publication bias and is likely a result of trim and fill analysis not considering reasons for funnel plot asymmetry beyond publication bias. It is built on the assumption that there should be a symmetrical funnel plot, a normal distribution of studies around a true or mean effect. The method is known to perform poorly in the presence of substantial between-study heterogeneity, therefore it is plausible here, given the high heterogeneity and breadth of study design characteristics that there are effects far from the global mean resulting in the asymmetry (Terrin et al., 2003, Peters et al., 2007).

This demonstrates that caution is required to ensure that adaptation of meta-analytical techniques from the clinical domain and it is necessary to ensure that they are appropriate for the characteristics of preclinical data i.e., high number of individually small studies that report large intervention effects. Unlike in clinical research, where meta-analyses have a small number of large studies with smaller intervention effects. Similarly, the inherent limitations and underlying assumptions need to be understood. The true mechanism for publication bias nor its magnitude cannot be known, however there is utility in attempting to provide empirical evidence of the impact of publication bias if the estimates are interpreted with caution. Failure to publish negative or null results is an ethical issue that continues to need to be addressed. This bias can seriously distort the literature leading to misguided research further contributing to research waste.

5.4.5. Utility of Preclinical Systematic Reviews

This systematic review and meta-analysis provided a comprehensive summary of studies in which cannabinoids, cannabis-based medicines, and endocannabinoid system modulators were assessed for antinociceptive efficacy in animal models of injury-related or pathological persistent pain. The overall behavioural data effect sizes are significant, and there is a spectrum of efficacy dependent on drug and pain model type, as well as on species, strain, sex. It has provided empirical evidence to support recommendations for the improvement of methodological quality for the design, conduct, analysis, and reporting of studies which assess interventions for anti-nociceptive efficacy. The stratified

analyses reveal that study design characteristics can influence the observed outcomes and therefore consideration should be given to the choosing the most appropriate model and outcome assessment that best reflect the clinical condition. There are also disconnects between the clinical and preclinical research and imbalances and gaps in the research e.g., external validity of animal models, drug types and dosing regimes, and assessment of side effects. This review demonstrates that there is a lot of breadth in terms of the interventions, disease models, animal characteristics and pain-associated behavioural outcomes. It appears that many siloed groups are working in several different directions with unclear rationales. There is a dominance of some drugs assessed repeatedly in the same model types and although rationale for conducting the research was not systematically investigated this hints at a lack of coordination. Coupled with the disconnect between preclinical and clinical research aims, there is a lack of multi-disciplinary collaboration which would facilitate forward and back translation. This review has provided recommendations to improve the rigor, transparency, and reproducibility of preclinical research as well as gaps related to understanding cannabinoid pharmacology and pharmacokinetics of drugs. The status of the findings did not support identification of priority candidates. It is outside the scope of a review of this kind to address the organisational and efficiency issues raised but is an important consideration for how research is conducted to prevent waste. Systematic reviews have an important role to play as they can ensure that new research builds upon the knowledge base and research effort is spent where required.

5.4.6. Limitations

Our systematic review has several limitations. Firstly, we can only rely upon what has been reported in publications. There were 99 studies that met the inclusion criteria but could not be included in the meta-analysis due to lack of reporting of key information e.g., variance, sample size or not having suitable controls. For the included studies, it is possible that methods were used to reduce the risk of bias, but not reported, conversely, these methodological quality criteria may have been reported but not performed. There are also other experimental design factors that will influence behavioural outcomes but are not included in our analyses e.g., housing, diet, handling, habituation (Mogil, 2017) and the sedative effect of the drugs.

We did not collect information on the aim of the studies. We collected all pain-associated behavioural outcomes of all interventions of any dose or route of administration that were being assessed for anti-nociception. The behavioural studies may have been conducted, not to determine efficacy, but for a different rationale and in some experiments the doses may have been lower than those required for anti-nociception. To not collect the aim of the studies was an oversight on our part however, it is rare that preclinical researchers report whether their studies are exploratory or confirmatory and nor do they state a primary outcome for which they perform a sample size calculation. To negate therapeutic

claims arising from efficacy studies that are not rigorously designed and prone to larger random and systematic variation, Kimmelman et al. (2014) make a compelling case for the disentanglement of exploratory and confirmatory research to improve translation. Doing so will simultaneously recognise the necessary freedom and resource constraints of exploratory research and allow resource intensive confirmatory studies be held to the highest of methodological standards.

5.4.6.1. Pain-Associated Behavioural Outcomes and Outcome Data Extraction

In our meta-analysis, we grouped together models by the same underlying biology. Similarly, we grouped together behavioural outcome measures. Most of the outcomes measured were limb withdrawal in response to mechanical or thermal stimuli and despite not having the same underlying biology, these were grouped together if a cohort was assessed in both. There is large variation in how these studies are reported and it is challenging to identify differences in study design when these are not often reported in detail. In addition, we chose to extract pain-associated behavioural outcome data at the time point where the difference was largest between vehicle and treatment groups. This allowed us to calculate treatment effects independent of the intervention's half-life, particularly pertinent as many studies did not report time course data. Although not feasible within this review, given the number of interventions, doses etc., a future approach may be to calculate area under the curve and percentage maximum possible effect for each experiment at all reported doses allowing more information, including dose-response, to be gleaned.

5.4.6.2. Crowd Science and Machine Learning

This review demonstrates that crowd science and machine learning are viable strategies to improve the feasibility of conducting a large review, both are discussed in following chapters. Briefly, crowd science offered a reduction in individual input by sharing the labour-intensive stages of the review (screening for study selection and annotation and data extraction phases). The limitations of individual crowd members have not undermined the findings of this review due to each study being reviewed by two independent reviewers and disagreements reconciled by a third independent reviewer. This review also demonstrated the usefulness of machine learning for study selection, albeit for error analysis. The machine algorithm performed with an eventual sensitivity of 95 % and specificity of 94 %. This high sensitivity made it unlikely to miss relevant literature.

The search for this review was conducted 2 years prior to publication; given the size of this review it is unlikely that the incorporation of more recent studies will change the overall conclusions. The continued development of online platforms and automation technologies are required to improve the feasibility of preclinical systematic reviews. This will create the possibility to incorporate the most recent data as it becomes available in the form of a living systematic review (Elliott et al., 2017)

thereby addressing the challenge for the future appraisal of the rapid and exponentially increasing volume of published preclinical literature.

5.5. Future Research

The number of studies that concurrently conducted and reported pharmacokinetic investigation (24 studies) including drug concentrations in tissues post administration (25 studies) was low at 7 % of studies. There is a need to conduct pharmacokinetic studies alongside pharmacodynamic studies to determine the relationship between plasma/tissue concentrations of treatments and reflexive or complex nociceptive behavioural measures. Few studies assessed the effects of the drugs on sedation with 124 studies (33 %) assessing impact of drugs on motor activity, which is a particular concern for direct acting CB_1 receptor agonists (but would not be expected to be a confound for CB_2 agonists, FAAH inhibitors, anandamide transport inhibitors or CB_1 allosteric modulators). Similarly, few studies assess the anxiolytic or depressive-like effects (11 studies, 3 %) of the drugs which may also compound pain-associated behavioural outcomes. A broader range of assessment is required to determine the full behavioural effect of these drugs (Negus, 2019). It was outside the scope of this systematic review and meta-analysis to assess the possible side effects of these treatments, but to assist future reviews, studies assessing pharmacokinetics, locomotor activity, anxiety and depression were annotated.

Rigorous preclinical design is required for internal validity (Huang et al., 2020). To limit threats to validity, we endorse conducting and reporting animal experiments as suggested by Andrews et al. (2016a), Vollert et al. (2020), and in accordance with the ARRIVE guidelines (Percie du Sert et al., 2019). To improve external validity, researchers should balance the sexes as stipulated by funding bodies including the USA's National Institutes of Health (Clayton and Collins, 2014, Tannenbaum et al., 2016). Historically, emphasis has been placed upon reflex withdrawal responses rather than measuring more complex, ethologically relevant behaviours, although the clinical relevance of these measures to analgesia remains uncertain. The development of sensory profiling for rodents and complex behavioural animal models specific to each pain condition is required to improve face and predictive validity and better reflect the clinical situation (Rice et al., 2017). Multicentre testing may offer a method to improve the generalisability of preclinical findings by increasing environmental heterogeneity and study samples (Wodarski et al., 2016, Voelkl et al., 2018). To address the potential issue of publication bias, we recommend researchers make available prespecified protocols and publish all results (i.e., positive, null, and negative data). To assist in the optimisation of experimental design we encourage primary researchers to conduct prospective systematic reviews and use the UK's National Centre for the Replacement, Refinement and Reduction of Animals of Research Experimental Design Assistant (EDA; https://eda.nc3rs.org.uk/) to inform their research design and protocol development (du Sert et al., 2017).

Recommendations were made in both the systematic review (Soliman et al., 2021) and a combined TF future research agenda (Haroutounian et al., 2021) was produced. The recommendations for research priorities and improved rigor and transparency as discussed above are summarised in tables 5.15 & 5.16. The dataset can be used to explore further the different potential sources of heterogeneity and factors that are likely to influence the observed outcomes e.g., drug route of administration, dosing regime, and drug dose response.

Table 5.16. Research priorities for assessing anti-nociceptive efficacy of cannabis-based medicines, cannabinoids, and endocannabinoid system modulators in animal models of persistent pain.

1	Better alignment between compounds tested in clinical trials with those tested preclinically, to allow improved understanding of translational (and back translational) pharmacology of targeting the endocannabinoid system for analgesia.
2	Detailed characterization of pharmacokinetic properties of cannabinoids, and determination of PK-PD relationship between plasma/tissue concentrations and antinociception in the context of specific preclinical models.
3	Improved external validity of animal models and outcome measures used to determine antinociceptive effects of cannabinoids and endocannabinoid system modulators.

Table 5.17. Research priorities relating to the improved rigor and transparency of design, conduct and reporting of preclinical trials assessing anti-nociceptive efficacy of cannabis-based medicines, cannabinoids, and endocannabinoid system modulators in animal models of persistent pain.

1	Protocol/Preregistration – primary outcomes, hypothesis generating or confirmatory, adherence with deviations reported, adequate sample sizes
2	Conduct in accordance with ARRIVE
3	Include all animals in analyses
4	Report in accordance with ARRIVE – including full reporting of animal adverse events/death with reasons
5	Full declaration of commercial and non-commercial interests held by the authors should be disclosed and declared. Any funding received directly for the study should also be disclosed

The dataset is shared on the open science framework and is freely available for researchers to use (https://osf.io/2qde5/). There are other analyses that can be performed particularly at a finer granularity which was not in scope for this review e.g., exploration of differences in model types, differences in dosing regimes. In terms of meta-analytical approaches, there are many choices that reviewers can make that will impact the conclusions. This dataset can be used to better understand meta-analysis methods and which methods are most appropriate for this type of dataset. It offers the opportunity to investigate the generalisability of the results reported in the study by Tanriver-Ayder et al. (2021) which compared the commonly used methods in random effects meta-analysis for preclinical drug discovery research.

5.6. Conclusion

This systematic review and meta-analysis provided a comprehensive summary of studies in which cannabinoids, cannabis-based medicines and endocannabinoid system modulators were assessed for antinociceptive efficacy in animal models of injury-related or pathological persistent pain. The behavioural data effect sizes are significant, and the evidence supports the hypothesis of cannabinoid-induced analgesia. Most drugs tested in animal models were small molecules, which is converse to the clinical situation where cannabis extracts have been evaluated most in clinical trials. The differences between the animal and clinical population highlights the importance for the development of better validated animal models. Behavioural assessments that have greater clinical relevance may also improve the likelihood of the development of effective therapeutic interventions. There is also a need to continually improve clinical trial design in a manner that is informed by high quality, mechanism-based preclinical research. Despite the 'unknown' predictive value of many animal studies, there is value in conducting a prospective systematic review to aid clinical trial decision making. The findings of this review support the need for preclinical living systematic reviews and closer, multi-disciplinary, cross sector collaboration to ensure that animal studies are rigorous to identify potential candidates and more accurately inform clinical trial design.

6. Crowd Science: Challenging Traditional Ways of Working

6.1. Introduction

Crowd science was employed to conduct the study selection and data extraction phases of the Cannabis & Pain review (Soliman et al., 2021). The review recruited and trained over 100 volunteer participants to perform the study selection and data extraction stages of the systematic review. The training was supported by guidance notes, newsletters, and personal communication. The participants of the crowd constituted the double screening and data extraction per study. Challenging the traditional ways of working, the aim was to determine if utilising a crowd could improve the efficiency and accuracy of a systematic review in comparison to a small, specialist author team (Franzoni and Sauermann, 2014). Utilising a crowd also offered the opportunity to engage with the wider pain research community.

A growing amount of research is conducted in an open collaborative fashion, in projects referred to as "crowd science", "citizen science," or "networked science". Crowd science aims to draw on the effort and knowledge inputs provided by a large and diverse base of contributors, potentially expanding the range of scientific problems that can be addressed at a relatively low cost, while also increasing the speed at which they can be solved (Franzoni and Sauermann, 2014). Human effort can contribute to systematic reviews in two ways, either by conducting specific tasks or providing examples that can be used to train machines which can than automate or semi-automate specific tasks. Systematic reviews are conventionally conducted by a small team of trained researchers working in a highly labour intensive, but time limited way. The most labour-intensive stages are screening, annotation, and outcome data extraction. Each stage requires a minimum of 2 reviewers and a 3rd independent reviewer to reconcile any disagreements thereby minimising the risk of bias and ensuring reliability. However, by breaking down the review process into microtasks, systematic reviews can be conducted more efficiently by a wider range of people and depending on the skill base, the crowd can be utilised to conduct tasks that require common skills through to specialised and expert skills.

An important feature of crowd science is that participation in a project is open; anyone who is interested and qualified is invited to join, often based upon self-selection in response to a general calling notice. A challenge, however, is attracting contributors and coordinating the contributions of many participants, although features of systematic review platforms including the Systematic Review Facility; SyRF assist with the coordination of many participants (Bahor et al., 2021).

Crowds have been shown to offer a less labour-intensive model allowing the authorship team to make more efficient use of their skillsets, allocate time to more specialist tasks within the systematic review and take advantage of emerging automation systems to reduce the workload. For example, study selection, a time-consuming and repetitive task, is not automated and therefore at this stage a crowd is particularly useful. Highly intensive human effort makes mistakes more likely, especially when fatigued by significant time spent on repetitive tasks. Therefore, utilising a crowd can achieve the necessary high recall and potentially be better, than that often achieved by using more conventional approaches. A good example of how the human crowd and machine can operate in mutually reinforcing ways is through the development of the Cochrane Randomised Control Trial classifier (Wallace et al., 2017). The crowd is supported by brief, interactive online training and accuracy is ensured using an agreement algorithm that requires each study to be classified multiple times before it is either submitted or rejected. As the Cochrane crowd screened more studies the classifier became more accurate thereby reducing the workload by removing the studies that are not eligible in turn making the crowd more efficient. More recently, Noel-Storr et al. (2021a) demonstrated that a crowd can perform with 100 % sensitivity i.e. correctly identify a specific randomised clinical trial, as per the PICO criteria (population, intervention, comparator, outcome), for inclusion. Together, these studies demonstrate the potential of crowd science for workload reduction in study selection.

6.2. Aims and Objectives

The systematic review aimed to identify and critically appraise all the available preclinical literature in which the efficacy of cannabis-based medicines, cannabinoids and endocannabinoid system modulators were assessed in animal models of persistent pain. A large and increasingly active field, it was anticipated that the search (which was purposefully inclusive) would return thousands of potentially relevant articles and study selection would identify hundreds of studies for inclusion. Given the volume of work, crowd science was employed for both study selection and data extraction phases of the review. The aims therefore were to recruit, train, and employ a specialist crowd in the screening and data extraction stages of the review and assess if utilising a crowd could improve the feasibility, efficiency, and accuracy of a systematic review.

6.3. Methods

6.3.1. Ethics

Ethical approval to employ a crowd was obtained in March 2019 from the Imperial College London, Surgery and Cancer, Head of Department and Science, Engineering and Technology Research Ethics Committee (SETREC; available on the Open Science Framework (osf)).

6.3.2. Recruitment

The recruitment strategy aimed to predominantly target bio-scientists who were interested in contributing to an important area of research and looking for professional development opportunities. There was not a pre-requisite criteria for joining, participants were required to pass training modules for both stages. An advertisement was developed which was shared widely through the International Association for the Study of Pain (IASP) network, collaborators and colleagues and students using direct communication, email, newsletters, social media and face-face interactions at conferences and workshops.

Participants were asked to email in response to the advertisement and once they received the participant information sheet, they were asked to complete the registration form and consent to participating. Once these enrolment activities were complete the participant was invited to commence the training. There was not a required minimum/maximum number of participants to meet the needs of the review, however the more people that were recruited the more broadly the workload could be shared.

Beyond professional development and international collaboration, other incentives including authorship were offered. Participants qualified for authorship upon completing more than 350 screening decisions and extracted data from 35 studies and because the recruitment stayed open throughout the duration of the project once screening was complete authorship was available for those that completed data extraction of 40 studies. The top 10 reviewers were offered a book of their choice from IASP press, the top 5 reviewers received free registration to IASP's 2020 World Congress and the top reviewer a 32 GB Apple iPad/equivalent tablet.

6.3.3. Training

Training modules were developed for both stages and hosted online on the platform (Learn to SyRF; Learn to Screen and Learn to Data Extract). These were supported by a manual with comprehensive instructions and guidance notes (https://osf.io/cn8ep/). Each module had pass rates set that were informed by previous reviews.

To pass the screening training, participants needed to correctly make the include or exclude decisions for 10 consecutive articles based upon the title and abstract. For data extraction, participants were presented with studies to extract data from, and their data extraction compared with a 'gold standard' derived following review of the set of studies conducted by the in-house team. Once their concordance was greater than 80 % for three consecutive studies, they were able to join the project. Participants were not able to contribute to the review if they did not complete or failed the training. However, they were able to continually conduct the training until they met the required standards.

6.3.4. Communication

Several communication strategies were employed to aid and encourage participation and to assist in further recruitment and retention. These included:

- Newsletters. Using mailchimp (an email marketing platform) regular newsletters were sent to all who had signed up. The newsletters aimed to keep everyone informed on the progress of the project, share key information, links to participatory instructions and share lessons learned. They also provided further training by contextualising elements of the review processes e.g., explaining the risk of bias, animal models for pain research, and included a reading corner for reference material.
- Individual email/video conference. Crowd participants were welcome to email at any time and when necessary, video calls were arranged.
- Collaboration Club. 2 h once a month were allocated for shared working using video conferencing. Crowd participants were invited to join me so that I could address any questions they had and so that we could work together (Pomodoro style) for that period; a form of committing/allocating time to the project.
- Feedback. Crowd scientists were continually encouraged to provide feedback throughout the course of the project. A formal feedback survey was sent to all those that contributed to the project upon its completion. Demographic data, experience, and impact of participating and accessibility and usability of SyRF were collected. Feedback was anonymous.

6.3.5. Efficiency

Efficiency is defined as achieving an end goal without wasting resources e.g., time, money, effort. In this context, efficiency is related to speed (time taken to complete the review stages), the quality (accuracy) and the cost (in this case predominantly manpower). The quicker the speed, the higher the accuracy and the lower the cost the more efficient a process/system is.

6.3.5.1. Speed

Average completion times of studies per week were calculated based upon reporting from crowd scientists. These were used to compare against the likely timelines for the traditional systematic review team approach of 2 independent reviewers for each stage.

6.3.5.2. Accuracy: Inter-rater reliability

It was not possible to perform quality control or assess agreement during the project however, inter-rater reliability was calculated using Cohen's kappa to assess, by proxy, the quality of the data extraction by crowd scientists prior to reconciliation. This could only be performed on the study level annotation questions due to the complex nature and structure of the other data types that were extracted.

6.3.5.3. Cost

Cost was calculated by totalling the approximate number of hours worked and multiplying by the Imperial College London's hourly rates for graduate student demonstrators (Imperial College London, 2020).

6.4. Results

6.4.1. Crowd Recruitment, Training and Contribution

453 people from 44 countries signed up to the project. Most crowd scientists were from the United States and Northern Europe, however lower-middle income countries were also represented (Figure 6.1). Most were early career researchers and had widely varied experiences of the topic and systematic review processes. Based upon the survey feedback most people joined the project through colleague referrals.



Figure 6.1. Locations of crowd scientists pinned to a world map.

Of the 453 recruited, 214 crowd scientists enrolled onto the training modules on Learn to SyRF. 51 % completed the screening training and 53 % completed the data extraction training. 100 went on to make a meaningful contribution to the project, 28 made a large enough contribution to meet the authorship criteria (Table 6.1).

For study selection, the number of studies screened by crowd scientists ranged from 1 to 7141. The mean was 630 studies, and the median was 234 studies. 89 % of the workload was shared between the 14 crowd scientists who contributed over the authorship threshold. For data extraction, the number of studies completed by crowd participants ranged from 1 to 172. The mean was 30 studies,

and the median was 25 studies. The 12 crowd scientists that contributed over the authorship threshold accounted for 60 % of the workload. Together the results show that a small share of contributors make the largest share of contributions. Additionally, 22 of the author crowd scientists stopped contributing once they met the threshold.

Table 6.1. The number of crowd scientists that enrolled for training, completed the training, and contributed to the project at each stage of the review process.

Stage	Enrolled	Passed Training	Contributed
Screening	91	46	36
Data Extraction	123	65	64

6.4.2. Communication

A total of 19 newsletters (available through Mailchimp) were sent to all crowd scientists who signed up to the project. Six Collaboration Clubs were held and were attended by 5 crowd scientists. Three (2 in person) lectures and interactive, trainer-led sessions were conducted for groups from the National University of Ireland Galway and King's College London. Over 1000 emails were sent to individuals, addressing individual needs. Once individual connectivity needs were met the questions were focused on specific studies and the complex way in which preclinical studies are reported. It was through these early studies that crowd scientists provided much needed and valued feedback for the improvement of the annotation and data extraction forms as well providing insight for more contextualised explanations of what was required - these were communicated in both the guidance notes and newsletters to ensure that all lessons were shared. For example, there was a request for further training for the use of tools for extracting data from graphs, so I recorded two videos demonstrating how to perform data extraction with webplotdigitiser and adobe ruler as well as clarification for which data points reviewers should be extracting. Other examples related to how to navigate and use the SyRF platform including an early lesson highlighting a security issue, whereby reviewers had the ability to change the form and corrupt the data extraction process.

6.4.3. Efficiency

6.4.3.1. Speed

Based upon crowd scientist feedback and real-time monitoring the approximate timeframes were calculated for each stage and due to the highly heterogenous nature of the studies across a full-time week (40 h). Screening was calculated to be 1000 studies per week per person, and data extraction was calculated to be 25 studies per week per person. The crowd took a total of 43 weeks to perform

both stages of the review process; 6 weeks to complete the screening, a saving of 4 person weeks (60 %) and 37 weeks to complete the data extraction: 7 more person weeks, an estimated 23 % increase in time compared to 2 specialist reviewers (Table 6.2).

Table 6.2. Time appreciation of both the screening and data extraction phases of the review. Includes comparison of the real time data of the crowd compared to theoretical timeframe of 2 specialist reviewers.

Stage	**Number of Studies (per week per person)**	**Crowd (weeks)**	**2 specialist reviewers (approx. weeks)**
Screening	1000	6	5 (10 person weeks)
Data Extraction	25	37	15 (30 person weeks)

6.4.3.2. Accuracy

Crowd scientists' inter-rater reliability was calculated using Cohen's Kappa and percent agreement for 15 study level annotation questions to which the possible answers were the categorical variables "yes", "no" or "not applicable". Cohen's Kappa ranged from 0.18 [95 % CI 0.00 – 0.44], slight agreement for 'Did the study investigate the effect of treatment(s) on dependency?' to 0.88 [95 % CI 0.78 – 0.97], near complete agreement for 'Did the study use electrophysiology?' Percent agreement ranged from 60 % for 'Did the study confirm the CB_1 or CB_2 receptor as the target?' to 98 % for 'Did the study use electrophysiology?' Overall, Cohen's Kappa suggests substantial agreement (Kappa = 0.70 [95 % CI 0.68-0.72] and percent agreement was 87 % (Table 6.3).

Table 6.3. Crowd scientists' inter-rater reliability calculated using Cohen's Kappa and % agreement.

Annotation Questions	Agreement	Disagreement	Cohen's Kappa (95 % CI)	% Agreement
Study Design				
Did the study investigate the potential toxic effects of the treatment(s)?	333	35	0.36, (95% CI: 0.15-0.56)	90.49
Did the study investigate the effects of treatment(s) on dependency?	335	33	0.18, (95% CI: 0.00-0.44)	90.15
Did the study investigate the pharmacokinetic properties of the treatment(s)?	332	36	0.37, (95 % CI: 0.17 - 0.56)	89.16
Did the study measure markers for neuronal activity?	339	29	0.47, (95% CI: 0.34-0.61)	91.45
Did the study investigate the effects of treatment(s) on motor activity?	306	62	0.62, (95% CI: 0.53-0.71)	79.74
Did the study use electrophysiology?	362	6	0.88, (95% CI: 0.78-0.97)	98.34
Did the study confirm the CB1/CB2 receptor as the target?	263	105	0.47, (95% CI: 0.38-0.55)	60.08
Did the study investigate the levels of drug(s) in tissue?	336	32	0.40, (95% CI: 0.20-0.60)	90.48
Methodological Quality				
Allocation concealment	297	71	0.25, (95% CI: 0.10-0.41)	76.09
Randomisation	338	31	0.79, (95% CI: 0.72-0.86)	90.83
Sample size calculation	340	18	0.36, (95% CI: 0.08-0.64)	94.71
Blinded assessment of outcome	313	55	0.70, (95% CI: 0.62-0.77)	82.43
Pre-determined animal inclusion criteria	318	50	0.23, (95% CI: 0.10-0.37)	84.28
Other				
Conflicts of interest	324	44	0.76, (95% CI: 0.70-0.83)	86.42
Compliance with animal welfare regulations	349	19	0.61, (95% CI: 0.44-0.78)	94.56
Overall	**5506**	**721**	**0.70, (95% CI: 0.68-0.72)**	**86.91**

6.4.3.3. Cost

This project gained approximately 1,720 hours of volunteer work, the cost of which has been calculated using Imperial College hourly rates (in effect since December 2020) for student demonstrators (Table 6.4).

Table 6.4. Estimated cost of crowd scientist labour based upon the hourly rate of a student demonstrator at Imperial College London.

Pay Rate	Pay Rate Description	Hourly Rate	Crowd 'Cost'
Higher rate	Duties requiring a larger element of preparatory work such as most types of tutorials and seminars	£27.05	£46,526
Lower rate	For duties requiring little preparation or marking such as demonstrating	£15.70	£27,004

6.4.4. Crowd feedback

Thirty-two crowd scientists completed the feedback survey providing insight into their experience of joining and contributing to the project (Figure 6.2) and the impact that it has had upon them (Figure 6.3 and Table 6.5). There were also a series of questions relating to the usability of the SyRF platform (Table 6.6).

6.4.4.1. Joining, contributing and impact

The feedback was very positive. In all cases most participants agreed or strongly agreed that the aims of the project were well explained (97 %), that the training adequately prepared them (100 % for screening and 84 % for data extraction), the newsletters were timely and informative (91 %), and they felt well supported (97 %). In terms of impact most respondents agreed or strongly agreed that contributing to the project has improved their understanding of preclinical systematic reviews (97 %), study design (50 %), and experimental bias (93 %), and contributing has enhanced their professional skillset (91 %). Of the 32 survey respondents 15 have since contributed to another systematic review, 11 have started their own systematic review and 20 report improving their own research practises. Eleven respondents provided free text responses to the question, 'Please provide any further feedback or reflections based on your experience contributing to the project.' The feedback was very positive with 8 respondents reporting enjoying the experience, 5 respondents reported that contributing provided learning opportunities, 4 respondents reported good communication, and 3 respondents reported it to be a great opportunity. One respondent reported being initially confused by the task

and the long instructions, another requested the inclusion of specific training for extracting data from graphs.

6.4.4.2. SyRF Feedback

Respondents were asked to rate on a scale from 0 - 'not easy' to 10 - 'very easy' the different functions of the SyRF platform. Responses are broken down into three categories; detractors (0-6), passives (7-8), or promoters (9-10) and responses were converted into Net Promoter Scores (NPS) reporting a number from -100 to +100. Most criteria received a low positive NPS, with most respondents being promoters or passive. However, 'how intuitive was the annotation form?' received a NPS score of -16, with more respondents being detractors than promoters. Seventeen (53 %) respondents said that they would consider using SyRF for their own systematic review and 20 (63 %) said that they would recommend SyRF to other researchers. Twenty-six respondents provided free text responses to the question, 'Please provide any further feedback or reflections based on your experience using the SyRF application. We are particularly interested in your suggestions for improvements e.g., new features and support resources.' In this section several issues were raised; the platform would slow down with large data entry (9 respondents), the platform would regularly crash (9 respondents), the ability to be able to return to previously started, incomplete reviews through the interface was requested (7 respondents), 2 respondents described the annotation form as 'not intuitive' and one respondent raised the issue of the time it takes for transcribing extracted outcome data.

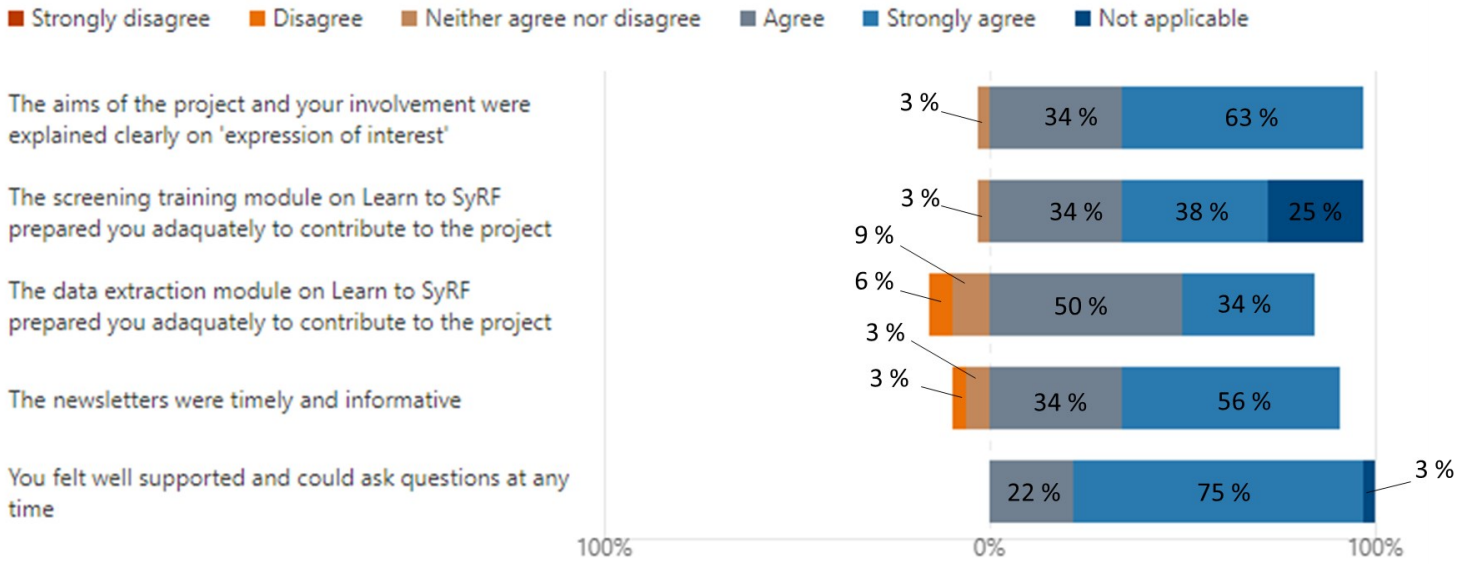


Figure 6.2. Crowd scientists' feedback on joining and contributing to the project. Survey results based upon the feedback of 32 participants.

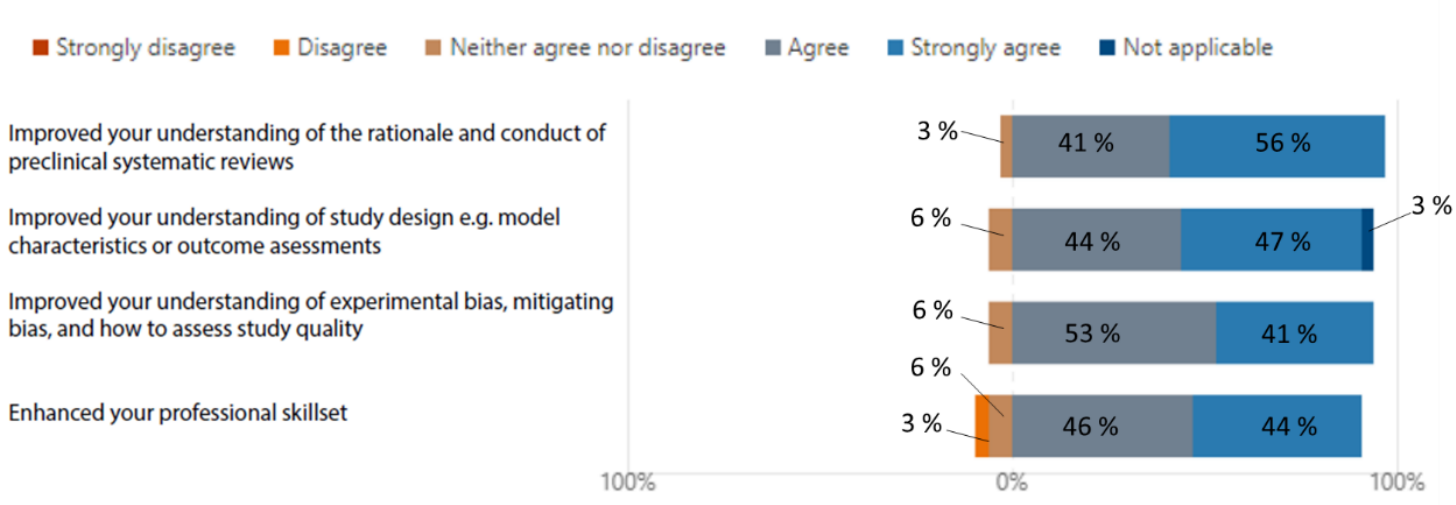


Figure 6.3. Crowd scientists' feedback on the impact that contributing to the project has had on them. Survey results based upon the feedback of 32 participants.

Table 6.5. Free text feedback from crowd scientists relating to their experience contributing to the project.

Please provide any further feedback or reflections based on your experience contributing to the project
"I have thoroughly enjoyed the overall process of the project from the beginning to an end. For example, in the screening part of the project, I was confused in the beginning of which specific criteria I should exclude or include the study. However, after several repetition of the screening process, I was well adapted to the learning curve and challenged myself to finish screening as much as I can. In the data extraction phase, I was also confused about the long page of instructions. However, after practicing several times and carefully looking at each journals, I have learned a lot from the data extraction process, learned new vocabularies and concepts as I was just an undergrad, and now I'm grateful to experience this huge opportunity and learn the process of what the PhD student looks like when I attend graduate school at sometime in the future."
"Nadia was very accessible via email, responded with great clarity to questions and very timely. A great experience, and best of luck with the publication."
"THE INVESTIGATION WAS CARRIED OUT IN AN ORDERLY AND CLEAR WAY."
"This type of approach to research is, I believe, the future for science. Especially in circumstances such as what we are currently facing with lockdowns locally and internationally - crowd science and online training is totally viable. Thank you for the great training and opportunity!"
"this was a very exciting experience. interesting to be a part of an international team"
"Overall, I really enjoyed being part of this project and I learned a lot from this experience :) The only suggestion I have is to slighlty change the data extraction training by adding a part in which trainees have to extract data from tables/figures using tools such as WebPlotDigitizer."
"I think the newesletters were great and helpful and solved many arising problems. Sometimes the Syrf platform crashed or for studies with a lot of data to extract saving and filling in took very long since syrf needed to load that long or crashed."
"Nadia was an excellent communicator and gave very clear feedback and responses. I do not work in animal studies, and so I really valued being trained in understanding methods of animal studies. I'm sure this will come in handy in the future for me. I am very pleased to be a part of this crowd source project!"
"contributing and participating in this project was an outstanding experience, it also improved my way of writing manuscripts and design studies to include essential information and to prevent bias."

"It was a fantastic experience, and nadia was always available in a timely matter; this made my contribution much easier."
"Really great to have been able to contribute to the project, thanks for the opportunity!"

Table 6.6. Crowd scientists' feedback on the usability of the SyRF application.

	Promoters (No. of respondents)	**Passives (No. of respondents)**	**Detractors (No. of respondents)**	**NPS**
How easy did you find registration and signing in?	12	15	5	22
How easy did you find the application to navigate?	13	16	3	32
How intuitive was the screening process?	12	13	7	16
How intuitive was the annotation form?	7	13	12	-16
How useful did you find the progress indicators for screening and annotation?	14	11	7	22
Did you feel supported using SyRF?	13	12	7	19

Table 6.7. Free text feedback from crowd scientists relating to their experience using SyRF.

Please provide any further feedback or reflections based on your experience using the SyRF application. We are particularly interested in your suggestions for improvements e.g., new features and support resources
"During the screening, please, add a possibility to select a reason for exclusion of a study; it would be great to implement functions to perform analyses within SyRF (or at least a possibility to make screening flow chart and some graphics and simple descriptive statistics on the included and excluded studies); the possibility to edit questions in the annotation form instead of deleting them and create the new ones (I think the main problem with designing the form for data extraction is an impossibility to edit even typos in the already created questions); the more cohorts I make, the slower the app works: if a study includes more than 7-8 cohorts, the app slows down and crash reports occur very often, so it would be great to have this fixed. I have not worked with the resulting database, so cannot suggest anything but the rest is great! Thank you very much for this really useful tool!"
"Sometimes the site was down, but otherwise it was a really good experience. It would be great if the system could keep a record of the manuscripts that each author previously edited (like an author home page with clickable links)."
"Continue with guidance and update newsletters"
"In the beginning, I was uncomfortable about using the app due to multiple times of deletion of the record in the data extraction phase. However, I'm relieved that I was able to finish entering the data much faster after developing new applications of SyRF."
"I did not like that I had to save the URL for each data extraction record, otherwise I would lose the record and unable to access it. Also, the program crashed often, and I had to save consistency to avoid losing work when it crashed."
"The annotation section was not user friendly at all"
"Was a great platform to use for such a large scale study"
"The only issue that arose was the server crashing with the use of the graph.syrf and app.syrf application which slowed progress!"
"Suggested functions: - Allow users to save and look at previously completed studies. - Make sure the server will not crash or slow down after a large amount of data is entered. - Provide feedback on the first few completed data extraction will be useful for improving further data extraction. Overall, it was a great application."

"I think for a new online software it was good. Obviously it crashed a lot which was often very frustrating when in the midst of submitting large quantities of data... but I understand it was a difficult task to undertake and it seems very adjustable for different needs so it seems it will take up more space and processing power I guess. I think if they worked on the robustness and user friendliness (the addition of tables which can be copied to excel etc.) it would be an amazing and powerful way to process systematic reviews and meta-analyses. I will definitely consider it's use in the future for individual and crowd based systemic reviews!"
"System becomes less responsive as the amount of data/cohorts increases"
"AT THE FIRST IT WAS DIFFICULT TO USE THE DATA EXTRACTION PLATFORM BUT OVER TIME EVERYTHING WAS EASIER"
"Adding a feature to allow users to re-visit studies which are still in progress for extraction."
"Unfortunately the SyRF application has a tough time handling very large studies that result in many entries. As a result the SyRF application became very slow, and with some studies I had to wait half a minute for every letter I wrote or every click that I made. For some very large studies I even had to give up after waiting for a long time for the application to respond. This problem has to be resolved before it for real can be used."
"Very clear and informative interface! Maybe make it easier to save and come back to the paper you're currently screening though"
"The application repeatedly crashed when entering more than 10 or 15 cohorts/datasets."
"there was maybe a few opportunities to train the extraction skills"
"When extracting data from a study, I had to save the study link on my PC, because SyRF doesn't allow users to go back to "in progress" studies. It would be great if users could access a study they're working on simply by clicking on "incomplete studies" on the progress indicator bar."
"less redundancy, less clicking, eg regarding sex. Sometimes syrf changed numbers after saving - eg wrong sample sizes or weight, than it needed to be rechecked and corrected and resaved. There was also sometimes a problem with decimals seperated by "," or "." depending if pasted into the table form or not the numbers were changed.And syrf took a long time for studies with many cohorts, or experiments. than just typing in data could take hours. It would be better to have a big table for all cohorts in one experiment and not opening one table for every cohort."
"The only thing that was tricky was, if you did not save the URL, you could not go back to that study. If I had a page where I could see all the studies I had started and my progress on them, that would be brilliant."
"When handling a lot of data in SyRF, the website got too slow. Sometimes an entry of a single letter would result in processing for several minutes. This is not sustainable and should be fixed ASAP."

"sometimes SYRF webpage aborted and crashed and especially with high data volume the working process was very slow or did not work, and data got lost. The typing of numbers in the data extraction field was sometimes not so intuitive"
"The platform was extremely slow when a lot of data was extracted"
"In order to use SyrF platform also for large datasets (many experimental groups) the server capacity must be increased to shorten the response time. This would significantly improve the workflow."
"System crashed on occasion, trying to minimise that. Otherwise, really good!"
"It was not very intuitive, after a pause I felt I had to start all over again"

6.5. Discussion

A crowd of scientists were successfully recruited, trained, and contributed to the study selection and data extraction stages of a preclinical systematic review (Soliman et al., 2021). Their contribution was successfully facilitated through the platform SyRF and formal and informal communication methods. In terms of efficiency savings, the crowd reduced the individual workload and consequently the financial burden, but the anticipated saving of increased speed was not met. This is coupled with unclear levels of accuracy. However, crowd feedback indicates a host of other benefits that are more difficult to quantify including, community engagement, training and education including professional skills development and feedback for tool development. The analysis provides insight for the development of future crowd science projects.

6.5.1. A small number of contributors make the largest share of the contributions

Of the 453 people who signed up, 100 people made a meaningful contribution. The training was a hurdle, with approx. 50 % not completing it. However, once individuals passed the training they remained committed to the project. This demonstrates the need to aim to recruit many more to account for attrition. Akin to other crowd-based projects e.g., Wikipedia (Ortega et al., 2008) and Zooniverse (Sauermann and Franzoni, 2015), this project was characterised by strong inequality in individual contributions with a small number of scientists taking the majority share of the workload. Similarly, analysis by Sauermann et al., 2013 suggests that most contributors only participate for a short period of time and the individuals that contribute for longer are then responsible for the project achievements. Therefore, the high attrition experienced, and the few contributors making the greatest contribution was to be expected. This raises the question of what the motivations of the top contributors were and if other factors could be influencing their higher levels of activity e.g., knowledge of systematic review processes or interest in cannabis preclinical research. A limitation is that individual crowd member motivations were not collected however, the motivations are likely to be mixed and dependent upon relevance, experience, subject matter interest, career stage, and capacity to contribute.

Sauermann and Franzoni (2015) analysed contributor patterns across seven online crowd science projects and report that top contributors speed up over time, although their contribution reflects more hours worked. Aligned with the crowd scientists' feedback, time on the project was required to understand what was required and to learn to how to navigate and use the platform. Therefore, future projects should more closely investigate the different types of contributors and their individual effort to identify means to reduce activation thresholds and increase participation in terms of number of contributors and the number of contributions they make. Individual motivations, complexity of the task and the usability of the platform are likely to influence participation.

6.5.2. Efficiency Savings May be Possible

Crowd science supplies a large labour force at a low cost. The estimated monetary value of the approximate 1,720 hours contributed was calculated at the hourly wage of a graduate student demonstrator at Imperial College London. However, this does not account for the seniority of individuals whose time has a higher cost and therefore the estimates are likely to be conservative. The analysis shows that this project attracted a considerable volunteer effort that would be difficult and costly to pursue through traditional academic funding/employment models. Therefore, drawing upon unpaid, motivated, volunteers is a viable strategy for performing research that would exceed modest research budgets.

This relates to speed and time saving. Although it is estimated that the crowd took 3 weeks longer than 2 specialist reviewers it is unlikely in an academic research setting that two full time staff will be committed to only one project, additionally, their salaries would exceed the projected cost saving of a crowd. In terms of time, project leads should also be cognisant that volunteers will have competing commitments and cannot be relied upon in the same way as employed staff. Similarly, as demonstrated by most of the authors, contributors are likely to commit to achieving a minimum target that they set for themselves within a timeframe that is plausible for them. Recruiting more crowd scientists will ensure that tasks are completed quicker than a specialist reviewer team. Lastly, time saving can be a fallacy if the quality or accuracy of the work is not to standard. It is for this reason, that focus should not be on time alone, and efforts spent trying to ascertain time spent on task may not be useful for gaining accurate estimates of the length of time it takes to complete a task.

With Cohen's Kappa the reviewers were determined to have 'substantial agreement' and a high percent agreement of 87 %. Agreement is a proxy for accuracy and could only be determined for the study level annotation questions. The reconciliation process identified ~20 % of completed studies that were not conducted to the standard and required re-reviewing. This was due to both technical and subject matter reasons. It is therefore likely at the more complex stages of the data extraction inter-rater agreement would reduce and due to the more complex nature and structure of the data it was not possible ascertain. This highlights that complexity of the task is an important consideration for crowd science, but the risks can be mitigated through suitable training and having a means to monitor project contributions and identify and mitigate issues as they arise.

There are two limitations to using Cohen's kappa to ascertain inter-rater reliability. Firstly, it assumes that the reviewers are two individuals that assessed the same studies, whereas in this instance, reviewer 1 and reviewer 2 are comprised of multiple crowd participants and therefore the same people are not always being compared. However, this may be offset by the scale of the review, the

randomisation of studies and the randomisation of reviewers which therefore assumes even distribution of reviewers to studies. Secondly, the cases ('yes', 'no', 'not applicable' are categorical variables) are not evenly distributed and Cohen's Kappa is sensitive to the distribution of the data (Belur et al., 2021). The propensity for reporting methodological quality criteria and experimental studies was low, resulting in many more 'no' answers than 'yes'. Hence, percentage agreement was also calculated. This analysis highlights the need to be able to monitor inter-rater agreement throughout the project using multiple methods.

6.5.3. Mutual Benefits

Efficiency savings are important and methods such as crowd science should be assessed accordingly, however there are also considerations outside of efficiency that will influence whether to instigate a project using crowd science or not. Participation in a crowd science project must be of mutual benefit to both the project lead(s) and the crowd scientists. Crowd scientists report general benefits as the opportunity to gain a more in depth understanding of systematic reviews, and a topic that is of very high interest, and the opportunity to collaborate on a large, international project. For me as the project lead, it was an opportunity to engage the wider pain research community and introduce a new generation of researchers (independent of career stage) to the utility and challenges of preclinical systematic reviews.

6.5.3.1. Training and Education

Individuals reported gains of skills and insight relevant to systematic review processes. They also reported improved understanding of research design and have since improved their own research practises. These findings are corroborated by Menon (2020) which as part of a ZonMw Netherlands 'More Knowledge with Fewer Animals' grant identified that conducting preclinical systematic reviews leads to a direct impact on research practises such as improved planning and reporting including increased understanding of for the 3Rs, improved transparency, and the need for better study quality appraisal. They also report impacts on other research behaviours such as promoting change within teams to improve research, team networking and collaboration.

The training materials consist of general information, training videos for data extraction, and specific contextualised modules on Learn to SyRF. Not all these educational resources are specific to the project and can be re-used e.g., general systematic review training or teaching, or adapted easily to the context of a new preclinical systematic review. They are also openly available and can be used to promote learning within both formal and informal education settings[1] as well as for other systematic reviews. To add, for the reasons discussed above, they provide opportunity to engender good

[1] Materials have been used for teaching critical analysis skills to undergraduate intercalating medical students

research practises as well as critical analysis skills and therefore there is great benefit to participating as a crowd scientist during undergraduate education. In addition, Collins (2014) provides evidence that school students who participated in a citizen science project expressed increased confidence in their ability to contribute to scientific research. Therefore, course designers should consider building such activities into the syllabus.

6.5.3.2. Rewards and incentives

As previously discussed, the motivations for joining the project were not collected. Authorship and acknowledgement were offered. Given that authorship is academic funding and promotion currency and the time and commitment required to meet the authorship threshold, it will undoubtedly have been an important motivation. The contributory acknowledgement is a common model for citizen science projects as it aligns with traditional research practices. Based upon interpretation of the International Committee of Medical Journal Editors (ICMJE) guidelines it was suggested that crowd participants should not qualify for authorship. This is an under valuation, and misrepresentation of the importance and complexity of the task and their commitment. I ensured all crowd scientists met the ICMJE authorship criteria through their contribution, revision of the manuscript, approval of the final version and agreement to be accountable for all aspects of the work. However, the ICMJE guidelines are general and limited in scope, and no longer reflect the way in which research conduct has evolved.

It is widely acknowledged that author teams are getting larger and increasingly interdisciplinary (Cummings and Kiesler, 2014) but how this reflects the division of labour across the team and how it feeds into hiring and promotions for scientists is not clear (Haeussler and Sauermann, 2020). Hence, to address the culture of individualised rewards the authors' contributions were reported in a statement in accordance with the Contributor Roles Taxonomy (CRediT) which represents the roles typically played by contributors to scientific scholarly outputs (Allen et al., 2019, McNutt et al., 2018, Soliman et al., 2021).

Reporting contributorship in accordance with the CRediT taxonomy fundamentally indicates who did what. It also provides data that can be used to deepen understanding of how the organisation of the team relates to the distribution of labour, specialist or generalist skillsets of team participants and the interdependence between activities and individuals. As contribution data become available from more systematic reviews, and a broader range of journals, it will be possible to study the division of labour and the relationships of scientists in a field. It will also provide further opportunity to study the functions of a project e.g., management, software development etc. and more directly related to challenges of preclinical systematic reviews it could provide insight into efficiency arguments for the

use of crowd science and automation technologies with consideration of the broader aims of crowd science such as education and increased uptake of preclinical reviews.

6.5.4. Crowd Science Facilitates Community Engagement, Networking, and Collaborative Working

Crowd science has been conceptualised as a way of performing better science, part of a market driven approach to the production of knowledge (Mirowski, 2018) and as a challenge to the hierarchical conceptions of expertise (Collins, 2014) and who can partake and contribute to research. Although this crowd was comprised of scientists, it still enabled a democratisation of the production and use of knowledge and provides an example as to how this can be extended beyond the immediate pain research community e.g., researchers from other fields, patients, and the public. Here, most crowd participants were from the United States and Northern Europe, which is likely a reflection of my own network. Participation and perspectives from the Global South are often underrepresented (Chan et al., 2020, Onie, 2020) and the few crowd scientists from lower-middle income countries e.g., Iran and Nigeria, demonstrates an opportunity to diversify participation and international scientific networks.

6.5.5. Crowd Science Requires Close Project Leadership and Management

Project lead(s) need to provide the infrastructure and technical support, as well as a nurturing environment to successfully build a virtual research community who are critical to the success of the project. For a large group of participants with diverse backgrounds to successfully contribute requires close project management, a very resource intensive task. During the planning phase of a project, because participants are likely to lack formal systematic review training and will have varied specialist knowledge, the project lead(s) need to consider the training requirements and develop appropriate training material. Similarly, they will need to consider the ethical issues including rewards and incentives and communicate the 'agreement' clearly to all participants before they contribute. Project lead(s) must also have instigated mechanisms to monitor progress and identification of issues for timely resolve. A risk assessment should also be conducted and identified risks mitigated as best as possible. During the project, the lead(s) needs to be flexible, adaptable, and approachable. They require strong inter-personal and communication skills to motivate and lead the crowd particularly pertinent for an isolated, geographically disparate, multinational, and culturally diverse crowd with varied experiences. Throughout they need to be cognisant that the crowd is comprised of volunteers and the project leads role is to facilitate their valued contribution and ensure that the individuals also feel valued.

6.6. Future Directions

A lot of research into citizen science for public engagement has been conducted, there is little for specific communities e.g., scientists with a specific skill set to contribute to a specified complex task. The citizen science platform Zooniverse has provided a test case to investigate the roles of project leaders and citizen scientists. These insights can be applied to the future use of crowd science in more specialist roles. These insights coupled with the findings from this project, provide further ideas to understand organisational structures for future research and knowledge production in the context of systematic reviews. There are several challenges to crowdsourcing for systematic reviews including recruitment, participation and retention and mechanisms for quality control.

For future projects, it would be useful to co-create a set of guiding principles for future crowd science projects. For example, Table 6.8 is based upon combining the principles of citizen science (Robinson L.D. et al., 2018) and the principles of Open and Collaborative Science in Development Network (Open and Collaborative Science in Development Network) and aims to remove barriers and strengthen diversity and equity in the production of knowledge.

Table 6.8. Guiding Principles for Crowd Science.

Principle	Description
Knowledge Generation	Crowd science involves participants in scientific endeavour that generates knowledge. It recognises the need for diverse understandings of knowledge making to co-exist in scientific production.
Inclusive Infrastructure	It incentivises use and development of inclusive infrastructure that empower people of all abilities to make and use accessible open-source technologies
Collaboration	Crowd science is mutually beneficial to the project lead(s) and crowd scientists. Crowd scientists may participate in multiple stages of the scientific process. It thereby advocates for every individual's right to research and enables different forms of participation. The project lead(s) must consider legal and ethical issues.
Open	Project practises are situated in openness. Participants can receive feedback from the project. Data and meta-data are made publicly available where possible, results are published in open access format
Acknowledgement	Crowd scientists are appropriately acknowledged (can include authorship) in project results and publications. Project leads must ensure fair and ethical rewards and incentives.
Research approach	It is a research method with limitations and biases like any other. Projects should be evaluated for their scientific output, participant experience, and wider impacts.

There are two key factors required to anchor crowd science practise in the conduct and production of timely preclinical systematic reviews: infrastructure and community networks and collaborative working, and despite volunteer labour both require funding.

6.6.1. *Infrastructure*

Infrastructure coupled with the design of the project can affect volunteer behaviour and therefore project success. The sustainability of this approach is dependent on our ability to design for volunteer engagement as well as for scientific efficiency.

SyRF was critical: it enabled access to the data (identified studies) and information that constituted both inputs and outputs of the research process. The platform provides flexibility to allow for breadth and diversity of project types. However, there were several SyRF related usability issues e.g., navigation, crashing of the platform, inability to return to previously started studies. These issues have since been resolved but may have compromised the experience for some crowd scientists, and some may have been disinclined to continue. SyRF was designed to answer the challenge of performing large-scale reviews and feedback from the crowd identified ways in which the platform can be improved. Additionally, SyRF has several features that remain under active development which crowd science can assist with e.g. testing of new features including training machines (Bahor et al., 2021). While this approach to working seeks to overcome geographical barriers for scientists from the Global South and those stereotypically considered outside of 'developed' countries, technological barriers such as lack of internet access can persist and therefore continued infrastructure development should also consider how to enable equitable access to technology.

6.6.2. *Community Networking and Collaborative Working*

Crowd science opens the ability to shape research projects in a way that matches research interests, thereby making science more responsive and inclusive. Studies show that researchers, funders and publishers fail to use prior research when preparing to start, fund or publish new studies (Robinson and Goodman, 2011, Clarke, 2004, Chalmers, 2005) and with an exponentially increasing volume of literature and the complexity of the systematic review process and the time that it takes makes preclinical systematic reviews unfeasible to most primary researchers. It is therefore incumbent upon us to challenge traditional ways of working to ensure that the full potential and benefits of systematic reviews can be realised. It would therefore be of benefit to the pain field to maintain an active pool of crowd scientists for contribution to preclinical systematic reviews. The most labour-intensive stages of the systematic review are most likely to be conducted by students and early career researchers. Therefore, a crowd is most likely to be comprised of those career stages and closer collaboration with primary research groups, coupled with the development and incorporation of such projects into formal teaching programmes is one strategy for growing a community crowd.

Future project leads of systematic reviews using crowd science will want to increase participation and achieve high numbers of contributors producing high quality contributions, so there is a requirement

to understand the volunteer experience and their motivations. Analysis by Sauermann and Franzoni (2013) shows that interest can be a powerful motivator of individuals' contributions. Spiers et al. (2019) analysed more than 60 online citizen science projects to study how project design affects volunteer behaviour and project success. They conclude a high level of heterogeneity across the projects and that variables beyond the platform and research domain play an important role in project success. Examples of variables include the difficultly of the task, level of project promotion, quality of researcher engagement and even the inclusion of 'charismatic' features. Quantifying these variables in the context of preclinical systematic reviews may help to increase project participation and success.

Cappa et al. (2018) investigated the feasibility of using reward mechanisms to positively influence participation and motivation. They demonstrated that both monetary and online acknowledgement rewards benefit citizen scientist participation and motivations. It is likely that these findings could be generalised and that rewards are a valid approach for increasing participation, motivations, and referral intention in crowd science projects. Future studies should identify what type of rewards are most valuable to maximise their effect. Concomitantly, many scientists that form the crowd here were gainfully employed and therefore the offer of authorship and public recognition may be a more economically sustainable incentive and reward, especially considering the potentially large number of contributors.

In this project, crowd participants were only involved in the screening and data extraction phases however, future projects could be more collaborative and encompass the complete systematic review process and realise the full intellectual potential of a specialist crowd. If a pool of crowd scientists were to be maintained it would be interesting to assess how the division of labour may evolve as individuals gain experience; it could lead to the identification of 'expert reviewers' who could assist with newcomers, answer questions, provide a quality control function and/or contribute to the reconciliation stage of the review process (Jackson et al., 2015). Although for reviews of this kind, it is necessary that reviewers conduct training specific to the research area. Additionally, as more automation technologies are integrated and utilised, studies will be required to assess, how the division of labour changes, how well inexperienced and experienced humans and machines work together and whether machines bring tangible benefits.

6.6.3. *Future research questions*

- What are different types of contributors and their motivations?
- What are the project characteristics required to ensure high recruitment, retention, and quality participation?
- What type of rewards are most valuable?

- What is the impact on researchers and their research by performing a systematic review? To include in depth analysis of research practises pre and post systematic review experience
- How reliable and feasible to use a crowd for the study selection and data extraction processes in evidence synthesis?
- People vs machine

6.7. Conclusion

The qualitative and quantitative analyses presented here provides evidence that crowd science can provide considerable salary savings, though did not improve efficiency through reduction of speed nor improved accuracy but did lead to professional and research related benefits. The observations made here provide strategies for improving recruitment, retention, participation, quality of contributions and appropriate rewards and incentives. Many project leads may be put off by the uncertainty surrounding a crowd reliability however, preclinical systematic reviews conducted using both crowd science and more traditional kinds of employment relationships are feasible and achievable with little risk to the project and doing so will build upon the lessons learned here. When considering the use of crowd science for systematic reviews and incorporating the observations made here, project leads must be cognisant of the dual aims of crowd science; to perform systematic reviews in the most efficient manner possible (scientific efficiency) and to allow a broad community to engage with the research (social inclusivity); the relationship between these aims is nuanced and may not be aligned. It is likely that many fields of research suffer from siloed effort, models and approaches which may be holding back progress. Crowd science with supportive infrastructure offers the opportunity to bring together researchers from across different fields of pain or more broadly biomedical research, thereby creating potential for cross-cutting, multi-disciplinary collaborative working and more efficient problem solving. More research should be conducted to better understand how to scale this approach and ensure appropriate rewards and incentives for a specialist crowd of scientists.

7. Machine Learning for Study Selection

7.1. Introduction

The search strategy for the Cannabis & Pain review was purposefully designed to be broad to capture all potentially relevant literature. The assessment of cannabis-based medicines, cannabinoids and endocannabinoid system modulators for anti-nociceptive efficacy has been of increasing research interest for over 30 years and hence it was anticipated that the search would return a large number of potentially relevant studies. Screening, the selection of studies in accordance with the inclusion criteria, is a labour-intensive stage in the systematic review process as each article needs to be screened in duplicate with discrepancies reconciled by a third independent reviewer. The use of text-mining and machine learning (ML) to aid systematic reviews is becoming an increasingly effective and popular approach to reduce the manual screening burden and improve efficiency of the systematic review process (Bannach-Brown et al., 2019, Tsafnat et al., 2014, Miwa et al., 2014, O'Mara-Eves et al., 2015). Therefore, an aim was to employ ML to reduce the manual screening burden.

In systematic reviews researchers perform a search of multiple databases using a comprehensive search strategy based upon free text words as well as controlled vocabulary to retrieve potentially relevant studies. Records identified from the search need to be screened and categorised independently against the predetermined inclusion criteria for eligibility. To ensure that all relevant research is identified reviewers can sacrifice elements of the specificity of their search for increased sensitivity, often leading to the retrieval of many thousands of studies. This is particularly challenging within preclinical context due to broad research questions, the complexity and large variation in animal experimentation, less precise or consistent terminology and the lack of controlled vocabularies. As a result, the specificity of sensitive electronic searches of bibliographic databases is low and reviewers often need to sift manually through thousands of irrelevant titles and abstracts in order to identify the small number of relevant ones (Lefebvre et al., 2013).

7.1.1. Screening

Following the search, the reviewer downloads a file with the records that require screening. It contains the titles and abstracts and other meta-data including author, journal, publication year and DOI. This file is uploaded to the systematic review platform and reviewers are presented an article at random. The reviewer reads the title and abstract and compares against the inclusion/exclusion criteria to decide. They are then required to click on either 'include' or 'exclude' and then click 'next' to be presented with the next article.

Screening best practice is considered manual screening conducted by two independent reviewers; the rationale that a single reviewer can introduce bias to the study selection either because of their interpretation of the inclusion criteria or through their understanding of the content of titles and abstracts. The discrepancies will be reconciled by a third independent reviewer thereby reducing the risk of bias and ensuring that all relevant studies are included (Higgins et al., 2016). An experienced reviewer can take between 30 seconds and several minutes to evaluate a citation (Wallace et al., 2010). CAMARADES' experience estimates that a systematic review of 10000 retrieved studies will take a minimum of 40 weeks to screen (Bannach-Brown et al., 2019). The need for 2 reviewers more than doubles the length of time that it would take for each reference to be screened by only one person. Hence, the screening stage is one of the most labour intensive and time-consuming tasks of the review process (Bastian et al., 2010).

Text mining offers many possibilities to assist in the identification of relevant research. Text mining is defined as the process of discovering knowledge and structure from unstructured data (i.e., text) and present it in a concise form (Ananiadou et al., 2006). Text mining technologies can facilitate the identification of relevant literature, its categorisation and summarisation. Greater efficiency in the systematic review process is most likely to be found in the identification of relevant studies rather than in their quality assessment and synthesis as such, the use of text mining to date has been focused on the early stages of the systematic review process, the identification of relevant studies and the initial descriptive stage. Text mining comprises of three activities; information retrieval, information extraction and data mining, activities that map well into the processes of systematic reviews; the retrieval of relevant studies, describing the study characteristics, identifying patterns within the studies.

7.1.2. Machine learning for Study Selection

Machine learning is the application of algorithms and mathematical models that provide the computer system the ability to learn and improve from experience without being explicitly programmed (Samuel, 1959). ML algorithms can be employed to learn how to categorise studies and any systematic review can take advantage of generic classifiers and review specific classifiers. Both generic and review specific classifiers are developed by humans and machines operating in mutually beneficial ways. There are techniques for using ML in new reviews where humans and computers operate in a workflow which prioritises the relevant studies for manual screening while simultaneously improving the accuracy of the machine model (Wallace et al., 2010). Humans categorise a sub-set of the records returned by the search which are then used to train the machine.

It is possible to build a model that directly predicts the probability of the study's eligibility and relevance to the review. However, lack of training data can be a limitation, and so the performance

of the model needs to be assessed and if the model does not meet the required threshold, then more records are screened and added to the training dataset. In addition, it is difficult to assess reliability because the machine will not have had the benefit of 'seeing' the full variety of relevant studies. Generic classifiers are highly accurate with respect to the task they have been trained to perform but they will not be able to classify studies in a way that directly aligns with the specific inclusion criteria defined by individual reviews. By contrast, review specific classifiers will tend to be highly attuned to the context of the criteria for review, but may be less accurate due to the extensiveness of training material (Thomas et al., 2017).

A systematic review of the use of text mining for study identification in systematic reviews conducted by O'Mara-Eves et al. (2015) presents the available research evidence on existing methods for text mining related to the screening of titles and abstracts, including the performance metrics used to evaluate these technologies, although the field is likely to have developed considerably since this review was published. There were 44 included studies and over the duration of 8 years, almost every year saw the evaluation of newly applied type of classifier or a new consideration of the application of text mining to screening. Most studies presented a new approach that distinguished it from previous studies therefore there are very few replications or comparisons between the studies. This suggests a rapidly evolving evidence base, that is both active and diverse and it is difficult to form conclusions about which approaches work best. The lack of replication between studies or collaboration between research teams, make it difficult to establish any overall conclusions about best approaches and is particularly limited in the context of preclinical systematic reviews.

There are several ways in which text mining and ML can reduce the workload, increasing the speed and efficiency of the screening process (summarised in Table 7.1):

7.1.2.1. Reducing the number of studies that need to be screened manually

Within the systematic review conducted by O'Mara-Eves et al. (2015) most studies (30/44) propose using text mining for the purpose of reducing the number of studies that need to be screened manually. There are two approaches to automatically apply eligibility criteria which aims to improve efficiency by automatically identifying which studies should be included.

1. The first approach, an automatic classifier, can make explicit include/exclude decisions by learning from an initial sample of decisions. This approach assumes that false positives do not matter as these will be identified later in the review process.
2. The second approach, screening prioritisation, can be used to support reviews that are required within a short timeframe. Following a search, the identified studies will

be prioritised by order of relevance against the inclusion criteria and studies that fall below a threshold, or that lie within a 'negative prediction zone' are excluded.

The prioritisation of relevant studies will not reduce the workload but will reduce the time from review commencement to completion. This method may bias the reviewer; they are likely to expect more studies at the beginning of the process to be included and be over-inclusive and likewise miss studies later in the list on the assumption that these studies are less relevant. Additionally, in the context of living systematic reviews this approach offers a method for identifying when likely relevant new studies are published and a review update is required (Cohen et al., 2012).

Both methods employ active learning. Active learning is an iterative process whereby the accuracy of the predictions made by machine is improved through interaction with reviewers. The machine 'learns' from an initial sample and subsequently generates a ranked list and requests the reviewer to provide decisions on items high up the list that the machine will learn the most from. The machine adapts its decision rule including the information from the additional items and generates a new list of items for the reviewer to screen which continues in an iterative manner. For example, for new reviews, Wallace et al. (2010) developed an 'active learning' system which uses automatic classification to identify included studies. They found that it was not necessary to manually screen more than half of the studies before the classifier was able to identify the remaining relevant studies; an estimated reduction of 40-50 % of irrelevant studies that needed to be manually screened (Wallace et al., 2010). Cohen et al. (2006) developed a system for updating a review that uses the included studies as training data and interested in identifying not only potentially relevant research but also research of high quality, therefore prioritised the ability of their system to recall as many relevant studies as possible, even if this resulted in some 'noise' for manual screening. They found that this reduced the time needed to identify studies significantly for all review updates by as much as 50 %.

7.1.2.2. Reducing the number of people needed to screen items

Several studies have advocated the use of text mining as a second screener: replacing or supplementing the additional human reviewer (e.g., (Bekhuis and Demner-Fushman, 2012, Bekhuis and Demner-Fushman, 2010, Frunza et al., 2011, Adeva et al., 2014). In this model, a human reviewer screens all records and the machine acts as an independent check. Results suggest the classifiers had good agreement with the human reviewers with a reported reduction of manual workload between 88 % and 98 % (Bekhuis et al., 2014). This approach is particularly appealing as it simultaneously reduces the resource required to screen studies and 100 % of screening is conducted by human reviewers therefore nullifying concerns about false negatives assigned by the machine. A further potential benefit of this approach is that this could deliver quality assurance by confirming concordant decisions and by identifying studies associated with discordant decisions for further consideration

(Bekhuis and Demner-Fushman, 2012). A limitation, however, is that this approach assumes that any mistakes made by the human screener are random and not because of a misapplication of the inclusion criteria, which would potentially be identified if two reviewers were screening concurrently.

7.1.2.3. Increasing the rate of screening

Another approach is to use visual data mining (VDM) to create a visual representation of the connections between records to assist the screener in identifying studies that are more likely to be like each other. The approach assumes that humans can decide a studies relevance faster when using visual information compared to using the textual information from the titles and abstracts alone (Felizardo et al., 2011). A second approach to increasing the rate of screening is through efficient citation assignment. Using active learning, Wallace et al. (2010) developed an algorithm that incorporates both information about the relevance of each study and the expected time that it will take to annotate; on that basis the algorithm selects citations specifically for expert and novice reviewers to label. This approach was reported to enable more records to be screened in the same amount of time compared with typical active learning approaches (Wallace et al., 2010).

Table 7.1. Summary of ML methods that can be employed to reduce the screening burden.

Approach	**Aim**	**Strengths**	**Limitations**
Automatic Classifier Include/Exclude decisions	Reduce the number of studies needed to be screened manually	Learns from training data and therefore decisions can be checked Provides opportunity to ensure that all relevant studies have been included	Requires initial sample of decisions to be 'trained' Will perform with greatest accuracy as generic classifier but accuracy will reduce with more complex classification requirements
Screening Prioritisation Prioritisation in order of relevance against the inclusion criteria Identifies most relevant studies first so that the remaining stages of the review process can be conducted concurrent to the less relevant studies being screened	Reduce the number of studies needed to be screened manually	Supports reviews that are required within a short timeframe Improves efficiency In the context of a LSR this offers a method for identifying when likely relevant studies are published and a review update required	Offers the option to establish a cut off to screening/'negative prediction zone' however this could compromise the principle of including all relevant research Unable to quantify the likely number of studies that have been missed Bias the reviewer
Second Screener (text mining – automatic classification)	Reduce the number of people needed to screen items	Reduces the resource required to manual screen Provides quality assurance by confirming concordant decisions and identifying studies associated with	Assumption is that mistakes made by the human screener are random and not due to misapplication of the inclusion criteria

		discordant decisions for further consideration	
Visual Data Mining (VDM) Creates visual representation of the connections between documents to assist the screener in identifying studies that are more likely to be like each other	Increasing the rate of screening	Once a relevant document has been identified the screener will be able to quickly determine if the other similar documents are relevant or not	Has only been employed in small studies so it is unclear how well it can perform if upscaled Assumes humans can decide a studies relevance more quickly based upon visual information compared to textual
Citation assignment Incorporates study relevance and the expected time to annotate thereby selecting studies specifically for expert or novice reviewers	Increasing the rate of screening	More items screened in a period Utility within large reviewer teams and when the level of expertise is known	

7.2. Aim

An aim of the Cannabis & Pain review was to employ ML to reduce the manual screening burden and add to the evidence-base of contextualised use of ML approaches in preclinical systematic reviews. The aim was for the machine to learn the characteristics of included and excluded studies, so a model that could predict whether a given record is more relevant or irrelevant was built. This prioritisation, in order of relevance would offer the opportunity to either increase the speed of screening of all retrieved records given that the records that reviewers are presented early in the process are most likely to be included so decisions can be made more quickly, or to reduce the number of studies that needed to be screened manually by establishing a cut off to screening.

7.3. Methods

The systematic search identified 10,816 unique records from three electronic databases; Embase, PubMed and Web of Science, which required screening against the inclusion/exclusion criteria (Soliman et al., 2021). Screening was performed by the crowd of scientists.

7.3.1. The Tool

The web-based tool EPPI-Reviewer (Thomas et al., 2010), developed at University College London was used. It had previously performed well in a preclinical systematic review context (Bannach-Brown et al., 2019). It facilitates study selection through text-mining and ML through the employment of 'active learning' in which active feedback (new include/exclude decisions) from reviewers improves the predictive accuracy. Active learning is an iterative process which involves using a small number of manually screened records from which the machine learns from the include/exclude decisions and generates a list of records for the reviewer to look at next; bringing the records that are most likely

relevant to the beginning of the screening process and pushing the records that are least likely to be relevant to the end (Miwa et al., 2014). The aim to reduce the manual effort by focusing the effort on the most relevant records, the aim to identify all relevant records as quickly as possible. There are two phases to making the machine learn from the include/exclude decisions as made by reviewers on a subset of records. Firstly, is 'feature selection' which is the transformation of text into a numerical format which can be used in a ML algorithm. The tool uses a 'tri-gram' 'bag of words' which involves listing every word, pair of words and triplet of words within each record. All records (title and abstracts) are then indexed against this list of words by counting the number of times a word occurs and uses 'term frequency – inverse document frequency' to capture more information about how well the word or word phrase distinguishes one record from another. Next, a statistical model is built, which aims to distinguish records relevant to the review from those that are not. A linear support vector machine algorithm is used as implemented in the Scikit-Learn Python Machine Learning library (Pedregosa et al., 2011). Once the model is built, all unscreened records are the scored according to the model. The records are then presented for screening in order of relevance, most relevant at the top of the list.

Although not formally investigated, based on their extensive experience of preclinical systematic reviews CAMARADES estimate two independent human reviewers to achieve 95 % sensitivity (Bannach-Brown et al., 2019). Therefore, the ML algorithm would need to perform to the same, if not better standard, to have confidence that relevant studies are identified. Once the level of sensitivity is achieved, the next priority is to improve specificity to the same 95 % threshold to limit over-sampling.

7.3.2. Machine Learning for Study Selection; Training and Validation

Using the RAND function in Excel I first selected 2000 records at random to form the first training dataset. The number of records was arbitrary as it was not possible to predict how many training instances were required for the machine to perform with the required sensitivity and specificity. These records were then screened by two independent crowd reviewers with discrepancies reconciled by a third independent crowd reviewer. The manual screening was conducted on the platform SyRF. Reviewers were randomly presented an article, with the title and abstract displayed. The reviewer decided against the inclusion/exclusion as to whether to include (1) or exclude (0) the article. A second reviewer was also presented the article to decide. If the article was given two 'include' or two 'exclude' decisions i.e., the two reviewers agreed on the classification of the record, then the screening of that record was complete. However, if the reviewers disagreed, then the record was screened by a third reviewer. Blinding was maintained, no reviewer was aware as to whether they were reviewer one, two or when applicable, three, so their decision making process could not be influenced.

The ML algorithm was then assessed. The training set consisted of 2000 records and the unseen validation set 8,816. The prevalence of included studies was low at 3 %, 60 included studies. The machine performed with 100 % sensitivity and 0 % specificity. High sensitivity has a low risk of missing relevant studies however, the low specificity is more likely to identify studies for inclusion (i.e., false positives). I then repeated this process, adding a further 2000 randomly selected and manually screened records and then assessed machine performance. In this instance, the training set consisted of 4000 records and the unseen validation set 6,816. The prevalence remained at 3 % and performance at 100 % sensitivity and 0 % specificity. This suggested the dataset was not balanced or could be biased preventing the machine from learning accurately. At this stage two options were available to me; firstly, to continue in this iterative manner, adding 2000 manually screened records or allow the crowd to manually screen the full set of returned records and then use ML for error analysis. Using a crowd meant that the burden was being spread across many reviewers and so I decided to keep them gainfully employed and engaged with the project by manually screening all the records.

7.3.3. Machine Learning for Error Analysis

In the absence of being able to use ML for screening prioritisation, I used ML to perform error analysis. All records that had been manually screened by the crowd of reviewers and were used to train the machine. The machine performs a linear classification and is therefore able to identify differences between the human decisions and present a list of studies that may be relevant for consideration by an expert reviewer (myself in this instance). I performed this in an iterative manner, manually screening the top 200 potentially relevant studies (an arbitrary criteria) before feedback to the machine, until there was no longer a change in the include/exclude decisions (Figure 7.1).

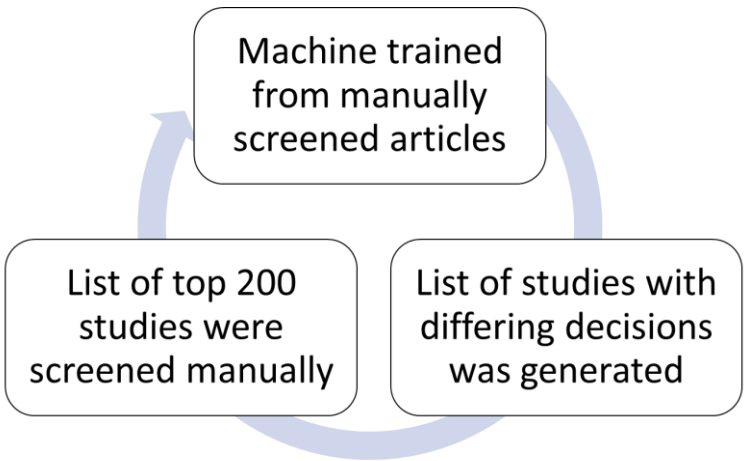


Figure 7.1 Schematic of the use of ML in an iterative manner for error analysis. Each step re-run until the stopping criteria was met (no more changes in screening decisions).

7.4. Results

The systematic search returned 10,816 records that were potentially relevant to the review that required screening against the inclusion/exclusion criteria. It was not possible to use prioritised screening and all records were screened manually by 36 crowd scientists. The inter-rater agreement was 90 %. All records were screened in duplicate and disagreements reconciled by a third independent reviewer.

7.4.1. Machine Learning could not be used for prioritised screening

After the first 2000 records had been screened against the inclusion/exclusion criteria by crowd members an attempt was made to use this as the training dataset. The prevalence of included studies was low, 3 %. The machine performed with 100 % sensitivity and 0 % specificity, so screening continued manually. Following the manual screening of 4000 records an attempt to train the machine was made again. The prevalence of included studies remained at 3 % and the machine performed with 100 % sensitivity and 0 % specificity. This suggested an unbalanced dataset due to the low prevalence which would prevent the machine from being able to accurately rank the records according to likely relevance. Therefore, the remaining ~6000 records were screened manually. Overall, the crowd identified 850 studies for inclusion. I performed an error analysis of the included studies manually and 228 (26.8 % of the included studies) had been wrongly included. The wrongly included studies all tested an intervention for anti-nociceptive efficacy in healthy animals not in pathological or injury-related persistent pain.

7.4.2. Machine Learning for Error Analysis

An aim of every systematic review is to identify all the possibly relevant studies. The human decisions therefore were used to train the machine for error analysis; offering the opportunity to ensure that studies had not been wrongly excluded. The manual error analysis (which identified 228 wrongly included studies) would facilitate the accuracy of the machine's performance. The human screening decisions of the 10,816 records were then used to train the ML algorithm. The machine ranked and presented the studies where there was a difference between human and machine decisions (Table 7.2). This allowed me to reassess records where the machine predictive score was divergent from the human decisions thereby identifying potential human errors.

Table 7.2. The difference between the number of included and excluded studies between human and machine decisions. 0; Exclude, 1; Include. Green denotes agreement, red disagreement.

Run	Exclude / Include	0	1
1	0	9141	6
	1	996	616
2	0	9425	7
	1	631	696
3	0	9481	7
	1	552	719
4	0	9480	7
	1	546	726
5	0	9450	7
	1	566	736
6	0	9465	7
	1	543	744

In the first iteration the machine performed with 95 % sensitivity and 89 % specificity which improved with each iteration to an eventual sensitivity of 95 % and specificity of 94 %. The performance of each re-trained algorithm compared to the previous is shown in Table 7.3. During the error analysis process a further 1200 studies were presented for screening. A total of 137 decisions changed. This represents a screening error of 1 %. The changed decisions led to an eventual inclusion of 751 studies (Table 7.3).

Following the third run, the machine performed with a 95 % sensitivity and 94 % specificity. This meant the machine was performing well and I could increase the sensitivity to 99.9 % which would in turn result in an 88 % specificity and the inclusion of the remaining records 'questioned' by the difference between machine and human decisions. However, I chose to continue with error analysis in an iterative manner because an additional 559 for full text screening and potentially incorrect data extraction due to using a crowd would potentially result in increased inefficiencies.

Table 7.3. ML algorithm performance for iterative error analysis. The number of differences represents the number of studies for which the human and machine decisions differ in either direction. The next two columns represent the number of studies that were changed from exclude to include and include to exclude. The final two columns represent the total number of records excluded and included after each round.

Machine Run	Accuracy (%)	Sensitivity (%)	95 % CI	Specificity (%)	Number of Differences	Exclude -> Include	Include -> Exclude	Total Changed Decisions	Exclude	Include
1	90	95	0.8813 - 0.9076	89	1002	82	1	83	10137	622
2	93	95	0.9212 - 0.9428	93	703	26	3	29	10056	703
3	94	95	0.9287 - 0.9493	94	559	7	0	7	10033	726
4	94	95	0.9302 - 0.9506	94	553	10	0	10	10026	733
5	94	95	0.9237 - 0.9450	94	573	8	0	8	10016	743
6	94	95	0.9272 - 0.9480	94	550	0	0	0	10008	751

7.5. Discussion

In the Cannabis & Pain review, ML was not able to reduce the screening burden however, the potential to do so is evident. The ML algorithm had high level of performance for determining the relevance of records that described the assessment of cannabinoids, cannabis-based medicines, and endocannabinoid system modulators in animal models of injury related or pathological persistent pain with 95 % sensitivity and 94 % specificity. The use of the model for error analysis proved utility in ensuring that all the relevant literature was captured, and studies not incorrectly excluded.

7.5.1. Error Analysis

Using the ML algorithm to classify the likelihood of an article being relevant to the review highlighted discrepancies between human and machine decisions. Identification and correction of human errors then updated the training set and improved the sensitivity of the algorithm.

The manual screening of all records returned by the systematic search was performed by a crowd of scientists. The algorithm facilitated the identification of screening errors in 1 % of the dataset. The decisions were made by 'majority vote', agreement of two reviewers and if there was disagreement the 'third' reviewer had the deciding vote, therefore each record had been dual screened by two independent reviewers and discrepancies reconciled by a third independent reviewer. In terms of the performance of crowd scientists, this was the same error rate as reported by Bannach-Brown et al. (2019) in the screening of records for a systematic review of animal experiments modelling depression. The are several reasons for the misclassification of records including the reviewers' understanding and interpretation of the review's inclusion/exclusion criteria; the complex, inconsistent and ambiguous way animal studies are reported or due to fatigue or distraction simply leading to the clicking of the wrong button on SyRF. The crowd members were trained and encouraged to reach out if they did not know how to apply the inclusion/exclusion criteria, but they were also encouraged to be over inclusive if in doubt. The approach assumed that false positives were not critical because they would be identified later in the review process.

The error analysis and subsequent corrections led to the improvement in the performance of the algorithm. The consequences of over sampling are evident from the error analysis process and if I had not performed a manual error analysis prior to building the model, the accuracy of its performance would have been further impaired. Specificity improved from 83 to 94 % without compromising sensitivity which, if reliant on ML to reduce the manual screening workload would have substantially reduced over-sampling of over 1000 records. This improvement of the algorithm following the iterative error correction demonstrates the impact that such errors can have on the learning of the machine.

This error analysis was conducted in the context of needing to complete the Cannabis & Pain review in a timely manner and therefore I had set a pragmatic cut off; to stop once I was no longer changing decisions. However, there were still differences between human and machine decisions and therefore it is likely that there are further errors in the dataset. The dataset could be updated with the full-text screening decisions, the exclusion of 278 records, further training the algorithm. A more in-depth analysis of all instances where the human and machine decisions differ may also lead to the identification of further errors and therefore upon correction improve the performance of the machine. This would be useful if the review were to be updated or a living systematic review instigated (Elliott et al., 2017). Optimisation of the algorithm will lead to further reduction in the human resource required for screening and subsequent stages of the systematic review process.

7.5.2. Considerations for Use

There are several ways in which a ML algorithm can be employed to reduce the screening burden in a preclinical systematic review; prioritisation for human screening, screening by machine either as a second reviewer or by truncation of manual screening, or error analysis. Firstly, it requires the manual screening of a sub-set of records to form a training set for the machine to learn and a model to be built. Depending on the programme of works there is the option for human and machine to work together, as human decisions are made these can be fed back to improve the performance of the algorithm.

As demonstrated in the error analysis, there are factors that can potentially reduce the predictive performance of the algorithm. Prior to use, all duplicates need to be removed from the dataset and consideration given to inclusion/exclusion criterion that may be inconsistent across relevant studies e.g., journal, year, country or sample size which may confuse the algorithm (Tsou et al., 2020). It is also important that human errors and biases are minimised so that these are not reflected in the training dataset and propagated into the ML algorithm. This can be mitigated by the training data used for ML should be based upon randomly, dual screened records with disagreements reconciled by a third independent reviewer thereby mitigating the risk of human error and ensuring that the dataset is balanced; there will not be an over representation of records that are at greater risk of misclassification.

7.5.2.1. Estimating Baseline Inclusion Rate

Methods for using ML for screening in systematic reviews are still being developed especially in the context of preclinical systematic reviews. Shemilt et al. (2014) highlight the utility of estimating the baseline inclusion rate; an estimate of the proportion of provisionally relevant records expected to be present within the full citation set. This will ensure an unbiased sample of training dataset and in

having an estimate for the likely number of included studies provides opportunity for monitoring the performance of the machine and screening progress and will assist decision making with regards to switching between manual screening and ML. It is calculated by performing manual screening on a subset of records and then performing a sample size calculation (Shemilt et al., 2014).

7.5.2.2. Humans versus Machine

High recall (the ML algorithm returning most of the relevant studies) is considered critical to systematic reviews for which a principle is to retrieve all relevant studies to avoid biasing the review findings. The importance of high recall of relevant studies is likely to be critical in acceptability and uptake of text mining techniques within the systematic review community. A potential drawback to these approaches is that a proportion of studies are likely to be incorrectly included/excluded. CAMARADES estimate that inter-screener agreement to be between 95 and 99 %. In the Cannabis & Pain review, the inter-rater agreement for screening was 90 %. The disagreements were reconciled by a third independent reviewer but such discrepancies may impact the optimisation of the algorithm. Liao et al. (2018) and Bannach-Brown et al. (2018) demonstrated that an algorithm can perform as well as human screeners and similarly to the Cannabis & Pain review has potential to identify human error. Key to having confidence in these techniques is how a ML algorithm is employed by systematic reviewers.

The use of ML to identify and rank the most relevant studies first offers several opportunities to reduce the screening burden. In a study comparing the predictive performance of two ML algorithms, EPPI-Reviewer was found to provide potential reductions in screening burden between 9 and 60 % depending upon how it was employed. Performance was also related to the context of the review and the inclusion/exclusion criteria (Tsou et al., 2020). Similarly, (Bannach-Brown et al., 2019) report an approximate 74 % reduction in screening, with the records requiring manual screening reducing from 70,000 to 18,500.

Reviewers may decide to continue screening manually with the prioritised list of records affording time savings because reviewers will know how likely an article is to be relevant.

A ML algorithm may also offer considerable efficiency savings if it fulfils the role of the 'second' reviewer. To mitigate the risk of human error and bias being promulgated to the algorithm the training dataset should be based upon dual screening and reconciliation of discrepancies by a third independent reviewer. In this review a crowd of reviewers performed the role of the first, second and third reviewer. Simulations could be performed with the ML algorithm to determine the accuracy of single or dual screening and provide an evidence base for the instances in which dual screening is not as critical thereby reducing resource requirements. Conversely, it may also provide insight into how

best to employ a 'crowd' and determine whether higher decision thresholds are required to identify all relevant literature E.g., for the Cannabis & Pain review double screening was employed. If there was disagreement between two reviewers, the decision was finalised by a third independent reviewer. Given the complexity of the review and the varied expertise within the crowd, this may still have led to false inclusion or exclusion decisions which could be avoided if further agreement in decisions were required. Such methods could lead to a more accurate development and optimisation of the algorithm.

Lastly, it offers the option to establish a cut off to screening i.e., records that fall into the "negative prediction zone" are not screened and excluded. A risk is the unknown impact of 'missed' relevant studies and deciding when to stop manual screening remains an open research question. In the assessment conducted by Tsou et al. (2020) the authors suggest a conservative threshold of being able to identify all relevant records after manually screening 70 – 80 % of citations using ML, although their results also suggest that after manually screening 50 % of citations reviewers would still identify 95 % of included studies. Several teams report truncation with methods to ensure high recall while making considerable resource savings e.g., vote counting (multiple classifiers run simultaneously, and then a 'vote' is taken on each item to determine whether it is likely to be relevant or not) led to an estimated reduction in screening effort between 40 and 50 % (Wallace et al., 2010); specialist algorithms that adjust the learning rate to penalise false negatives e.g. Cohen et al. (2006) resulted in a reduction in manual screening of 50 %, and with human input, the algorithm continues to learn as more items are manually screened and so the decision rule is adaptable and less reliant on the initial training set. Bannach-Brown et al. (2019) determined cut off points based upon the decisions on the validation set to achieve the desired performance, based on true scores, an option that would not be available in practise. They recommend using the threshold identified in the validation set and apply it to the remaining 'unknown' records but reaching the necessary sensitivity is not guaranteed.

In summary, ML approaches have been evaluated in the context of systematic reviews of several medical problems (Cohen et al., 2012, Cohen et al., 2006, Howard et al., 2016a). However, in the preclinical domain there have only been a few attempts to apply such techniques (Bannach-Brown et al., 2018, Howard et al., 2016b). Nevertheless, these approaches offer several advantages including being able to search more sensitively than is currently the norm and screen the same number of studies therefore the effort spent screening manually is focused on those studies identified as being the most relevant retrieved by the search. They also offer the opportunity for the sensitivity of the search to remain the same but for screening to be conducted with reduced resources.

7.5.3. Limitations

The performance of the machine will vary significantly depending on the composition of the training dataset. If studies in the training dataset are not representative, then the predictive algorithm may be skewed, affecting its performance. The use of 2000 then 4000 records, although arbitrary, aimed to address this limitation. The inability to train the machine following manual screening of a sub-set of studies was likely due to the low prevalence of inclusion. This can be indication of an over inclusive search string. However, in this context it was particularly challenging to narrow the search terms further, not only due to wanting to identify all relevant literature but because it was difficult to identify the population due to the challenge of differentiating studies in which an intervention(s) was tested for antinociception in healthy animals, animal models of acute pain or animal models of injury-related or pathological persistent pain (the desired population).

When considering the use of ML in the context of improving the feasibility and efficiency of systematic reviews there are several considerations that may influence wider uptake despite the potential for significant resource saving by reducing the screening burden. Firstly, it is not possible to know in advance whether the tool will perform and meet the required threshold. It was anticipated in the Cannabis & Pain review that there may not be enough instances to train the machine (high volume of studies and complex categorisations are complicating factors) therefore a further random sample of studies may have been required before optimal sensitivity was reached. This led to the eventual requirement of full manual screening, a possibility reviewers should be prepared for.

The machine algorithm now has good predictive performance if an update to the search and review was conducted, however the potential may not translate into savings in a different review. In addition, to use the EPPI-Reviewer tool I had to download the screening decisions from SyRF and upload to EPPI-Reviewer. SyRF would benefit from the integration of a similar tool to facilitate ease of use. This relates to the concerns about how best to employ ML without specific ML expertise that many primary researchers and systematic reviewers are likely to share. It is evident that ML for prioritisation, error analysis and for review updates is a safe choice and will ensure that relevant studies will not be missed. Further research and continued evaluations are required to determine the accuracy and impact of using such tools as second reviewer or for truncation. This research will provide developers and reviewers with guidance and how to assess validity for different user cases thereby building trust and increasing uptake of automation technologies to improve the feasibility, accuracy, and efficiency of preclinical systematic reviews (O'Connor et al., 2019).

7.6. Future Research Questions

- Further evaluation of ML algorithms using validated datasets

- Evaluation of the impact of losing relevant studies that are not correctly classified by the algorithm particularly pertinent if applying a cut off to manual screening
- Comparison of search sensitivity and specificity and ML algorithms
- Evaluation of methods used in conjunction with truncation to mitigate the risk of missing studies
- Evaluation of using crowd science in combination with ML

7.7. Conclusion

The use of ML for error analysis confirms the performance and application of ML algorithms for screening in preclinical systematic reviews. Not being able to train the machine earlier in the review process is likely a specific limitation of the Cannabis & Pain review, however, using ML for error analysis highlights the potential to identify human errors. Machine learning performance can be review specific and it is not clear which technologies and strategies work best and therefore there is a need for further comparison trials and systematic reviews of the available automated techniques. Although the evidence suggests that these can offer opportunities to make searches more sensitive or maintain the sensitivity but reduce the resource required for screening and they have a definite safe role for screening prioritisation and error analysis, there remain technological and adoption barriers to using automation methods in systematic reviews. Further research is required to determine their validity and identify different ways in which they can be used. The development and incorporation of machine-assisted tools in open-source, online systematic review software like SyRF would be beneficial and necessary to facilitate the production of accurate and timely evidence synthesis, reduce research waste, and improve decision making concurrent to contributing to the future functionality and success of living systematic reviews. These powerful tools will assist in making sense of an exponentially growing body of data which will ultimately, make what was previously thought of as unattainable, attainable.

8. Graph2Data: semi-automation of data extraction from graphs

8.1. Introduction

Academic publishing rates are accelerating and in preclinical research figures often provide much needed numerical data for meta-analysis. Current manual or semi-automated methods for extracting data from figures can compromise accuracy and are time consuming. Given the volume of research, the complexity, number of figures within a publication and the risk of inaccuracy with current extraction methods, an automated or semi-automated system capable of extracting data from graphs is needed.

8.1.1. Why a Meta-Analysis?

Where appropriate, a meta-analysis can be conducted whereby the quantitative data is extracted from the relevant studies and statistical techniques are used to provide objective sense of the meta-data. A meta-analysis is a statistical method for combining data from multiple studies to provide a quantitative summary of the research findings and assess the strength of evidence present. By combining the data from individual studies through a meta-analysis the power and precision of the results can be increased (Noordzij et al., 2009). Animal studies are generally small, with cohorts typically having a sample size of 4-10 animals per group. To add, an intervention can be assessed in slightly different ways across many independent laboratories; different model types, different intervention types, doses, route of administration, dosing regime, and different outcome assessments with variation in how the assessment is performed and reported. Erdogan et al. (2020) present examples of how easily unit of measure and scaling of the y-axis can affect our perception of results and can result in misleading impressions about the outcomes of a study. Extracting data and performing a meta-analysis removes this perception bias created by inadequate or misleading graphical depiction of data.

Animal studies are homogenous, with a lot of emphasis placed on minimising variance. This contrasts with clinic trials which have larger, more heterogenous samples in single studies conducted across multiple centres. The differences between individual animal studies have implications for the conduct, analysis, and interpretation of meta-analysis of data from preclinical studies. It is therefore a requirement of meta-analytical techniques that the data of the many individual studies of varied experimental designs are combined to determine how the effect varies between studies and identify the variation (Vesterinen et al., 2014, Macleod et al., 2020).

8.1.2. Extracting Data from Graphs for Meta-Analyses

Preclinical researchers typically present numerical data, very often aggregate or summary data, in graphs for visual analysis. However, this increases the complexity for meta-analysis because the raw data (aggregate or individual animal data) is rarely available and if presented in a table or in the text

it is not in a machine-readable format. The data within graphs are not always presented in a clear format e.g., overlapping data points or the variance encompassed within the point for the average, therefore extracting preclinical outcome data is challenging. To add, in contrast to the relative homogeneity of clinical trials preclinical studies can present relevant outcome data in different forms, whether that is by providing multiple measures of the same outcome, measures of multiple timepoints, or in multiple statistical forms. A further complication is the presentation of the data within different studies using different types of graphical plots, tables, and text. Hence the data extraction process can be very time consuming and prone to error (Buscemi et al., 2006, Mathes et al., 2017). For a meta-analysis outcome data need to be extracted in a systematic, accurate, efficient, and standardised way. To ensure reliability and reduce the risk of bias, data are usually extracted by 2 reviewers, with a third to adjudicate discrepancies.

For data only presented in graphs the reviewers must manually measure values as accurately as they can to extract the results. Skill and attention are required to minimise bias and error in the selection and extraction of the data for inclusion in the analysis. Transcription errors are a common problem, with some errors not being detected until after the publishing of a systematic review. For example, a case study comparing 3 systematic reviews reported the rate of data extraction errors ranged between 8 % and 42 % (Carroll et al., 2013). This is problematic because reliability is an important factor for a meta-analysis. Data extraction must be completed with a high degree of accuracy and there should be close agreement between 2 reviewers, to give confidence in the reliability. It is vital that the data extracted are reliable as errors made at this stage of the review can affect estimates and review conclusions (Mathes et al., 2017). There is clear potential to improve both the speed and accuracy of extraction of outcome data. To assist researchers in extracting data accurately, automation technologies have been utilised in the development of various software programmes e.g., Ungraph (Biosoft, 2021), GetData(GetData, 2013). Some of these programmes have individual user associated costs and there remains the risk of transcription errors. To mitigate the risk of transcription errors and to improve reliability it is necessary for the automated software for data extraction to be integrated into the review platform and workflow.

Graph2Data is a tool developed by EPPI-Centre and CAMARADES in collaboration on the SLIM grant (MR/N015665/1). It streamlines and semi-automates axes identification and data point identification, reducing the time required for data extraction, as well as reducing extraction and transcription errors associated with manual methods. A non-inferiority trial demonstrated that this machine-assisted approach represents a considerable time saving and is also more accurate

than current methods; for 10 participants to extract data from 23 graphs the mean time overall was 5 min and 52 s less and 29 % more accurate using the tools compared with current methods. In addition, it was not inferior to users' preferred current methods for data extraction (Cramond et al., 2019). The non-inferiority to current tools and preference of users to existing tools demonstrates that Graph2Data is worth developing and disseminating.

The development and incorporation of machine-assisted tools in open-source, online systematic software such as SyRF is necessary to facilitate the production of accurate and timely evidence synthesis and will contribute to the future functionality and success of living systematic reviews. However, as with all new methods, there is a requirement to conduct further studies to ascertain the impact on efficiency and data accuracy and to assist the systematic review community to determine which automation tools fits their data extraction needs best.

8.2. Aim

The Cramond et al. (2019) study was small. It was therefore necessary to expand, validate, refine, and operationalise Graph2Data. The integration of Graph2Data into SyRF required the adaptation of the platform's systematic review workflow. The aims were as follows:

- Update and enhance Graph2Data in preparation for integration into SyRF
- Plan and conduct a randomised controlled trial to assess whether Graph2Data reduces the time taken for data extraction and improves agreement between reviewers as a proxy for accuracy and reliability in the context of a preclinical systematic review

Throughout I have collaborated closely with the CAMARADES software engineer, Chris Sena, and two consultant developers, Lee Doran-Constant and Laurie Young. Throughout the development my role has been to provide end user guidance and facilitate user research.

8.3. Methods

8.3.1. Update and Enhance Graph2Data

For the initial development of the tool, the format type of graphs which typically featured in research publications were identified from 34 publications from two different preclinical fields (animal models of neuropathic pain and animal models of D-galactose induced aging) to be dot, scatter, bar, line, box, and area under the curve (Figure 8.1). These had been selected to ensure relevance to the corpus for the trial (Cramond et al., 2019). Therefore, it was necessary to conduct a more accurate confirmation of which structures of graph typically feature in pain focussed research papers using the chemotherapy induced peripheral neuropathy data set (n=341 primary research studies).

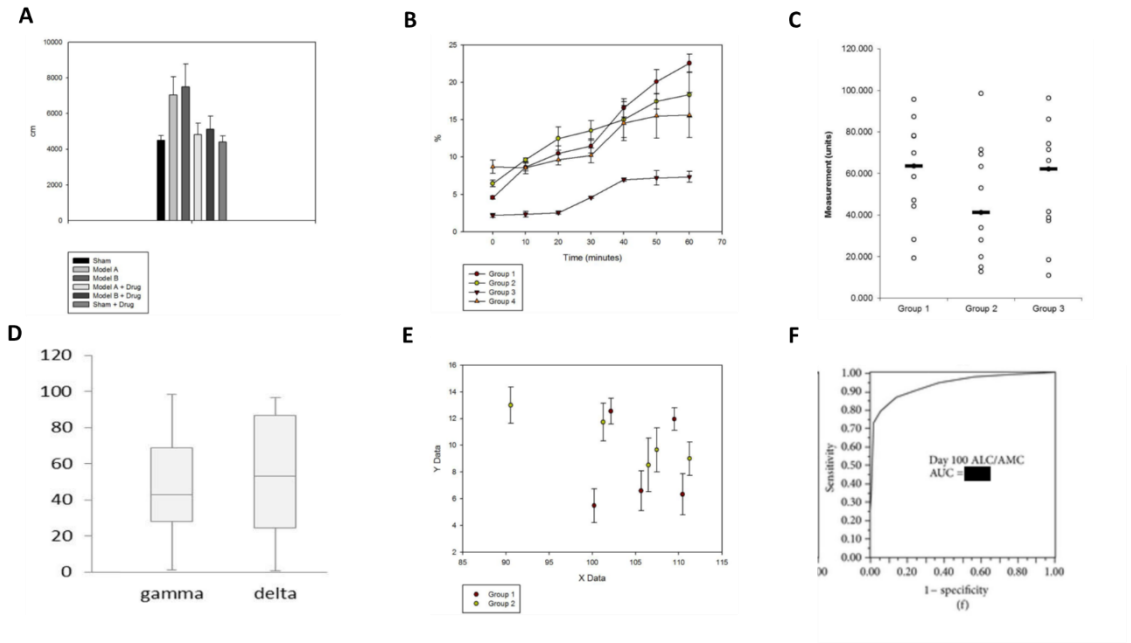


Figure 8.1. **Different graph types that feature in preclinical research publications.** (A) Bar plot, (B), line graph, (C) dot plot, (D) box plot, (E) scatter plot and (F) area under the curve. Graphs pictured were plotted for the Graph2Data non-inferiority trial (Cramond et al., 2019).

The developers (LDG and LY) were employed to improve the functionality and the user interface of Graph2Data as well as build the application for integration into SyRF. Throughout their development I provided testing, explained how reviewers are likely to perform tasks and the features required to meet user needs. At several stages of the development, I organised opportunities for future users to test the application and provide feedback to the developers. The sessions allowed the developers to improve their understanding of user expectations, behaviours, needs and motivations through methodological investigative approaches. This included identifying user accessibility, workflow,

usability (what works well and areas for improvement), overall experience, and how it fits into the broader context of a preclinical systematic review.

8.3.2. Randomised Controlled Trial

In close collaboration with the CAMARADES developer (CS) a protocol has been developed for the assessment of randomised controlled trial of Graph2Data. It contains the rationale for the trial, trial objectives, the methodology, data collection and analysis plans. My role was to design the trial, determine the outcomes, conduct and analysis. This trial will also demonstrate the broad utility of SyRF; a platform on which a randomised controlled trial can be conducted. CS is developing the technical functionality to support the trial which includes a sorting and randomisation algorithm to ensure even distribution of studies to intervention and reviewer to mitigate bias. The aim was for the trial to be instigated as part of the Cannabis & Pain preclinical systematic review however the review timelines could not be extended to incorporate delays in the development and integration. A new systematic review(s) will be identified for the trial. The protocol will be amended to reflect the change of review and ethical approval will be sought for the recruitment of new reviewers into the trial.

8.4. Results

8.4.1. Identification of graph types

Using the chemotherapy induced peripheral neuropathy data set (n=341 primary research studies) (Currie et al., 2019) the frequency of each graph was recorded for the presentation of behavioural assessment experimental data. For behavioural assessments, the line graph is most frequently used graph, closely followed by bar graphs. Scatter and dot plots are used infrequently, and the box plot only used once. AUC data was presented in bar charts (58/954) (Table 8.1). This confirms that the Graph2Data application includes the ability to extract data from all graph structures that commonly appear in preclinical research.

Table 8.1. Frequency of graph types presenting *in vivo* experimental data.

	Graph Type					
Frequency	**AUC**	**Bar**	**Box**	**Dot**	**Line**	**Scatter**
Behavioural	0	954	1	16	1166	15

8.4.2. Development of Graph2Data

Graph2Data was developed as a browser-based solution for ease of integration into a web-based systematic review software. The version of Graph2Data used in the Cramond et al. (2019) trial required considerable programming for integration into the SyRF platform. While the general functionality and main features have remained the same, it has evolved to improve usability, interoperability, and integration into SyRF and the systematic review workflow. In addition, to improved usability two supporting applications were developed. They provide the ability to identify and select figures of interest from PDFs to then pass in the relevant image(s) to Graph2Data and the ability of SyRF to retain the figure that corresponds to data extracted; the Graph Selector allows reviewers to select a graph of interest from a PDF (with corresponding outcome measure selection) and the Graph Displayer displays the image to aid navigation and reconciliation (Fig 8.2). The three components are being integrated into the SyRF workflow.

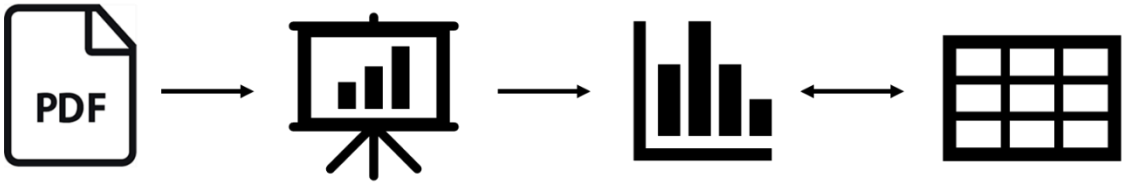


Figure 8.2. The SyRF – Graph2Data workflow. An image(s) of a graph is selected from the PDF of a study. The graph can be tagged with an associated outcome measure. The graph is then displayed as an image. Upon clicking the image, the Graph2Data application is launched. Once data extraction is complete, the data are then passed in to SyRF and populate a table in SyRF. A thumbnail image of the selected graph and corresponding coordinates is also 'saved' on SyRF which allows for Graph2Data application to be re-opened and amend the data extraction, if necessary.

8.4.2.1. Graph2Data Architecture

Each Graph2Data component was designed to be a web-based solution that can be incorporated into and hosted on existing web platforms e.g., SyRF. The components were developed using existing libraries. The Graph Selector and Graph Displayer were both developed using the widely used PDF.js library (version 2.2.228 at the time of writing) which displays PDF files in web-browsers. For the Graph Selector the PDF is rendered and coordinates are applied. The application uses Google's open-source Angular framework (version 12.0.5 at the time of writing). This allows a reviewer to select a graph by drawing a box around it. The Graph Displayer uses the corresponding coordinates to display the image of the graph. Graph2Data was developed using React. React is a JavaScript library for building user interfaces.

The SyRF application acts as a coordinator between the components. SyRF has a reference to a PDF file which is passed into the Graph Selector component. The Graph Selector component is responsible for finding the coordinates of the graph which it emits. Those coordinates with the PDF reference are passed to the Graph Displayer component enabling it to display the selected graph only. When a reviewer clicks on the image, the image is emitted and passed into the Graph2Data component. The reviewer can then use Graph2Data for the extraction of outcome data, once complete Graph2Data will emit the outcome data values. The SyRF front end code will save the outcome data within that annotation session by making an API call.

Once generally available on SyRF, multiple users will be able to use the Graph2Data application simultaneously. The application is running through individuals' browsers, using their own computers resources. The application produces data that gets saved along with the annotation session on SyRF (Figure 8.3).

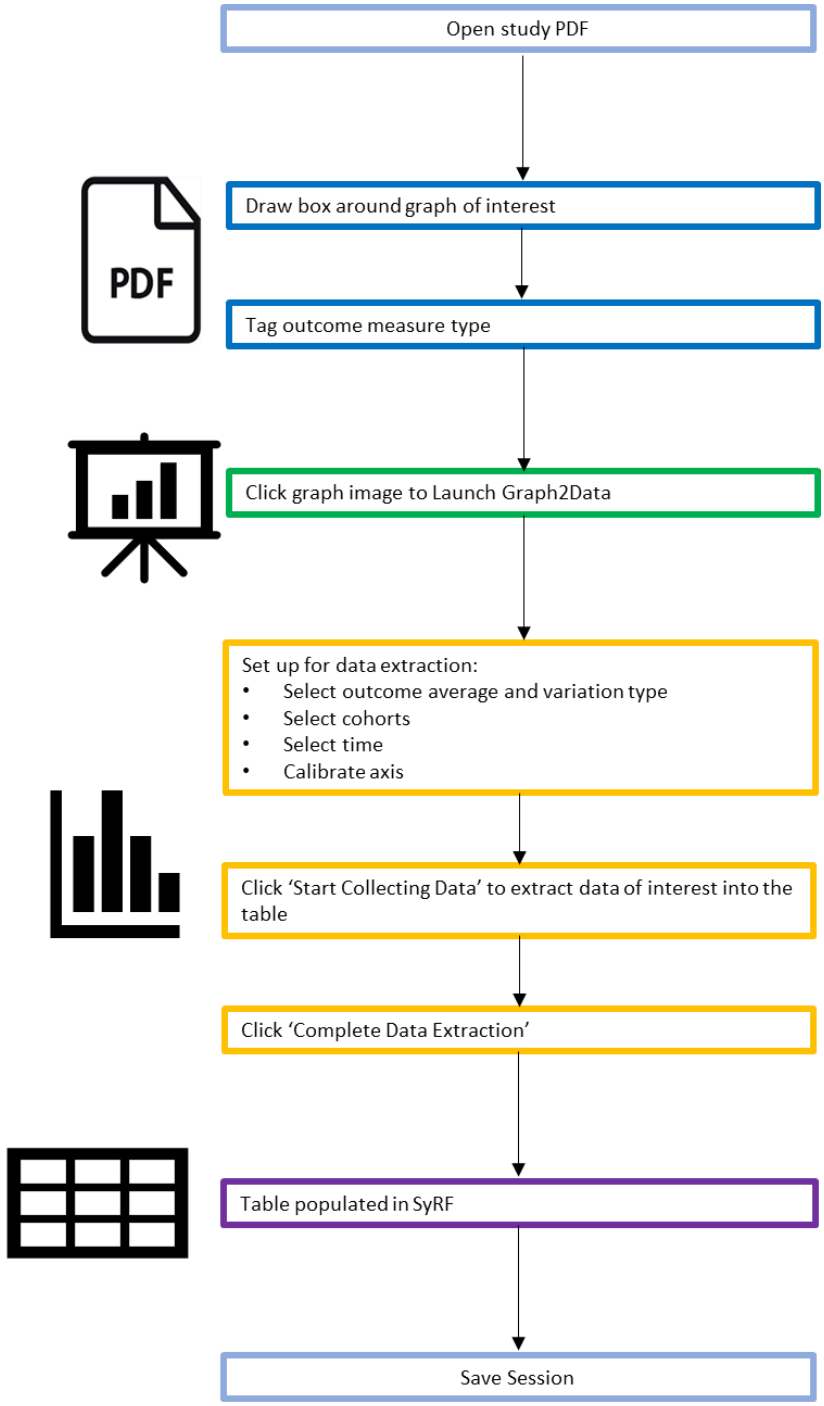


Figure 8.3. SyRF and Graph2Data component workflow. A reviewer selects to open the study PDF. SyRF makes an API call to the server to load a JSON document which displays the PDF on the screen. The reviewer is then able to use the Graph Selector and on clicking the image of the graph, Graph2Data is launched. Once data extraction is completed, the data is emitted to SyRF, and a table of outcome data is populated. When the reviewer saves the session, SyRF makes an API call, and the session is saved on the server.

8.4.2.2. Features of Graph2Data

The starting point for data extraction is to select a figure or figures from the study pdf from which numeric data is to be extracted. The reviewer does this by 'drawing' a box around the figure of interest using their mouse. The reviewer has the option to select multiple figures and they can label each figure's image prior to launching Graph2Data (SyRF 'sends' the image(s) to Graph2Data). A thumbnail image(s) of the figure will be presented in the application. Clicking on the image will display the graph in the main Graph2Data user interface (Figure 8.4). Here the reviewer, will be taken through a series of selection boxes which will identify the specific data types and structures from the outset. They will then be required to calibrate the axes. They are then able to select the relevant data into a table. The data table is interactive. The structure of the data table matches the parameters entered by the reviewer. Once all data points have been extracted and the data table is complete, the table is exported back to SyRF (Figure 8.4). Other features include a zoom in window on the top right-hand side as well as the ability to fine tune a data point using the cursor keys rather which is more accurate than a mouse.

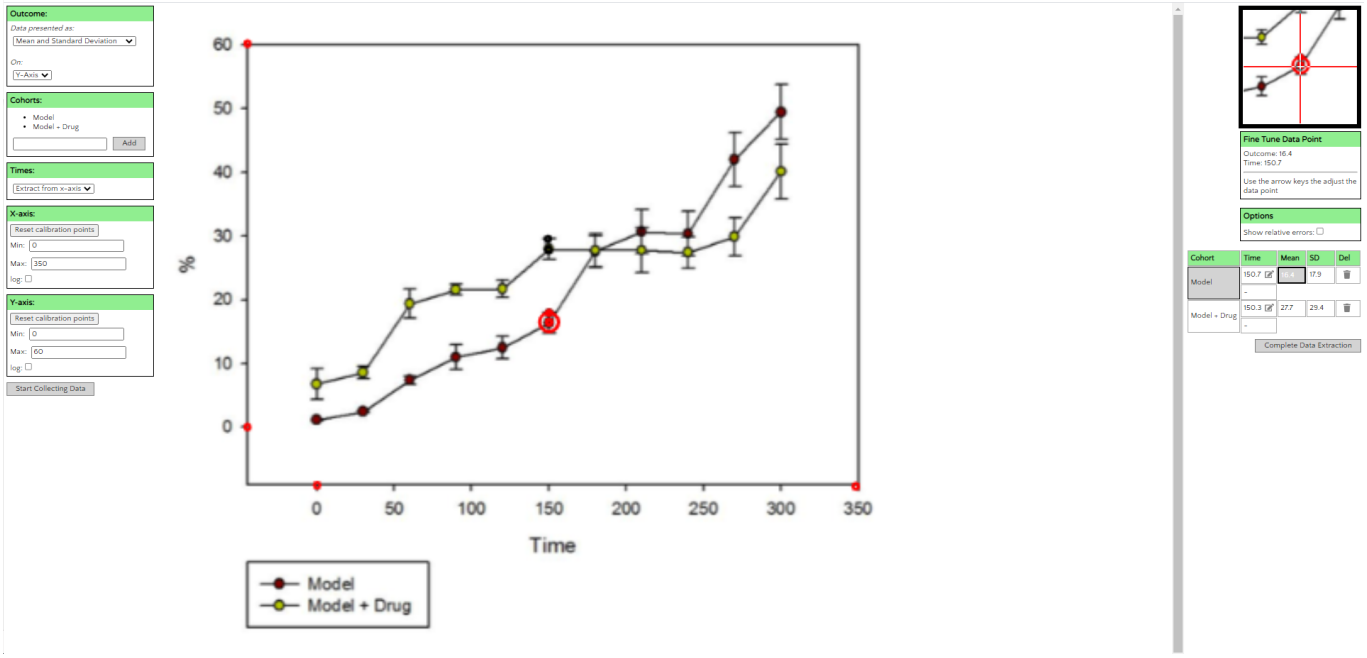


Figure 8.4. The Graph2Data user interface. The series of selection boxes on the left-hand side need to be completed prior to being able to start collecting data. Under outcome, the data type and structure can be selected from a drop-down list e.g., mean and standard deviation. Cohorts can be entered manually or pre-populated based upon previous entries into SyRF. Some graphs have time series data and there is the option to select whether to use the application to calibrate the axes or extract the values manually. Lastly, is the option to calibrate the axes by selecting the corresponding minimum and maximum values and inputting their values. Selecting 'Start Collecting Data' builds the table on the right-hand side and the individual data points can be selected to populate the table.

8.4.3. Randomised Controlled Trial Protocol

A protocol for the randomised controlled trial of Graph2Data has been developed (depicted in Fig 8.5) in close collaboration with the CAMARADES lead SyRF developer (CS). It will be registered on the Open Science Framework. The aims are two-fold; to assess if Graph2Data reduces the time taken for data extraction in a preclinical systematic review on SyRF and whether the use of Graph2Data increases agreement between reviewers, a proxy for accuracy. The initial aim was to test Graph2Data through the Cannabis & Pain preclinical systematic review however, this was not possible due to delays in the development and integration into SyRF.

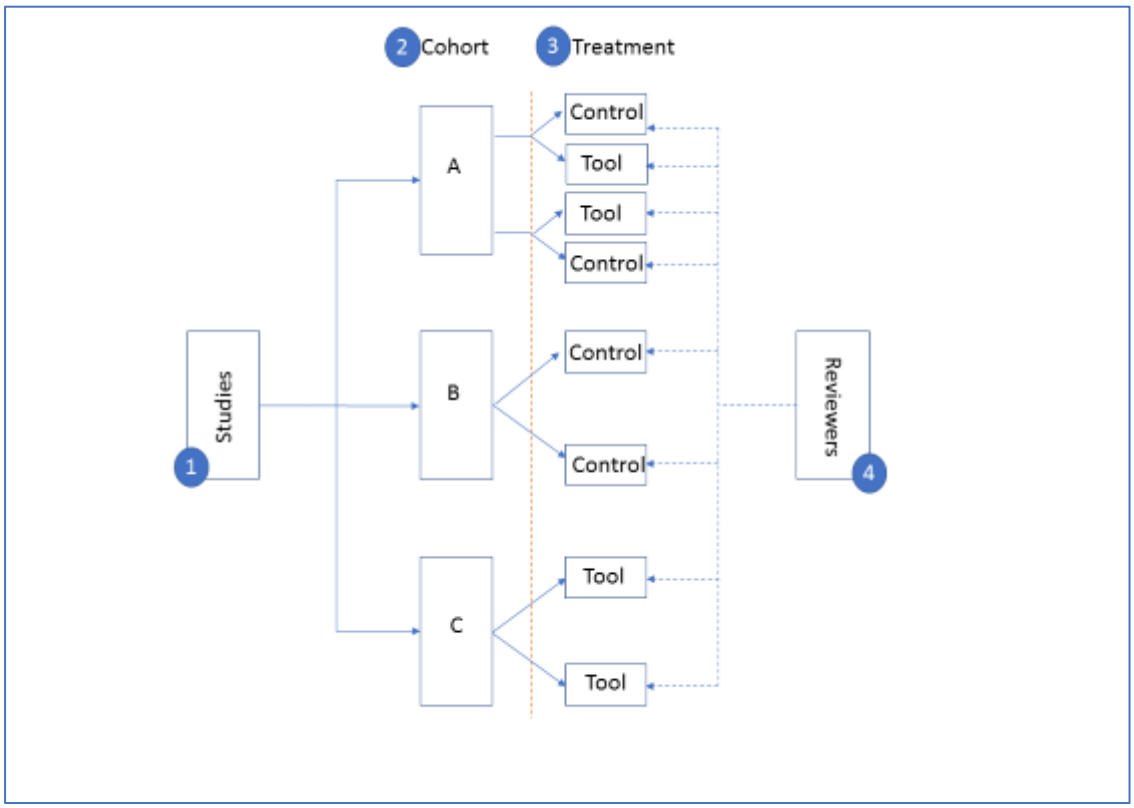


Figure 8.5. The schematic for the randomised controlled trial of Graph2Data: 1. Studies. Studies will be randomly assigned to a cohort using multivariate random sampling to ensure even distribution between the cohorts of both journal and year. **2. Cohort.** Cohort A contains half of the available studies and cohorts B and C each have one quarter of the available studies. **3. Treatment**. The cohorts are organised to aid the between and within group comparisons. Within each cohort studies are assigned a treatment; either the control (traditional methods of data extraction) or the Graph2Data tool (intervention). A) Cohort A. Each study is allocated to both control and intervention for data extraction. B) Cohort B. Data is extracted from these studies using the current methods only (control). C) Cohort C. data is extracted from these studies using Graph2Data (intervention). **4. Reviewers**. The dashed lines from the reviewers denotes that the type of study the reviewer is directed to will depend upon what they have already completed. This is to ensure that reviewers are presented evenly with both an intervention and control study thereby reducing the impact of a single reviewer on the results.

8.4.3.1. Study Design

Aim

To determine whether the use of the customised toolkit reduces the time taken for data extraction in a preclinical systematic review on SyRF and whether the use of the toolkit increases agreement between reviewers, a proxy for accuracy.

Population

Studies that were determined in the screening process to meet the inclusion criteria will contain pain-associated behavioural outcome data in graphs for extraction.

Intervention

Customised data extraction toolkit. The user will be able to extract data from the graphs by specifying axes values and data types and then click the appropriate data points on the screen using their mouse.

Comparator

Current methods for manual data extraction as preferred by the reviewer (e.g. manual measuring, Adobe ruler, WebPlotDigitizer)

Outcome

The primary outcome is agreement between the data extracted by 2 reviewers and the secondary outcome the time that is takes to conduct the data extraction.

8.4.3.2. Randomisation of studies to intervention or comparator

Studies which meet the inclusion criteria for the systematic review will sorted into sub-groups; firstly, by journal, categorised by type into 5 subgroups based upon journal titles;

- pain
- pharmacology
- medicinal chemistry
- neuroscience
- other

From the five journal subgroups the studies will then be grouped into 5-year blocks.

To ensure that journal and year published are evenly distributed between cohorts, studies these sub-groups will be randomly allocated into one of three cohorts; half the studies will be allocated to cohort A and the remaining half split between cohorts B and C. This is:

1. **Cohort A.** Each study is allocated to both control and intervention for data extraction
2. **Cohort B**. Data is extracted from these studies using the current methods only (control)
3. **Cohort C**. data is extracted from these studies using the customised data extraction toolkit (intervention)

The SyRF platform supports the random allocation of studies, the upload of publications and outcome assessment for the project.

The included studies will be stratified into groups according to journal and publication year. The objective of this is to improve the precision of the sample within each of the cohorts by reducing the sampling error. The percentage of each category combination will be calculated, and the percentage used to define the probability of the randomisation result i.e., the allocation of studies to respective cohorts.

For studies allocated to **intervention** participants will use the integrated customised data extraction toolkit kit. The participants will be able to select the graph of interest, label the axes and select the data point for extraction. The graphical data extraction application will automatically measure the time that participants spend extracting data from each graph.

For studies allocated to **control** participants will be asked to extract data using their preferred current method. This is because the current method varies from person to person, it may be a specific toolkit e.g., Desktop Ruler, Adobe measuring toolkit or WebPlotDigitizer or simply holding a ruler to the computer monitor. The SyRF platform will collect time taken per graph; the participant will be asked to press a 'start' button, at the point of selecting the relevant graph for data extraction and once the extracted data has been transcribed, the participant will 'stop' the timer.

All participants will be instructed not to leave the platform running as this will affect the accuracy of the time measurements.

8.4.3.3. Outcome Assessment

As is standard for systematic reviews study data extraction should be conducted by a minimum of two reviewers (i.e., each study will have data extracted twice) and any discrepancies reconciled by a third independent reviewer (Higgins et al., 2016). Within cohort A each study has data extracted once by the traditional method and once using Graph2Data. Studies within cohort B will have data extracted twice using the traditional method, similarly, studies within cohort C will have data extracted twice using Graph2Data.

To extract data from a study, the participant will logon to SyRF and will be allocated the next study requiring data extraction for the first time. The study can be from any of the groups and assigned to either intervention or control but studies. For maximal objectivity and ambiguity reviewers are required to conduct data extraction using both control and intervention methods to an equal number of studies. In order to ensure this, an algorithm will randomly allocate studies that have not been

reviewed based upon the reviewer's annotation history (i.e. proportion of control or intervention that they have completed).

8.4.3.4. Participants and Recruitment

This systematic review aims to employ a "crowd" with the aspiration that it will improve efficiency of the review process. The recruitment strategy aims to target people who are interested in contributing to an important area of research and looking for professional development opportunities. As such, participants will have widely varied experience of both the systematic review process and data extraction. However, participants will receive training using an online platform (Learn to SyRF) which will be supported by a training manual with comprehensive instructions. During their training, they will be presented with studies to extract data from, and their data extraction will be compared with a 'gold standard' derived following review of a set of studies conducted in house by the team. Once their concordance with that gold standard is greater than 80% for three successive studies they will be considered to have been trained to a sufficient standard. They will be required to achieve this using both the current method(s) and the customised data extraction toolkit before proceeding into the live systematic review and the trial.

Study randomisation into blocks will allow for performance management and quality control by members of the study committee. As previously discussed, two independent reviewers will extract data from the same graphs and discrepancies reconciled by a third independent reviewer. It will be possible to identify reviewers who are continually inaccurate and conduct remedial training. Anyone suspected of being purposefully malicious will be removed.

8.4.3.5. Primary Outcome Measure

The time taken for data extraction using the current method and the customised data extraction toolkit.

8.4.3.6. Secondary Outcome Measure

The level of agreement between 2 reviewers between the methods, within the current method and within the customise data extraction toolkit.

8.4.3.7. Statistical considerations

Power Calculations

The aim of this trial is to test the customised toolkit within a systematic review. The pilot study (Cramond et al., 2019) forms the basis of this trial, used a non-inferiority trial design. The graphs were created for the purpose of the trial and therefore the true values were known. The 10 individuals in the study assessed 23 graphs using both current methods and the customised toolkit therefore allowing for direct comparisons and accuracy to be determined. Use of the toolkit reduced the mean time taken by 5 min and 52 secs. The effect size calculated from the results derives a sample size of

22 graphs (calculated using G*power; paired t-test, desired power = 95 %, significance level (α) = 0.01). However, to account for the likely increased variation (more reviewers of differing experiences, and the context of a systematic review) we have doubled the variation and require a sample size of 75 graphs.

It is anticipated that screening will identify a few hundred studies which meet the inclusion criteria which therefore require data extraction. Hence the sample will exceed what is required to determine if Graph2Data is superior to current methods.

8.4.3.8. Statistical Analysis Plan

As previously discussed, study data extraction is conducted by a minimum of two reviewers (i.e. each study will have data extracted twice) and any discrepancies reconciled by a third independent reviewer (Higgins et al., 2016). Firstly, it is necessary to determine the tolerable bounds for agreement that will represent an 'accurate' extraction. In the review, the data extraction will be performed by two individuals working independently and errors of one extractor can be detected when disagreement is observed with the other extractor, and these data points identified for third person reconciliation. An extracted data point from each reviewer must be within 10% or less, and an average of the 2 measurements is taken. If the two values have a difference greater than 10% then the discrepancy will be reconciled.

Each study within Cohort A has data extracted by the traditional method and using the customised toolkit. This will allow for direct comparison for the time that it takes for data extraction in addition to the agreement between the two methods. Studies within cohort 2 will have data extracted twice using the traditional method, similarly, studies within cohort 3 will have data extracted twice using the Graph2Data, both allowing the opportunity to calculate the agreement within the method. Comparison of cohorts 2 and 3 will explore reliability, time taken, and agreement between reviewers.

8.4.3.9. Primary Outcome

For studies within Cohort A, to establish the time taken to extract the data for each method a within-subjects design will be used. Participants are required to extract the same data from the same graphs in each condition, it will be possible to directly compare how long it took using each method of extraction. To measure differences between approaches, for each study, subtraction will be used to calculate the difference in time taken between each method, such that a positive value will indicate that the current methods will take longer than the new method. For each study, a mean difference in time taken across participants and the standard deviation will be calculated. In addition, the total time taken for data extraction for studies in both conditions (all 3 cohorts) will be calculated and expressed as minutes (overall time for control or intervention).

8.4.3.10. Secondary Outcome

It is not possible to establish the accuracy of data extraction because the true values are not known. However, to establish the level of agreement between reviewers the intra-class coefficient (ICC) will be used for each method. From the ICC values the standard error of measurement will also be calculated.

To give a summary estimate of differences in the agreement between reviewers using the different methods, the percentage of agreement using the new method and that using conventional methods will be calculated.

8.4.3.11. Qualitative analysis plan

A secondary aim is to consider the user's reactions to the customised data extraction toolkit. Analysis of the multiple-choice questions will involve frequency and percentage of the participant responses (Annex A). Analysis of the open-ended text questions will involve coding the text responses into categories derived from the data (i.e., not *a priori*); for example, if the participant comments on how quickly they extracted data, this will be coded relating to the theme of 'speed'. The frequencies of themes will then be examined.

8.4.3.12. Ethical approval

Ethical approval will again be sought from Imperial College London's Science, Engineering and Technology Research Ethics Committee (SETREC)**.** Participation within this study is considered of low ethical concern to the individual.

8.4.3.13. Consent

Participants are self-selecting and will have expressed interest in participating in response to the IASP advertisement or other recruitment methods. They will receive a detailed participant information sheet (Annex B) and are given the opportunity to ask further questions prior to completing an online registration and returning a signed informed consent form (Annex C). We consider the 'subject' for this study to be the data extraction method not any of the individuals involved.

8.4.3.14. Blinding

Reviewers cannot be blinded as to whether a study is within the intervention or control. Knowledge of the data extraction toolkits is highly likely to alter the behaviour of reviewers. This is likely mitigated by the necessity of performing data extraction for the purpose of achieving the aims of the systematic review. However, for any discrepancies requiring resolution, the third independent reviewer will be blinded to whether the studies are from intervention or control as well as blinded to the identity of the reviewer.

- PROTOCOL END -

8.5. Discussion

Graph2Data has the features required to support the semi-automation of a broad range of preclinical data types presented in different graph types in studies. The integration into SyRF and the adaptation of the SyRF workflow was complex and reliant on the expert skill and availability of the CAMARADES SyRF developer who was also balancing other complex developmental needs of SyRF against several projects. It remains a high priority to bring this feature to preclinical systematic reviewers given the anticipated benefits. It is now necessary to identify a systematic review(s) in which the performance of the tool can be evaluated in a 'real-life' setting.

The lessons learned highlighted by Cramond et al. (2019) have been incorporated into the trial design and accompanying processes. Firstly, the aim is to conduct a larger study to determine superiority. The trial also aims to better reflect reality of the systematic review and data extraction process. It therefore requires a review(s) with several reviewers (the higher number of reviewers the more generalisable the findings) and the included studies will need to be equally distributed (by journal and year) between 'treatment' groups. This is because certain graph types may be more prevalent in certain journals and reporting trends are likely to have changed over time. Additionally, depending on research question, reviewers may only be extracting a single time point from a graph or will be extracting other information. Reviewers' decisions will also be influenced by reading the rest of the paper and therefore there are several confounding factors to data extraction that were not accounted for in the pilot study. All reviewers will receive training to understand how to use the application and ensure that they are able to select the correct graph and data types when using the tool. The primary outcome is the time taken to extract the data and a within-subjects design will be used for the analysis. It will not be possible to establish accuracy because the true values of the data presented in studies will not be known, however the level of agreement between the reviewers will be determined as a proxy for accuracy using an intra-class coefficient. Participants in the trial will be asked to feedback their experience including their perceptions on the ease, speed, and features of the tool. They will also be asked to state which methods they prefer to use for data extraction. This information will identify further user needs and inform future development.

For the integration we have built a graph selection feature to optimise the workflow. It requires the reviewer to select the graph and corresponding outcome measure prior to launching the Graph2Data application. It is unlikely that this feature will ever be fully automated because most reviews have specific requirements and studies can present varied graph types and not all outcome data are required, so human intelligence will be required to determine which graphs are relevant. However, there are several steps within the workflow that could be automated to assist us in getting closer to the goal of minimal human effort e.g., the identification of graph axes and their values so that a

reviewer does not need to enter these manually, recognition of graphs as potentially relevant or potentially irrelevant to aid reviewers as discussed by Cramond et al. (2019). We also built a thumbnail image of the graph which can launch the Graph2Data application with a reviewer's 'stored' axes calibration and extracted data. This will aid in the identification and cause of discrepancies so that time will be saved during reconciliation and resolution of reviewer disagreements.

Graph2Data aims to improve the abstraction of data presented in graphs. Results from the non-inferiority trial demonstrated that such an application can reduce errors and save time, but it remains that not having access to the raw data will limit precision. The need for automation tools could be negated with the sharing of data in a machine-readable format. Funders and journals are increasingly requesting researchers to share their data in accordance with the FAIR (findable, accessible, interoperable, and reusable) principles (Wilkinson et al., 2016). Data sharing for re-use in meta-analyses presents a different array of challenges; data will be required to have adequate description and attributes, clear and accessible data usage licence, detailed provenance and meet domain-relevant community standards. Data will be formatted in a way that a computer can automatically read and process it e.g., CSV, JSON, XML, thereby increasing accuracy by removing the need for manual extraction, machine-assisted or not, which will be further aided by the potential availability of individual animal data not just aggregate data.

The need to extract data from already published studies will remain and for new studies until machine-readable data sharing is more widely adopted or even mandator, the new additional features of Graoh2Data and its integration into the SyRF workflow offers substantial data extraction assistance. If the findings of the Graph2Data trial confirm those of the pilot study (Cramond et al., 2019) and superiority is demonstrated, then this new feature will have a considerable impact on improving the feasibility, accuracy and efficiency of data extraction for preclinical systematic reviews.

9. Discussion

"Pain is an unpleasant sensory and emotional experience associated with, or resembling that associated with, actual or potential damage" (Raja et al., 2020). It diminishes quality of life and is the leading cause of disability worldwide (Rice et al., 2016). Current analgesics often lack mechanistic rationales and are characterised by limited efficacy and unwanted side effects (Finnerup et al., 2015, Aldington and Eccleston, 2019), thus there is an urgent unmet medical need for improved treatment. Despite recent collaborative efforts between academic institutions and industry to develop better analgesics, only a few novel candidates have shown promise in animal models and these have failed to demonstrate efficacy in clinical trials (reviewed by Borsook et al. (2014), Mao (2012), and Attal and Bouhassira (2019). This demonstrates there is a fundamental difference between preclinical and clinical trials, and it comes at a high cost, both in terms of research waste and to patients. To address the translational failures, it is paramount that researchers can assimilate the published research to reliably inform next steps.

Systematic reviews and meta-analyses are well-established strategies for answering specific research questions by use of explicit methods to identify, combine, and integrate primary research. They are vital research tools that can provide evidence to provide an overview of available evidence, identify knowledge gaps, improve methodological quality, determine efficacy, identify factors influencing efficacy, translate animal data to the clinic and inform future research agendas (Sena et al., 2014). Preclinical research is characterised by high volume, rapidly accruing data from studies with wide variation in the design, conduct, analysis, and reporting, coupled with time and resource intensive systematic review processes presents researchers with several feasibility challenges. Overcoming these challenges will ensure the full potential of preclinical systematic reviews can be realised and the value and translational ability of preclinical results improve (Soliman et al., 2020).

The aims of this PhD were two-fold: (i) address key areas of neurobiology to provide empirical evidence to improve and guide future preclinical research which offered the opportunities (ii) to evaluate and operationalise methods improve the feasibility, efficiency, and accuracy of preclinical systematic reviews, such as crowd science, machine learning for study selection, and semi-automation of data extraction (Graph2Data; a data extraction application) as well as exploring how systematic methods can be used to improve narrative reviews.

There is a spectrum of evidence synthesis methods available to assist researchers in their critical appraisal and review of a literature base. A traditional narrative review and a systematic review were conducted during this PhD and their characteristics are compared in table 9.1. Through the VGF review (Soliman et al., 2019b) I have highlighted how the methodological quality of a traditional

narrative review can be improved using systematic review methods without undermining the aims and functions of a narrative review. The Cannabis & Pain review (Soliman et al., 2021) has provided an overview of the available evidence pertaining to the anti-nociceptive efficacy of cannabinoids, cannabis-based medicines, and endocannabinoid system modulators. I have demonstrated the broad utility of preclinical systematic reviews including identification of knowledge gaps, methodological quality of studies, factors that influence treatment efficacy, and the misalignment between preclinical and clinical trial design. The empirical evidence has been used to inform a future research agenda (Haroutounian et al., 2021), including the experimental design of new studies and translation of animal data to the clinic. The review provided the opportunity to assess different methods for improving the feasibility, accuracy, and efficiency of preclinical systematic reviews. I report that while further development and refinement of machine learning for study selection and crowd science is needed, both demonstrated potential for improving the feasibility and efficiency of the systematic review process; either by increasing the size of a systematic review and/or reducing the resource requirements for evidence synthesis. In this chapter, I review my main findings and how they inform translational pain research improvement. I also discuss and critically appraise the benefits and challenges of preclinical systematic reviews and the emerging methods and technologies that aim to improve the feasibility, accuracy, and efficiency of preclinical systematic reviews.

Table 9.1. Comparison highlighting the differences between the two reviews undertaken during this PhD.

Review	Review Type	Aims	Methods	Synthesis
VGF	Narrative/Mixed methods	Examine the literature	Systematic search and study selection (single reviewer)	Qualitative – thematic narrative synthesis. Tabular summaries, included quality assessment
Cannabis & Pain	Systematic	Systematic search, appraisal and synthesis of all evidence pertaining to the anti-nociceptive efficacy of cannabinoids, cannabis -based medicines, endocannabinoid system modulators assessed in animal models of injury-related or pathological persistent pain	Systematic search and study selection Structured data extraction (Methods conducted in duplication with disagreements reconciled by third independent reviewer)	Quality assessment Narrative and tabular synthesis of study characteristics Quantitative meta-analysis including influence of study design characteristics, methodological quality criteria on the observed effects

9.1. The Role of Animals in Pain Research

Animals are used to assist in determining the aetiology and pathophysiology of pain. Animal experimentation affords the benefit of controlling influencing factors to systematically determine the impact a single factor has on the initiation, development, and maintenance of chronic pain. It is not possible to perform such controlled experiments in humans, often the type of research being conducted is not practicable nor ethical in human subjects e.g., dissection of the brain or spinal cord, or administration of a drug with an unknown mechanism of action. Discoveries related to the pathophysiology of pain are primarily disseminated to the research community and beyond, in peer reviewed journals. These studies inform the field and leads to the generation of new hypotheses for the discovery and development of new therapeutics. However, clinical trial failures have led to questions about the predictive validity of preclinical research; how relevant and accurate the animal models at are predicting the human condition. It is also postulated that the low rate of successful translation of preclinical research findings into safe and effective therapies, across many fields of medicine, not just isolated to pain research, are results of inadequate study design (Andrews et al., 2016b), lack of scientific rigour (Currie et al., 2019, Hirst et al., 2014, Crossley et al., 2008, Rooke et al., 2011), and transparent reporting (Rice et al., 2008, Clark, 2016). To add, reports suggest that a substantial amount of published literature is likely to be biased and irreproducible (Andrews et al., 2016a, Freedman et al., 2015, Macleod et al., 2014). Understanding the causes of nonreproducible and nonreplicable research through systematic reviews should assist pain researchers to improve the reproducibility, replicability, and translatability of their work. Lack of transparency, underpowered studies and researcher degrees of freedom have been highlighted as common factors that drive nonreproducible and nonreplicable research (Lee et al., 2018). In addition, a lack of valid preclinical disease models of pain and assessment strategies including the limitations of current methods to assess nociception and pain in non-human animals, and species and sex differences are cited as important factors (Mouraux et al., 2021). These factors were evident from both the VGF and Cannabis & Pain reviews.

9.2. Methodological Quality: poor reporting practises?

The findings presented here support the notion that the methodological quality of preclinical pain research may be hampering translational efforts. Both the VGF and Cannabis & Pain reviews report a low propensity to report methodological quality measures to reduce the risk of bias, findings that are commensurate with other preclinical systematic reviews (Currie et al., 2019, Macleod et al., 2015). The reliability of the results is underpinned by the mitigation of experimental bias. Failure to mitigate risks of bias can lead researchers to incorrectly attribute a change in the observed outcome

to the independent variable (i.e., the intervention under investigation) when it was in fact a result of confounding factors (Higgins et al., 2011).

Table 9.2. Comparison of the prevalence of reporting of methodological quality criteria in the VGF and Cannabis & Pain reviews. Additional comparison with another preclinical systematic review from the pain research field.

Methodological Quality Criteria	**VGF (%)**	**Cannabis & Pain (%)**	**Chemotherapy Induced Peripheral Neuropathy (%)**
Sample size calculation	0	3	2
Allocation concealment	0	14	1
Randomisation	22	13	28
Blinding	67	47	51
Animal Exclusions	11	4	17
Conflicts of Interest	44	56	50
Animal welfare regulations	100	94	97

Of concern, is that the low prevalence of the reporting of measures to mitigate bias tend to give higher estimates of treatment effects (Crossley et al., 2008, Hirst et al., 2014). This was identified in the Cannabis & Pain review too. The studies that reported conducting a sample size calculation and allocation concealment were associated with significantly smaller effect sizes than those that did not. Bias in an experiment may result in wrongful conclusions in both the primary research and in subsequent systematic reviews and meta-analyses. An overestimate of treatment efficacy may lead to the candidate being taken forward to clinical trial putting participants and patients at unwarranted risk.

The most frequently under reported criterion is the performance of a sample size calculation, again commensurate with the findings of other preclinical systematic reviews (Currie et al., 2019, Bahor et al., 2017, Macleod et al., 2015). Researchers must ensure that the sample size is large enough to have a high probability of detecting a treatment effect of a given size while balancing against the legal, ethical, and practical considerations in following the principles of the 3Rs (replacement, reduction, and refinement). Animal studies tend to use small group sizes, as was reflected in the reviewed literature, which can increase their imprecision and give less accurate estimates of effect (Button et al., 2013, Ioannidis, 2005). Button et al. (2013) estimated the average power in neuroscience animal studies to be around 20 % which signifies a risk of missing a genuine effect because only 1 in 5 experiments would have a chance of detecting it. The use of sample sizes that are too small reduces the reliability of conclusions from an individual experiment. The smaller the study, the less likely the research findings are to be true (Ioannidis, 2005). To add, although not investigated systematically here, animal research authors rarely state or report the primary outcome for which they should be

powered. Underpowering of studies can lead to spurious false-positive or false-negative effects. As demonstrated, justification for the choice of sample size is rarely included in publications (Kilkenny et al., 2009). It is necessary for primary researchers declare their primary outcome (Kimmelman et al., 2014) and ensure that the number of animals used is adequate to detect an anticipated effect so that effects are not misrepresented (Percie du Sert et al., 2020).

The methodological quality criteria used to assess risk of bias of the preclinical literature have been informed by clinical trials. Investigation of methodological approaches of clinical trials demonstrated that authors that did not report randomisation, adequately conceal treatment allocation or use double blinding yielded larger estimates of treatment effects than trials which did report these criteria (Jüni et al., 2001). The analysis of how a methodological quality criterion influences effect size is limited because it assumes independence and does not account for interactions between the quality criteria. It is unlikely that they operate independently and performing measures to reduce the risk of bias may be a proxy for other unreported factors that may also influence the observed outcome (van der Worp et al., 2010). The relative importance of the possible sources of bias are not yet known and mitigation measures may vary in relevance for different types of experiments.

To perform an assessment of quality requires authors to transparently report all experiment details. Within the Cannabis & Pain systematic review all studies were rated 'unclear risk of bias' because methods to mitigate experimental bias were not reported in sufficient detail. The inherent risk of bias can affect both the individual study conclusions and meaningful interpretation of the reported results. It is critical that all experimental details are reported so that the quality of research can be assessed (Sena and Macleod, 2007).

Several reporting guidelines have been developed to improve the conduct and reporting of preclinical research. The ARRIVE guidelines (Animal Research: Reporting In Vivo Experiments) were developed in 2010 to help authors and journals identify the minimum information that should be reported of research involving animals. They aim to improve the reporting of research to maximise the quality and reliability of published research and enable others to critically appraise and reproduce it (Percie du Sert et al., 2019, Percie du Sert et al., 2020). Similarly, Landis et al. (2012) proposed a core set of reporting standards; randomisation, blinding, sample size estimation and full information relating to data handling (known as the Landis 4). They posit that improving how animal studies are reported will raise awareness of the importance of rigorous study design which will accelerate scientific progress and the development of new therapies. Specific to preclinical pain research Andrews et al. (2016a) proposed the PPRECISE considerations.

Although there is consensus, the ARRIVE guidelines are the most comprehensive and include a compliance checklist for use by journal staff, editors, and peer reviewers. It has long been documented that we need to improve. Systematic reviews have instigated a cultural shift however, with increasing guidelines and little enforcement a research question requiring answering is how effective are these interventions? What is required beyond guidelines to improve the quality of preclinical research? The high compliance with reporting conflicts of interest and compliance with animal welfare regulations as mandated by many journals is an example of stringent editorial policy having a positive effect. However, disappointedly a randomised controlled trial found that journal requested completion of an ARRIVE checklist did not improve compliance with the guidelines suggesting the editorial process alone is not enough to improve reporting, their findings consistent with prior reports that endorsement by editors and reviewers did not improve compliance with ARRIVE (Baker et al., 2014, Gulin et al., 2015, Avey et al., 2016). Other approaches need to be considered based upon understanding the barriers to implementing quality check lists for animal experiments (Hair et al., 2019). Training and availability and accessibility of other tools designed to improve research quality are likely to facilitate implementation of guidelines over time (Vollert et al., 2020). As more journals require authors to clearly state their hypothesis, experimental methods, statistical analysis methods, and disclose conflicts of interest, Lee et al. (2018) have provided a table detailing the transparency standards of ten leading pain journals; at the time of writing only one journal requires ARRIVE guideline compliance and three journals encourage its use. Systematic reviews will continue to play a vital role in the assessment and critical appraisal of study quality as well as the impact of interventions designed to improve reporting of animal research.

Mitigating bias is also pertinent for evidence synthesis methods. The VGF review employed a systematic search and study selection to ensure that all the relevant literature was identified and included. The Cannabis & Pain review, as a systematic review, followed the prescribed methods (Sena et al., 2014, Vesterinen et al., 2014, Hooijmans et al., 2014). It was reported in line with the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRIMSA) guidelines. All systematic reviews should be reported in adherence to the protocol with explanation of deviations. It is also helpful, when writing or reviewing systematic reviews to refer to the AMSTAR 2 critical appraisal tool for assessing the methodological quality of systematic reviews (Shea et al., 2017). In addition, The Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) approach has been adapted to rate the certainty of the evidence and the strength of recommendations from preclinical studies in the context of therapeutic interventions (Hooijmans et al., 2018). To ensure transparency and sustainability of the Cannabis & Pain systematic review I made the data accessible by uploading

data to the Open Science Framework https://osf.io/2qde5/ . The data is shared under a CC-BY license and free to other researchers to use.

9.3. Animal Models for Pain Research

An animal model consists of three components; the inherent characteristics of the animal e.g., species, sex, strain, age; the injury or disease that is induced, and the assessment. Each component will be discussed independently throughout this discussion therefore references to animal model are references to the injury or disease induced.

The VGF review highlighted the broad range of uses for animals in pain research including exploration of the basic physiological mechanisms of pain and the identification and validation of new pharmacological targets. The Cannabis & Pain review focused on the role of animal models of injury-related or pathological persistent pain used to predict anti-nociceptive efficacy of cannabis-based medicines, cannabinoids, and endocannabinoid system modulators to lead to clinical drug development. The findings of the reviews suggest that there are several priorities to address translation challenges: (i) bridge the gap between preclinical and clinical presentation, (ii) develop outcome assessments that incorporate the psycho-social components of pain, (iii) explore options to improve both preclinical and clinical research through forward and back translation.

The limited clinical trial success of drugs that passed preclinical evaluation, points to the need for optimising the current preclinical approaches to better mimic the complexity of human disease and pain sensory phenotypes. To increase probability of success the choice of animal model and outcome measures are important, these need to reflect the target population as closely as possible. In addition, evaluation of potential drug candidates should include pharmacodynamic, pharmacokinetic and toxicology studies to ensure anti-nociceptive efficacy is not associated with unwanted side effects.

9.3.1. Poor Animal Models

The models used in preclinical pain research are not often well matched to the clinical population (du Sert and Rice, 2014, Rice et al., 2018). Firstly, animal cohorts are genetically very similar converse to the widely heterogenous patient population. The Cannabis & Pain highlights several weaknesses in the face and construct validity of the animal models. The data also demonstrates that there are several study characteristics that can impact pain-associated behavioural outcome measures and the assessment of novel anti-nociceptive efficacy.

A criticism frequently levelled at pain researchers is that there is a dominance of nerve injury models (Mogil et al., 2010). Nerve injury models were used to assess cannabinoid, cannabis-based medicine, or endocannabinoid system modulators in 74 % of the included studies. In the VGF review, only nerve

injury models were used to assess VGF expression and to assess the anti-nociceptive potential of the antagonism of VGF-derived neuropeptides. The concern is that an intervention that successfully attenuates nociception will not predict a similar outcome for clinical pain patients where the prevalence of nerve injury is much lower than that of disease (Smith et al., 2020, Finnerup et al., 2015) and animal models of tissue damage do not reflect the multi-dimensional characteristics of pain. In relation, another criticism is that sequelae and co-morbidities are not modelled. Chronic pain patients can experience for example, disability, anxiety, depression and sleep loss which could be modelled in animals that are experiencing chronic inflammatory or neuropathic states (Mogil et al., 2010).

In the Cannabis & Pain review, stratified analysis showed that rodent strain had a significant impact on pain-associated behavioural outcomes and that most studies used either Sprague Dawley or Wistar rats. The difference between strains has also been observed in other animal models (Mogil et al., 2005). There was also a major bias toward male animals. No female animals were reported in the VGF literature and in the Cannabis & Pain review female animals accounted for only 7 %. It was not possible to reliably assess sex differences given the disparity in the numbers however, it has been posited that female animals are more sensitive to painful stimulation than male animals (Mogil and Chanda, 2005, Fillingim et al., 2009). Of additional concern is that chronic pain is more prevalent in women than men and therefore the clinical situation is not reflected (Mogil, 2012b, Berkley, 1997).

Injury-related or pathological persistent pain presents over a long period of time, months to years in human patients (Treede et al., 2015). The prevalence of chronic pain is greater in middle-aged and elderly people. This was not systematically investigated in either review, however, represents another mismatch between chronic pain patients and choice of animal models. Preclinical research is predominantly conducted in young, healthy animals and is more immediate in the investigation of cellular, molecular, and behavioural changes, usually after days or weeks. As such there is uncertainty as to whether these changes could be responsible for a persistent pain state. Researchers investigating the role of VGF is an example of translational research exploring the relationship between initial responses to tissue damage and a chronic pain state, longer experiments are needed to determine the latter stages.

9.3.2. Outcome Measures

The subjective nature of pain experiences makes it difficult to produce experimental pain in animals and there is lack of effective tools to detect the different dimensions of pain in animals. It is not possible to measure in pain in animals and so researchers are reliant on surrogate measures. In the Cannabis & Pain review I recorded five different pain-associated behavioural outcome measures, grouped into evoked; mechanical, cold, heat and non-evoked; complex, formalin. The most frequently

reported outcome was evoked limb withdrawal to mechanical stimulation, accounting for 50 % of all outcome measures. The second most frequently reported outcome measures were evoked limb withdrawal to heat which accounted for 24 % of the report outcomes. This observation is consistent with a much earlier review which identified that reflex hypersensitivity was the most commonly assessed outcome in preclinical studies, with 90 % of the studies exclusively reporting either thermal (48% of studies) or mechanical (42% of studies) hypersensitivity (Mogil and Crager, 2004).

Researchers have focused upon measuring nociception related to the application of noxious stimuli, predominantly limb withdrawal to mechanical and heat stimulation. Although susceptible to experimenter bias, these assessments are rapid, reproducible and inexpensive to perform however, their dominance is criticised because mechanical allodynia and thermal hyperalgesia are not the sole drivers nor most common patient complaint and do not reflect patient sensory phenotypes (Vierck et al., 2008) (Bove, 2006, Chesler et al., 2002). Evoked limb withdrawal to mechanical or thermal stimulation is a measure of hypersensitivity, sensory gain, whereas many patients experience pain due to sensory loss or spontaneous pain. These assessments correspond to some indications e.g., post-operative or inflammation associated pain (Pogatzki-Zahn et al., 2018, Pogatzki-Zahn et al., 2017) but bear little relevance in other clinical indications. For example, HIV patients experience pain due to sensory loss not sensory gain (Phillips et al., 2014). Patients with postherpetic neuralgia also experience thermal and mechanical sensory loss but may also experience dynamic and static mechanical allodynia (Maier et al., 2010). In general, most neuropathic pain patients experience spontaneous pain and sensory gain is much less frequent (Maier et al., 2010, Vollert et al., 2017, Baron et al., 2017).

While in preclinical studies withdrawal response to mechanical or thermal stimulation is the most commonly pain-associated behavioural outcome measure used to assess nociceptive behaviour, patients in the clinic are most frequently asked to self-report their pain experiences. This demonstrates that different parameters are being measured which is problematic for translation. For example, three major sensory phenotypes have been identified in patients with neuropathic pain across a range of underlying diseases; sensory loss, thermal hyperalgesia, and mechanical hyperalgesia (Baron et al., 2017). It would therefore be a flawed decision to take to clinical trial pharmacological interventions assessed to be efficacious based on attenuation of a single outcome and raises questions as to whether it could have clinical application. It is therefore necessary for animal modelling and outcome assessments to be refined to reflect both the disease or injury and the sensory phenotype (Rice et al., 2018).

Given the complexity of the pain experience preclinical pain research needs to move beyond testing nociception, anti-nociception is often not sufficient for the treatment of chronic pain and a distinction between pain and nociception needs to be made. There is a need to develop new objective pain-assessment and monitoring tools. Including assessment parameters that may reflect the perception of nociception in animal models, such as spontaneous behaviours, will be a step toward improving translation by bridging the preclinical clinical gap. It is important multiple assessments are used so that a broader spectrum of pain symptoms is represented. The battery of tests applied should be consistent across laboratories and studies so that meaningful comparisons can be made (Ioannidis 2005). Assessments should include how the brain processes pain perception and related motor, autonomic and psychological responses. There is some evidence to suggest that animals too are affected by chronic inflammatory and neuropathic states (Mogil, 2009). Increasingly researchers are adding ethologically relevant outcome measures to the battery of behavioural assessments e.g., thigmotaxis as a measure of anxiety-like behaviour and burrowing, a measure of animal wellbeing and global impact of pain (Wodarski et al., 2016). Recent advances to automatically track animal behaviour (e.g. grooming, drinking, eating, socialising and resting) in their home cage are allowing 24 hour-a-day assessment of animal injury or disease models e.g., Hasriadi et al. (2021) and Pitzer et al. (2016). In addition, brain imaging is being used to compare the similarities and differences in various brain regions in response to peripheral nociceptive input to improve understanding of the mechanisms of pain processing, modulation, treatment effects and treatment targets (Da Silva and Seminowicz, 2019).

Historically, emphasis has been placed upon reflex withdrawal responses rather than measuring more complex, ethologically relevant behaviours. The development of complex behavioural animal models specific to each condition is required to improve face and predictive validity.

9.3.3. Environmental Factors

There are other experimental design factors that are likely to influence behavioural outcomes that were not included in the Cannabis & Pain review including e.g., housing, diet, handling and habituation, experimental set-up and timings (Mogil, 2017). Housing and handling of animals is rarely reported in the preclinical literature (Avey et al., 2016) however, there is growing recognition for the importance of the housing and testing environment and changes to those environments because health, wellbeing and behaviour of the animal can be effected, thus influence outcome data (Reardon, 2016, Gerdin et al., 2012). The handling of animals is an indeterminant concept, and despite some standardisation will vary between individual experimenters, which makes it difficult to introduce evaluation criteria. Similarly, the experimental apparatus used and the timelines of experiments are rarely and inconsistently reported in the literature (Egan et al., 2016). Aligned with this, from

experience of conducting a systematic review there is a lot of variation in how the 'same' pain-associated behavioural assessment is performed e.g., there are differences in temperature ranges, cut-off times, apparatus, methods, habituation times, no habituation etc. These are all factors that may influence the observed results and hence should be reported so that their impact can be determined and if required, standardised to improve the comparability of data collected from multiple labs.

9.3.4. Paradigm Shift Needed

The poor external validity of animal models used in pain research coupled with the translational failures should be forcing us to acknowledge the weakness of the current system and spur exploration of the theoretical premise needed to improve and progress. Laboratory experiments are strictly controlled which makes it easy to repeat an experiment and replicate or reproduce results. The high control is required to establish a causal link between a single variation and an outcome. The more an experiment is controlled and standardised the more difficult it is to translate the findings to a different context. In pain research homogeneity dominates; the predominant use of male, young, healthy, Sprague Dawley rats in which nerve injury is induced and evoked limb withdrawal to mechanical and thermal stimuli is measured is an extreme example of complexity reduction. It has limited utility for improving understanding of human pain biology.

Increasing heterogeneity of experiments is a potential solution to remedy low reproducibility and improve generalisability of findings. Standardisation is considered the gold standard of experimental conduct because it improves the precision of outcomes. However, as discussed, using young, healthy, male, genetically similar animals, raised in the same environment means that these animals are not representative of the whole population and the findings from experiments may not be generalisable. Voelkl et al. (2020) advocate increasing biological variation to improve reproducibility through "systematic heterogenization of study samples and conditions by actively incorporating biological variation into study design through diversity study samples and conditions." They provide evidence that supports increasing biological variation increases chances of finding robust effects concurrent to reducing the risk of obtaining spurious results (Voelkl et al., 2018). Increasing heterogeneity comprises of using animals of different genetic backgrounds from diverse environments, animals of different species, strain, sex, and age. Recognising heterogeneity is not just necessary for reproducibility but is fundamental to reflecting the clinical heterogeneity of pain patients: the differences in terms of underlying disease, clinical presentation, pain mechanisms and different responses to treatment. Increased heterogeneity can be achieved by active and controlled heterogeneity within a lab or by conducting multi-centre experiments which will capitalise on the natural differences between labs e.g., Wodarski et al. (2016) demonstrated the multi-centre approach to evaluating suppressed

burrowing as robust and reproducible and therefore have supported the use of burrowing as an outcome to infer the global effect of pain on rodents.

9.4. Benefits of Systematic Reviews

Both the VGF and Cannabis & Pain reviews, consistent with prior reviews, have highlighted fundamental mismatch between preclinical research and the patient population. This mismatch has scientific, economic, and ethical costs. To improve the value of research and reduce research waste we need to be more strategic in how we use research to inform future research; systematic reviews and meta-analyses are the vital research tools for the job. They form the framework for the assimilation and critical appraisal of large volumes of primary research. They can provide evidence to inform a variety of decisions from identifying knowledge gaps, to improving methodological quality and translation of animal trial data to the clinic. Of note, they can also determine whether there is already enough data on a topic thereby nullifying the need for further experimental investigation. The table below, using the key findings from the Cannabis & Pain review (discussed in more detail in Chapter 5) as an example, is a summary that demonstrates the broad utility of systematic reviews.

Table 9.3. Utility of preclinical systematic reviews and meta-analysis based upon the Cannabis & Pain review.

Research Question Type	Empirical evidence	Recommendations
Provide an overview of the available evidence	Summarised and critically praised all available and relevant literature to answer the research questions	Only possible using systematic review methods. Improve reporting especially PICO elements in the title abstract to ensure that all relevant literature can be identified, synthesised, and appraised using rigorous systematic review methods
Determine efficacy of intervention(s)	The overall behavioural data effect sizes are significant, and there is a spectrum of efficacy dependent on drug and pain model type, species, strain, sex, and other experimental factors.	Not possible to identify highest priority/most promising candidates. Prospective experiments are required taking into consideration study design characteristics
Identify knowledge gaps	Disconnect between clinical and preclinical research e.g., different drug types and the external validity of animal models and outcome measures Low performance of pharmacokinetic studies in conjunction with pain-associated behavioural assessment Low performance of studies to assess side effects of potential treatments e.g., on motor activity, anxiety- and depression-like behaviours	Improve understanding of mechanisms e.g., pharmacokinetic and pharmacodynamic studies, a broader range of behavioural assessments to assess effects on motor, anxiety, and depression Improve external validity of animal models e.g., use both male and female models, disease models, appropriate behavioural outcomes and multi-centre testing to increase heterogeneity and therefore generalisability
Improve methodological quality	Low paucity to report in sufficient detail methods used to reduce the risk of bias	Improve rigor and transparency. Conduct and report measures to reduce experimental bias e.g., preregistration of protocols and report in accordance with the ARRIVE guidelines

Identify factors influencing treatment efficacy	Model, species, strain, sex, and outcome measure are study design characteristics that influenced the observed treatment efficacy	Care required. Researchers should look to increase external validity and biological variation to improve likelihood of generalisability of results
Inform experimental design of new studies	(From above)	Increase biological variation Reduce risks of bias and power studies appropriately Close the gap between clinical and preclinical research Perform pharmacokinetic, motor, anxiolytic, and depression assessments in conjunction with pain-associated behavioural outcomes
Translation of animal data to the clinic	For example, FAAH inhibitors were shown to be least effective in osteoarthritis models	Perform prospective systematic review and meta-analysis of the available preclinical data before proceeding to clinical trial
Implementation of the 3Rs	Research questions can be answered without the need for new experiments and data from the review can be used e.g., to perform power analyses and ensures the ethical use of animals	Perform prospective systematic review and meta-analysis of the available preclinical data before planning a new experiment/study
Prevent research waste	In addition to the above, data can be used for futility analyses to determine stable effects	Perform prospective systematic review and meta-analysis of the available preclinical data

9.5. The Feasibility Challenges of Systematic Reviews

When assessing what is already known in a field, justifying and designing experiments, and making evidence-based funding and policy decisions, it is essential that scientists, funders and policy makers can assimilate this volume of rapidly accruing data in a comprehensive and unbiased manner. Searching the term "pain" in PubMed for articles up to and including 31/12/2020 returns 865,565 articles. It is simply not possible for an individual to integrate research findings in an organised manner.

The 'evidence synthesis' problem could not have been better characterised than in response to the COVID-19 pandemic. The pandemic highlighted the weaknesses in the production and use of research-based evidence (Pearson, 2021) and the limited ability to keep up and make sense of the tens of thousands of primary research and systematic review articles that have been published in relation to COVID-19. Many of which containing data that could inform policy, practise, and future research decisions. Additionally, a recent review of the evidence synthesis articles, that were published at pace in response to the pandemic e.g., systematic review, rapid review, overview, meta-analysis, or qualitative evidence synthesis, had considerable overlap, were of low quality, lacked robust methods, and had discordant findings which ultimately undermines trust (Abbott et al., 2021). Arguably we did not/do not have the research capacity or infrastructure in place to do all the systematic reviews and meta-analyses that were required to respond to the pandemic and robustly assimilate the evidence as it became available to make timely decisions.

There are several challenges associated with systematically reviewing all the literature. The processes are labour intensive and time-consuming. Depending on the research question(s) and the size of the review there are several rate limiting stages e.g., study selection, full text retrieval, annotation, and data extraction, and despite the increased ease provided by systematic review platforms the stages still require considerable labour resource to complete. The need in both the clinical and preclinical domains for timely evidence synthesis is driving development of automation technologies to improve the feasibility, accuracy, and efficiency of these systematic review stages. The use of text mining tools and machine learning algorithms are becoming an increasingly popular approach to reduce human effort, cost, and time to complete. These developments are currently being assessed in both clinical and preclinical systematic reviews to determine their effectiveness and appropriate uses (Howard et al., 2016b, Tsafnat et al., 2014, O'Mara-Eves et al., 2015, Cramond et al., 2019, Bannach-Brown et al., 2018). In response to the COVID-19 pandemic several groups started to use novel tools and methods for evidence synthesis e.g., establishing databases, combining human effort including crowd sourcing and machine learning to identify, classify and map COVID-19 research (Voorhees et al., 2021, Shemilt et al., 2021, Keenan et al., 2021, Hair, 2021, Noel-Storr et al., 2021c). The aim of developing and using

such technologies is to facilitate the development of living systematic reviews, systematic reviews in which new evidence is incorporated into the review as and when it becomes available (Elliott et al., 2017, Thomas et al., 2017). Living systematic reviews will ensure that decisions are dynamic and based upon the full body of evidence. Most machine learning and text mining technologies have been developed to improve the feasibility and efficiency of the early stages of the review process e.g., search, study selection and annotation. For example, the CAMARDES COVID SOLES project developed and integrated a series of automation tools and methodologies to enable continual collection, categorisation and quality assessment of COVID-19 primary research to enable the timely synthesis of COVID-19 evidence for use by researchers and policy makers (Hair, 2021). However, automation tools can be used without integration and making the review living as was the case in this PhD, a living search was instigated for the VGF review and machine learning for the Cannabis & Pain review study selection.

9.5.1. Machine Learning for Study Selection

Machine learning algorithms have been applied to systematic review processes in several different ways e.g., study selection (Bannach-Brown et al., 2019), study classification (Wallace et al., 2017) and for annotation (Bahor et al., 2017). The aim in the Cannabis & Pain review was to use machine learning to facilitate the screening of articles retrieved by the systematic search and improve the accuracy of screening. Screening for study selection is a very time-consuming labour-intensive process. Searches can return tens of thousands of potentially relevant articles. In the chemotherapy induced peripheral neuropathy systematic review they report that screening in 2012 took 18 person months (Currie et al., 2019). In the cannabis review, the crowd of 36 scientists took 6 weeks to complete the screening which equates to 11 person weeks (Soliman et al., 2021). The findings from the Cannabis & Pain systematic review are commensurate with the study by Liao et al. (2018). They compared the performance of machine learning algorithms for screening of preclinical literature. It highlights how performance will vary based upon complexity of the inclusion/exclusion criteria, the size of the training dataset; with a positive correlation between dataset size and performance, inclusion prevalence and the type of the machine learning classifier (Liao et al., 2018).

The use of machine learning for study selection during the Cannabis & Pain review did not lead to a reduction in workload as initially hoped. It was not possible to train the machine algorithm to perform with the required sensitivity and specificity until all the retrieved articles had been manually screened. Its utility was demonstrated through the error analysis which aimed to ensure that we identified all the relevant literature. It resulted in the inclusion of 129 wrongly excluded studies (discussed in detail in Chapter 7). A challenge is the size of the training dataset that is required for the machine to learn

and perform favourably. This is a problem echoed by Hair et al., 2021 in The CAMARADES COVID-SOLES project. The relevant COVID-19 studies' objectives and methodologies were annotated by a crowd and the human decisions will be used to train machine learning classifiers. Yet, despite having 1,174 fully annotated primary studies, the number is too low to train classifiers to identify specific objectives or study methodologies and they are not yet able to classify research automatically (Hair, 2021). To reach similar machine performance as the Cannabis & Pain review, the chemotherapy induced peripheral neuropathy preclinical systematic review, used a larger dataset of 33,814 articles (10,816 for Cannabis & Pain) with known status for 'include' or 'exclude' to train the machine. There was also a larger inclusion prevalence of 19 % (3 % for Cannabis & Pain)(Currie et al., 2019). Currie et al. (2019) were then able to successfully demonstrate the utility of machine learning for study selection and used the algorithm to reduce the labour resource and time required for screening.

Such approaches are also being utilised for categorisation and annotation to further reduce the human workload. Text analytic approaches have been used to automatically ascertain the reporting of measures to reduce the risk of bias in preclinical stroke research (Bahor et al., 2017). A recent study evaluating bias in clinical trials found that using a semi-automated approach; the machine presented possible results to the reviewer, resulted in a 25% reduction in the time taken for risk of bias assessment, compared to the fully manual approach (Soboczenski et al., 2019).

Future research to determine most appropriate and reliable utility of machine learning algorithms and text mining is still required (discussed in detail in Chapter 7), but as is being demonstrated by the CAMARADES COVID SOLES project future projects, systematic reviews or living systematic reviews, will include the integration of machine learning and text mining tools which support screening and annotation into online review platforms with crowdsourced decisions and annotations performed in the context of reviews contributing to the pool data available for tool development.

9.5.2. Crowd Science

The Cannabis & Pain review used crowd science for the most labour and resource intensive processes of the review, screening and annotation and data extraction, in the anticipation that doing so would improve the efficiency of the systematic review process. Throughout Chapter 6 I made comparisons to citizen science projects including Cochrane Crowd however, there are key differences in the demographics of the crowd and what was asked of them. The Cannabis & Pain review employed scientists who volunteered their time, and their specialist skillset was required. These factors pose their own unique challenges to the recruitment, training, and contribution of scientists to improve efficiency of systematic reviews. It is without doubt that employing a crowd spreads the workload and reduces individual input requirements. Additionally, relying on volunteer effort provided a

considerable saving and represents a viable strategy for performing research that would otherwise exceed modest research budgets. However, efficiency can be undermined by reliance on a volunteer force. The Cannabis & Pain crowd did not save time. The crowd had substantial agreement for study level questions, a proxy for accuracy, but it was not possible to ascertain their overall accuracy.

Using a crowd creates opportunities to train and evaluate automation technologies for evidence synthesis (Hair, 2021, Noel-Storr et al., 2021a). Manual classifications are time-consuming and labour intensive whether they are being developed for screening against an inclusion/exclusion criteria or for categorising research. Machine learning algorithms are dependent on training and validation datasets: humans and machines work together to train, calibrate, and validate machine learning algorithms. The results is an automation technology that can significantly reduce the human effort required e.g. the Cochrane Crowd assisted in the development of the Cochrane Randomised Controlled Trial classifier and it's estimated to reduce screening burden by 20 – 25 % (Thomas et al., 2021b).

A crowd approach will be necessary to achieve the aims of the randomised controlled trial of the Graph2Data application (discussed in detail in Chapter 8). Having a diverse range of reviewers with varying systematic review experience plus the ability to trial across many studies will ensure that we identify the benefits of semi-automated data extraction, identify further user needs and inform future development including how to further reduce human input e.g., identification of graphs as potentially relevant or not relevant or identification of graph axes and their values (Cramond et al., 2019).

The way the Cannabis & Pain crowd was employed was more high risk than that of the Cochrane Crowd and demonstrates what can be achieved with close project management and risk mitigations. The Cochrane Crowd are asked to classify studies (Noel-Storr et al., 2021b) whereas the Cannabis & Pain crowd were asked to understand complex information and abstract considerable information from studies which often tested multiple drugs, of different doses and routes of administration, in multiple models, with efficacy assessed in several behavioural assessments. The biomedical science expertise of the Cannabis & Pain crowd was a necessity. In response to the COVID-19 pandemic the Cochrane Crowd's tasked were adapted and extended to include classifying COVID-19 studies, identifying studies for inclusion in rapid reviews and categorisation by PICO elements (Noel-Storr et al., 2021c). The crowd were described as invaluable in enabling Cochrane to keep up with the rapidly accruing evidence based. Noel-Storr et al. (2021c) explain that the COVID-19 pandemic accelerated their development of crowd sourcing tasks and advanced their expectations of crowdsourced capability for evidence synthesis. However, the complexity of the task and the time commitment required are likely important influencing factors of crowd success. There is a balance to be struck; ask

too much and risk losing motivated and competent individuals and/or risk inaccuracy as peoples' limits are stretched. Effective employment of a crowd requires supportive and accessible infrastructure and technical support, training, consideration of crowd needs including rewards and incentives, and good lines of communication. The Cannabis & Pain crowd were invaluable to the systematic review of the preclinical literature in which cannabinoids, cannabis-based medicines, and endocannabinoid system modulators were assessed for efficacy in animal models of injury-related or persistent pain (Soliman et al., 2021).

The Cannabis & Pain review demonstrated that employment of a crowd is more than improving the efficiency of a systematic review. An important strength of translational pain research is collaboration across the many different disciplines involved and using crowd science can create opportunities for the sharing of methodological and mechanistic insights between the different disciplines for forward and back translation. It has also been posited that for standards to be widely adopted and guidelines to be effective community consensus and alignment is needed (Freedman and Gibson, 2015). Crowd scientists reported that their critical analysis skills improved and enhanced their professional skillset. It has led them to improve their own research practises including an improved understanding of animal study design, conduct and reporting and experimental bias, including mitigation and assessment of study quality. Therefore, employing crowds of scientists in preclinical systematic reviews offer the opportunities to raise awareness of the extent of the problems and how they can be improved; they also offer the opportunity to identify areas where standards are most needed, what they should be and how maximum benefit can be provided; systematic reviews can also be used to assess the effectiveness of standards/guidelines for the continued development of policy and practise.

9.6. Limitations

9.6.1. Systematic Reviews

The gold standard prescribed methods were followed throughout the Cannabis & Pain review, but systematic reviews are vulnerable to bias due to the decisions made throughout the process. Of note, are the appropriateness of meta-analytical techniques and the strength of conclusions that are made. Preclinical systematic reviews should be considered observational research and therefore hypothesis generating, and so the findings related to experimental design should be confirmed in prospective studies. This is particularly pertinent given that systematic reviews can only be performed on the literature and data that is publicly available. The quality of the review findings will be influenced by the quality of the original studies.

The VGF review was performed by a single reviewer, whereas in the Cannabis & Pain review the considered gold standard of performing study selection and annotation and data extraction in

duplicate with discrepancies reconciled by a third independent reviewer was conducted. The VGF review was updated for this thesis however it has been over 2 years since the search for the Cannabis & Pain review systematic review was conducted. An update of the search may lead to the identification of new studies that reveal new and different insights. Given the size of the review, although uncertain, it is likely that new findings will have a minimal impact. If the review were to be updated, the trained machine learning algorithm could be used to reduce the screening burden, either by acting as the second screener or by instigating a cut-off and only screen a percentage of the studies that are most likely to be included. However, a future reviewer will need to be aware of concept drift; this is that the literature may change over time e.g., a new model type or treatment that was not present in the earlier references or database indexing and article retrieval changes. Continued performance testing will be required, updates to the training dataset necessary so as to reduce impact of these evolutionary changes (Liao et al., 2018).

Both reviews show that there was a paucity of the reporting of methodological quality criteria and therefore the reported data are at an unclear risk of bias. Therefore, the conclusions based upon this data have an unclear risk of bias that may affect the precision and reliability of estimates of effect. It is not known whether the researchers performed measures to mitigate bias but didn't report them, and for those that did report the measures, they were often not reported in enough detail to assess their appropriateness. Unlike clinical systematic reviews, it would not have been appropriate to limit the review and meta-analysis to only high-quality studies because most studies did not sufficiently report methodological quality criteria. It should also be noted that current risk of bias assessments for preclinical research are based upon expert opinion rather than evidence-based methods and there may be other factors that have not been assessed that can influence the reliability of the results and conclusions.

In the Cannabis & Pain field the between-study heterogeneity was high. It was not possible to account for all sources of heterogeneity. Studies were grouped together based upon their characteristics but there was undoubtedly the grouping of vastly different characteristics including the drug class and dosing regime, model type, species, strain, sex, and outcomes. This limits the usefulness of the summary estimates, but exploration of the heterogeneity has helped to characterise differences in efficacy and demonstrate the effect of study design characteristics.

Lastly, it was only possible to review the evidence as reported in publications. Details omitted from the published record may be influencing the observed outcome measures, but they are not possible to assess. In addition, the results may be susceptible to publication bias depriving researchers of null

or negative data. There is a critical and ethical need for transparency of reporting all experimental details to complete the story (Sena and Macleod, 2007).

9.6.2. Investigation of Crowd Science and Machine Learning

I have demonstrated that for study selection crowd science and machine learning can be used to reduce the time it takes to perform that stage and improve accuracy. However, crowd science alone did not reduce the overall time that it takes to perform a review and overall accuracy could not be determined. The crowd were varied in their experience of systematic reviews and knowledge of the topic. Errors by crowd members were detected during the reconciliation process and ~20% of studies required re-review by expert reviewers. The limitations of individual crowd members therefore will not have undermined the findings of the Cannabis & Pain review. The crowd manually screened all articles identified by the systematic search. The inability to train the machine following manual screening of a sub-set of studies was likely due to the low prevalence of inclusion. This can be indicative of an over inclusive search string. However, in this context it was particularly challenging to narrow the search terms further, not only due to wanting to identify all relevant literature but because it was difficult to identify the population due to the challenge of differentiating studies in which an intervention(s) was tested for antinociception in healthy animals, animal models of acute pain or animal models of injury-related or pathological persistent pain (the desired population). Importantly, this review has also demonstrated the usefulness of machine learning for study selection, albeit for error analysis. The machine algorithm performed with an eventual sensitivity of 95 % and specificity of 94 %. This high sensitivity made it unlikely to miss relevant literature.

Both approaches require further development and optimisation for their use in future preclinical systematic reviews. The way in which these methods were explored for improving feasibility was exploratory and therefore the findings susceptible to bias. Future preregistered research and trials are required to confirm findings and advance understanding of how best to employ such methods. However, this experience supports a recent Cochrane study which demonstrates the feasibility of study identification for inclusion in a systematic review using crowd science and machine learning (Noel-Storr et al., 2020). The continuing development of online platforms and emerging automation technologies will improve the feasibility of preclinical systematic reviews and create the possibility to incorporate the most recent data as it becomes available in the form of a living systematic review Elliot et al., 2017).

9.7. Conclusion: Future Research & Future of Systematic Reviews

A lot of research has and is being conducted amassing data to improve our understanding of the pathophysiology of pain and how we may treat it. In some contexts, the aetiology and pathophysiology remain elusive and the pharmacological treatments for those effected provide only limited efficacy and are associated with unwanted side effects. Preclinical research is essential to the protection of participants or patients in clinical trials and therefore it must be conducted to the highest standards of rigor. Translational failures can be traced in part to how we select and use preclinical tools and the way in which we report our findings. The mismatch between preclinical and clinical research focuses the need to be more strategic in the way in which we use research findings to determine what to do next and prevent unnecessary harm and research waste.

To address the need to systematically synthesise preclinical evidence, systematic reviews offer the methodological framework to combine and integrate all research claims to inform decisions. However, the exponentially increasing volume and complexity of preclinical evidence, coupled with the resource requirements of systematic reviews present feasibility challenges to overcome. Improving feasibility should increase uptake of preclinical systematic reviews and will provide more comprehensive understandings of the evidence base.

The results of this PhD support the notion that crowd science and automation technologies can be employed to improve the feasibility, efficiency, and accuracy of preclinical systematic reviews. Crowd science can reduce the individual workload and increase efficiency. It offers the opportunity to bring together researchers from across different fields of pain and even more broadly, increasing the diversity and democratisation of knowledge production while creating potential for cross-cutting, multi-disciplinary collaborative working and more efficient problem solving. Machine learning and text mining technologies are force multipliers. They offer the opportunity to make searches more sensitive or maintain the sensitivity but reduce the resource required for screening. They are also being developed for classifying and categorising research as well as for risk of bias annotations. These methodological and technological solutions will have limited utility unless improvements are made to the way in which research is designed, conducted, analysed, and reported.

There are two aims of crowd science; to perform systematic reviews in the most efficient manner possible (scientific efficiency) and to allow a broad community to engage with the research (social inclusivity); the relationship between these aims is nuanced and may not be aligned. More research is required to understand those who contribute to crowd science, what their motivations are and what types of incentives and rewards are of value to them. It is necessary to identify the project characteristics that are required to ensure high recruitment, retention, and diligent contribution.

Importantly, given the secondary benefits understanding the impact of participating in systematic reviews on researchers and their research may provide information to broaden participation in systematic reviews concurrent to improving research quality.

The development and incorporation of automation technologies and machine-assisted tools in open-source, online systematic review software like SyRF will reduce adoption barriers and facilitate the production of accurate and timely evidence synthesis. Further research is required to determine their validity and best uses e.g., use validated datasets to evaluate machine learning algorithms for study selection, assess the impact of losing relevant studies that are not correctly classified, understand how search sensitivity and specificity can impact machine learning algorithms and evaluate the use of crowd science in combination with machine learning. As new online platforms and tools for machine learning and automation become available, preclinical living systematic reviews become more feasible (Elliott et al., 2017).

Often discussed at the meetings of the International Collaboration for the Automation of Systematic Reviews (ICASR) is that research and development efforts should be focused on the eventual redundancy of systematic reviews. Research will no longer be communicated using the journal article, but information flow will be improved through interoperable standards and tools. Research, including all the experimental details i.e., model, species, strain, sex, interventions tested, and the outcomes measured and the outcome data, as well as methodological quality criteria, will be automatically indexed and curated in a central repository. To answer a specific research question all that will be needed is to select the parameters and a meta-analysis will be performed at the hit of a button. This means that new research will always be informed by previous knowledge and documented accordingly, and new data shared and added to the cumulating evidence base. Decisions will always be up to date because it will not need synthesising but will reflect the status of the current knowledge. Currently, automation technologies and tools are being developed to address the challenges resultant of the current communication system. However, for systematic reviews to become redundant a much larger cultural shift and reform is required to how we produce and share knowledge.

Until such time that systematic reviews become redundant, it is necessary to continue to improve the infrastructure to improve the feasibility, efficiency, and accuracy of systematic reviews. Only then will more researchers be able to perform systematic reviews but it also imperative to understand what the barriers are to increased uptake (Figure 9.1). Systematic reviews have proven their utility in clinical research and fundamentally revolutionised the conduct and reporting of clinical trials. In the preclinical space their utility is increasingly being demonstrated and given the large volume of primary research and the research waste and translational challenges is it now time for new policies and

further funding support (Macleod et al., 2014). It has been suggested that policy and incentive tend to follow infrastructure, experience, and norms, on the basis that a policy intervention without infrastructure and support is likely to fail (Nosek et al., 2021). Are we now at the tipping point?

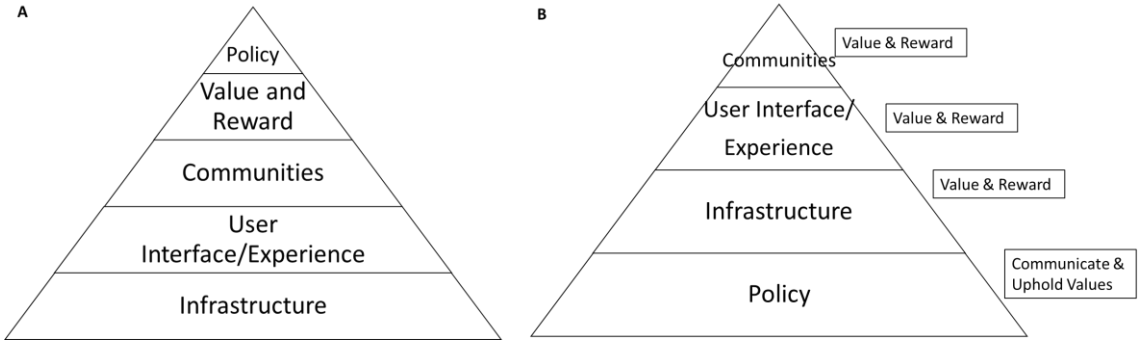


Figure 9.1. Improving the feasibility, accuracy, and efficiency for the potential of preclinical systematic reviews to be realised. (A) This pyramid reflects our current research culture. The first step is making systematic reviews possible with infrastructure e.g., SyRF and development of automated technologies including Graph2Data for data extraction from graphs and text mining and machine learning for screening and annotation. Second, is make it easy. The use of new technologies will be easier if integrated into a systematic review platform with a user-friendly interface thereby negating the need for subject matter expertise to use. Thirdly, performing preclinical systematic reviews need to become normative. Building a community through crowd science can create and facilitate the opportunities to share, learn and perform systematic reviews together. Fourth, it needs to be rewarding or rewarded. As the uptake increases, more researchers will be able to shape their use and gain to inform their research practises. Concurrently, development of new technologies will increase. Concomitantly, if we researchers say that we value being able to synthesise evidence in a timely manner and identify research conducted with the highest rigour from which we can build upon it will lead to further reward through scientific progress. For the last criteria, policy; the conservative approach is that policy comes once all other steps have been developed. However, this will come at a continued research waste cost. (B) A more radical approach is to have policy on the bottom. Make preclinical systematic reviews a condition of funding. Ensure funding is made available for their conduct and continued development. It is likely that doing so will accelerate development of new technology to improve infrastructure etc. and increase uptake. Similarly, each step of the pyramid should be valued and rewarded.

It is my aspiration to build a new community for pain research that comprises of primary researchers who perform research synthesis with support from systematic reviewers, librarians, and statisticians. Primary researchers will use systematic reviews to generate hypotheses and inform future research design. This will ensure that systematic reviews address research questions of importance and have the necessary subject matter expertise. It will also ensure that the final report is accessible to other primary researchers so that they can incorporate the findings into their research planning, conduct, analysis, and reporting. Evidence synthesis will also be recognised as an end goal of the research

process. Research will be designed, conducted, analysed, and reported accordingly (thereby mitigating biases and reducing research waste (Seidler et al., 2019).

Conducting systematic reviews will enable the continued development of automation technologies to automate several systematic review processes so that research questions can be addressed at a much lower effort and cost. Conducting systematic reviews will also lead to improvements in education, practice, and communication of pain research, improve the predictive validity of animal research and reduce research waste, to ultimately lead to improved pain outcomes for patients.

References

.

ABBOTT, R., BETHEL, A., ROGERS, M., WHEAR, R., ORR, N., SHAW, L., STEIN, K. & THOMPSON COON, J. 2021. Characteristics, quality and volume of the first 5 months of the COVID-19 evidence synthesis infodemic: a meta-research study. *BMJ Evidence-Based Medicine***,** bmjebm-2021-111710.

ADEVA, J. J. G., ATXA, J. M. P., CARRILLO, M. U. & ZENGOTITABENGOA, E. A. 2014. Automatic text classification to support systematic reviews in medicine. *Expert Systems with Applications,* 41**,** 1498-1508.

ALABAS, O. A., TASHANI, O. A. & JOHNSON, M. I. 2013. Effects of ethnicity and gender role expectations of pain on experimental pain: A cross-cultural study. *European Journal of Pain,* 17**,** 776-786.

ALDER, J., THAKKER-VARIA, S., BANGASSER, D. A., KUROIWA, M., PLUMMER, M. R., SHORS, T. J. & BLACK, I. B. 2003. Brain-derived neurotrophic factor-induced gene expression reveals novel actions of VGF in hippocampal synaptic plasticity. *Journal of Neuroscience,* 23**,** 10800-10808.

ALDINGTON, D. & ECCLESTON, C. 2019. Evidence-Based Pain Management: Building on the Foundations of Cochrane Systematic Reviews. *American journal of public health,* 109**,** 46-49.

ALLEN, L., O'CONNELL, A. & KIERMER, V. 2019. How can we ensure visibility and diversity in research contributions? How the Contributor Role Taxonomy (CRediT) is helping the shift from authorship to contributorship. *Learned Publishing,* 32**,** 71-74.

ANANIADOU, S., KELL, D. B. & TSUJII, J. 2006. Text mining and its potential applications in systems biology. *Trends in Biotechnology,* 24**,** 571-579.

ANDREWS, N., LEGG, E., LISAK, D., ISSOP, Y., RICHARDSON, D., HARPER, S., PHEBY, T., HUANG, W., BURGESS, G., MACHIN, I. & RICE, A. S. C. 2012. Spontaneous burrowing behaviour in the rat is reduced by peripheral nerve injury or inflammation associated pain. *European Journal of Pain,* 16**,** 485-495.

ANDREWS, N. A., LATREMOLIERE, A., BASBAUM, A. I., MOGIL, J. S., PORRECA, F., RICE, A. S. C., WOOLF, C. J., CURRIE, G. L., DWORKIN, R. H., EISENACH, J. C., EVANS, S., GEWANDTER, J. S., GOVER, T. D., HANDWERKER, H., HUANG, W., IYENGAR, S., JENSEN, M. P., KENNEDY, J. D., LEE, N., LEVINE, J., LIDSTER, K., MACHIN, I., MCDERMOTT, M. P., MCMAHON, S. B., PRICE, T. J., ROSS, S. E., SCHERRER, G., SEAL, R. P., SENA, E. S., SILVA, E., STONE, L., SVENSSON, C. I., TURK, D. C. & WHITESIDE, G. 2016a. Ensuring transparency and minimization of methodologic bias in preclinical pain research: PPRECISE considerations. *Pain,* 157**,** 901-909.

ANDREWS, N. A., LATRÉMOLIÈRE, A., BASBAUM, A. I., MOGIL, J. S., PORRECA, F., RICE, A. S. C., WOOLF, C. J., CURRIE, G. L., DWORKIN, R. H., EISENACH, J. C., EVANS, S., GEWANDTER, J. S., GOVER, T. D., HANDWERKER, H., HUANG, W., IYENGAR, S., JENSEN, M. P., KENNEDY, J. D., LEE, N., LEVINE, J., LIDSTER, K., MACHIN, I., MCDERMOTT, M. P., MCMAHON, S. B., PRICE, T. J., ROSS, S. E., SCHERRER, G., SEAL, R. P., SENA, E. S., SILVA, E., STONE, L., SVENSSON, C. I., TURK, D. C. & WHITESIDE, G. 2016b. Ensuring transparency and minimization of methodologic bias in preclinical pain research: PPRECISE considerations. *PAIN,* 157**,** 901-909.

ASHRAF, S., BOUHANA, K. S., PHENEGER, J., ANDREWS, S. W. & WALSH, D. A. 2016. Selective inhibition of tropomyosin-receptor-kinase A (TrkA) reduces pain and joint damage in two rat models of inflammatory arthritis. *Arthritis Research & Therapy,* 18.

ATKINS, D., BEST, D., BRISS, P. A., ECCLES, M., FALCK-YTTER, Y., FLOTTORP, S., GUYATT, G. H., HARBOUR, R. T., HAUGH, M. C., HENRY, D., HILL, S., JAESCHKE, R., LENG, G., LIBERATI, A., MAGRINI, N., MASON, J., MIDDLETON, P., MRUKOWICZ, J., O'CONNELL, D., OXMAN, A. D., PHILLIPS, B., SCHÜNEMANN, H. J., EDEJER, T. T.-T., VARONEN, H., VIST, G. E., WILLIAMS, J. W., JR., ZAZA, S. & GROUP, G. W. 2004. Grading quality of evidence and strength of recommendations. *BMJ (Clinical research ed.),* 328**,** 1490-1490.

ATTAL, N. & BOUHASSIRA, D. 2019. Translational neuropathic pain research. *PAIN,* 160.

AVEY, M. T., MOHER, D., SULLIVAN, K. J., FERGUSSON, D., GRIFFIN, G., GRIMSHAW, J. M., HUTTON, B., LALU, M. M., MACLEOD, M., MARSHALL, J., MEI, S. H. J., RUDNICKI, M., STEWART, D. J., TURGEON, A. F., MCINTYRE, L. & CANADIAN CRITICAL CARE TRANSLATIONAL BIOLOGY, G. 2016. The Devil Is in the Details: Incomplete Reporting in Preclinical Animal Research. *PLOS ONE,* 11**,** e0166733.

BAHOR, Z., LIAO, J., CURRIE, G., AYDER, C., MACLEOD, M., MCCANN, S. K., BANNACH-BROWN, A., WEVER, K., SOLIMAN, N., WANG, Q., DORAN-CONSTANT, L., YOUNG, L., SENA, E. S. & SENA, C. 2021. Development and uptake of an online systematic review platform: the early years of the CAMARADES Systematic Review Facility (SyRF). *BMJ Open Science,* 5**,** e100103.

BAHOR, Z., LIAO, J., MACLEOD, M. R., BANNACH-BROWN, A., MCCANN, S. K., WEVER, K. E., THOMAS, J., OTTAVI, T., HOWELLS, D. W., RICE, A., ANANIADOU, S. & SENA, E. 2017. Risk of bias reporting in the recent animal focal cerebral ischaemia literature. *Clinical Science,* 131**,** 2525-2532.

BAKER, C. 2018. *wosr: Clients to the 'Web of Science' and 'InCites' APIs* [Online]. Available: https://CRAN.R-project.org/package=wosr [Accessed 31/03/21].

BAKER, D., LIDSTER, K., SOTTOMAYOR, A. & AMOR, S. 2014. Two Years Later: Journals Are Not Yet Enforcing the ARRIVE Guidelines on Reporting Standards for Pre-Clinical Animal Studies. *Plos Biology,* 12.

BANNACH-BROWN, A., HAIR, K., BAHOR, Z., SOLIMAN, N., M., M. & LIAO, J. 2021. Technological advances in preclinical meta-research. *BMJ Open Science,* 0**,** e100131.

BANNACH-BROWN, A., PRZYBYŁA, P., THOMAS, J., RICE, A. S. C., ANANIADOU, S., LIAO, J. & MACLEOD, M. R. 2018. The use of text-mining and machine learning algorithms in systematic reviews: reducing workload in preclinical biomedical sciences and reducing human screening error. *bioRxiv***,** 255760.

BANNACH-BROWN, A., PRZYBYŁA, P., THOMAS, J., RICE, A. S. C., ANANIADOU, S., LIAO, J. & MACLEOD, M. R. 2019. Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. *Systematic Reviews,* 8**,** 23.

BARBER, R. P., VAUGHN, J. E., SLEMMON, J. R., SALVATERRA, P. M., ROBERTS, E. & LEEMAN, S. E. 1979. ORIGIN, DISTRIBUTION AND SYNAPTIC RELATIONSHIPS OF SUBSTANCE-P AXONS IN RAT SPINAL-CORD. *Journal of Comparative Neurology,* 184**,** 331-351.

BARON, R., BINDER, A. & WASNER, G. 2010a. Neuropathic pain: diagnosis, pathophysiological mechanisms, and treatment. *Lancet Neurol,* 9**,** 807-19.

BARON, R., BINDER, A. & WASNER, G. 2010b. Neuropathic pain: diagnosis, pathophysiological mechanisms, and treatment. *Lancet Neurology,* 9**,** 807-819.

BARON, R., MAIER, C., ATTAL, N., BINDER, A., BOUHASSIRA, D., CRUCCU, G., FINNERUP, N. B., HAANPÄÄ, M., HANSSON, P., HÜLLEMANN, P., JENSEN, T. S., FREYNHAGEN, R., KENNEDY, J. D., MAGERL, W., MAINKA, T., REIMER, M., RICE, A. S. C., SEGERDAHL, M., SERRA, J., SINDRUP, S., SOMMER, C., TÖLLE, T., VOLLERT, J., TREEDE, R.-D., GERMAN NEUROPATHIC PAIN RESEARCH, N., THE, E. & CONSORTIA, N. 2017. Peripheral neuropathic pain: a mechanism-related organizing principle based on sensory profiles. *Pain,* 158**,** 261-272.

BARTOLOMUCCI, A., POSSENTI, R., MAHATA, S. K., FISCHER-COLBRIE, R., LOH, Y. P. & SALTON, S. R. J. 2011. The Extended Granin Family: Structure, Function, and Biomedical Implications. *Endocrine Reviews,* 32**,** 755-797.

BASTIAN, H., GLASZIOU, P. & CHALMERS, I. 2010. Seventy-Five Trials and Eleven Systematic Reviews a Day: How Will We Ever Keep Up? *Plos Medicine,* 7.

BATE, S. & KARP, N. A. 2014. A Common Control Group - Optimising the Experiment Design to Maximise Sensitivity. *PLOS ONE,* 9**,** e114872.

BEKHUIS, T. & DEMNER-FUSHMAN, D. 2010. Towards Automating the Initial Screening Phase of a Systematic Review. *Medinfo 2010, Pts I and Ii,* 160**,** 146-150.

BEKHUIS, T. & DEMNER-FUSHMAN, D. 2012. Screening nonrandomized studies for medical systematic reviews: A comparative study of classifiers. *Artificial Intelligence in Medicine,* 55**,** 197-207.
BEKHUIS, T., TSEYTLIN, E., MITCHELL, K. J. & DEMNER-FUSHMAN, D. 2014. Feature Engineering and a Proposed Decision-Support System for Systematic Reviewers of Medical Evidence. *Plos One,* 9.
BELUR, J., TOMPSON, L., THORNTON, A. & SIMON, M. 2021. Interrater Reliability in Systematic Review Methodology: Exploring Variation in Coder Decision-Making. *Sociological Methods & Research,* 50**,** 837-865.
BERGE, O. G. 2011. Predictive validity of behavioural animal models for chronic pain. *Br J Pharmacol,* 164**,** 1195-206.
BERKLEY, K. J. 1997. Sex differences in pain. *Behav Brain Sci,* 20**,** 371-80; discussion 435-513.
BERMAN, P., FUTORAN, K., LEWITUS, G. M., MUKHA, D., BENAMI, M., SHLOMI, T. & MEIRI, D. 2018. A new ESI-LC/MS approach for comprehensive metabolic profiling of phytocannabinoids in Cannabis. *Scientific Reports,* 8**,** 14280.
BIOSOFT. 2021. *Ungraph* [Online]. Available: http://www.biosoft.com/w/ungraph.htm [Accessed 20/05/21].
BOOS, L., CAMPBELL, L. L., AMES, R., WETSEL, R. A. & BARNUM, S. R. 2004. Deletion of the complement anaphylatoxin C3a receptor attenuates, whereas ectopic expression of C3a in the brain exacerbates, experimental autoimmune encephalomyelitis. *Journal of Immunology,* 173**,** 4708-4714.
BORENSTEIN, M., HEDGES, L. V., HIGGINS J.P.T. & ROTHESTEIN, H. R. 2009. *Introduction to Meta-Analysis,* United Kingdom, John Wiley & Sons, Ltd.
BORSOOK, D., HARGREAVES, R., BOUNTRA, C. & PORRECA, F. 2014. Lost but making progress--Where will new analgesic drugs come from? *Science translational medicine,* 6**,** 249sr3-249sr3.
BOUTRON, I., ALTMAN, D. G., HOPEWELL, S., VERA-BADILLO, F., TANNOCK, I. & RAVAUD, P. 2014. Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of cancer: the SPIIN randomized controlled trial. *J Clin Oncol,* 32**,** 4120-6.
BOUTRON, I., DUTTON, S., RAVAUD, P. & ALTMAN, D. G. 2010. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *Jama,* 303**,** 2058-64.
BOVE, G. 2006. Mechanical sensory threshold testing using nylon monofilaments: The pain field's “Tin Standard”. *PAIN,* 124**,** 13-17.
BOZDAGI, O., RICH, E., TRONEL, S., SADAHIRO, M., PATTERSON, K., SHAPIRO, M. L., ALBERINI, C. M., HUNTLEY, G. W. & SALTON, S. R. J. 2008. The neurotrophin-inducible gene Vgf regulates hippocampal function and behavior through a brain-derived neurotrophic factor-dependent mechanism. *Journal of Neuroscience,* 28**,** 9857-9869.
BOZDOGAN, H. 1987. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika,* 52**,** 345-370.
BRITO, R. G., ARAUJO, A. A. S., QUINTANS, J. S. S., SLUKA, K. A. & QUINTANS, L. J. 2015. Enhanced analgesic activity by cyclodextrins - a systematic review and meta-analysis. *Expert Opinion on Drug Delivery,* 12**,** 1677-1688.
BULLOCK, D., JESUTHASAN, A., GONZÁLEZ-CANO, R. & COSTIGAN, M. 2019. Reading and writing: the evolution of molecular pain genetics. *PAIN,* 160.
BUSCEMI, N., HARDING, L., VANDERMEER, B., TJOSVOLD, L. & KLASSEN, T. P. 2006. Single data extraction generated more errors than double data extraction in systematic reviews. *Journal of Clinical Epidemiology,* 59**,** 697-703.
BUSSE, S., STEINER, J., GLORIUS, S., DOBROWOLNY, H., GREINER-BOHL, S., MAWRIN, C., BOMMHARDT, U., HARTIG, R., BOGERTS, B. & BUSSE, M. 2015. VGF expression by T lymphocytes in patients with Alzheimer's disease. *Oncotarget,* 6**,** 14843-14851.

BUSSE, S., STEINER, J., MICHEEL, J., DOBROWOLNY, H., MAWRIN, C., KRAUSE, T. J., ADAMASZEK, M., BOGERTS, B., BOMMHARDT, U., HARTIG, R. & BUSSE, M. 2014. Age-related increase of VGF-expression in T lymphocytes. *Aging-Us,* 6**,** 440-453.

BUTTON, K. S., IOANNIDIS, J. P., MOKRYSZ, C., NOSEK, B. A., FLINT, J., ROBINSON, E. S. & MUNAFÒ, M. R. 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci,* 14**,** 365-76.

CALVO, M., DAVIES, A. J., HÉBERT, H. L., WEIR, G. A., CHESLER, E. J., FINNERUP, N. B., LEVITT, R. C., SMITH, B. H., NEELY, G. G., COSTIGAN, M. & BENNETT, D. L. 2019. The Genetics of Neuropathic Pain from Model Organisms to Clinical Application. *Neuron,* 104**,** 637-653.

CAMPBELL, J. N. & MEYER, R. A. 2006. Mechanisms of neuropathic pain. *Neuron,* 52**,** 77-92.

CAMPBELL, M., MCKENZIE, J. E., SOWDEN, A., KATIKIREDDI, S. V., BRENNAN, S. E., ELLIS, S., HARTMANN-BOYCE, J., RYAN, R., SHEPPERD, S., THOMAS, J., WELCH, V. & THOMSON, H. 2020. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. *BMJ,* 368**,** l6890.

CAMPBELL, R., POUND, P., POPE, C., BRITTEN, N., PILL, R., MORGAN, M. & DONOVAN, J. 2003. Evaluating meta-ethnography: a synthesis of qualitative research on lay experiences of diabetes and diabetes care. *Social Science & Medicine,* 56**,** 671-684.

CAPPA, F., LAUT, J., PORFIRI, M. & GIUSTINIANO, L. 2018. Bring them aboard: Rewarding participation in technology-mediated citizen science projects. *Computers in Human Behavior,* 89**,** 246-257.

CARROLL, C., SCOPE, A. & KALTENTHALER, E. 2013. A case study of binary outcome data extraction across three systematic reviews of hip arthroplasty: errors and differences of selection. *BMC Research Notes,* 6**,** 539.

CATTANEO, A., SESTA, A., CALABRESE, F., NIELSEN, G., RIVA, M. A. & GENNARELLI, M. 2010. The Expression of VGF is Reduced in Leukocytes of Depressed Patients and it is Restored by Effective Antidepressant Treatment. *Neuropsychopharmacology,* 35**,** 1423-1428.

CERO, C. & BARTOLOMUCCI, A. 2014. Beta-Adrenergic Receptors-Activation Is Required for the Pro-Lipolytic and Anti-Obesity Effect of the Vgf-Derived Peptide Tlqp-21. *Endocrine Reviews,* 35.

CERO, C., VOSTRIKOV, V. V., VERARDI, R., SEVERINI, C., GOPINATH, T., BRAUN, P. D., SASSANO, M. F., GURNEY, A., ROTH, B. L., VULCHANOVA, L., POSSENTI, R., VEGLIA, G. & BARTOLOMUCCI, A. 2014. The TLQP-21 Peptide Activates the G-Protein-Coupled Receptor C3aR1 via a Folding-upon-Binding Mechanism. *Structure,* 22**,** 1744-1753.

CHALMERS, I. 2005. Academia's failure to support systematic reviews. *Lancet,* 365**,** 469-469.

CHALMERS, I. & GLASZIOU, P. 2009. Avoidable Waste in the Production and Reporting of Research Evidence. *Obstetrics and Gynecology,* 114**,** 1341-1345.

CHAN, L., HALL, B., F, P., TANDON, R. & WILLIAMS, W. 2020. Open Science Beyond Open Access: For and with communities, A step towards the decolonization of knowledge. *zenodo*.

CHEN, S.-Y., FENG, Z. & YI, X. 2017. A general introduction to adjustment for multiple comparisons. *Journal of thoracic disease,* 9**,** 1725-1729.

CHEN, Y. C., PRISTERA, A., AYUB, M., SWANWICK, R. S., KARU, K., HAMADA, Y., RICE, A. S. C. & OKUSE, K. 2013. Identification of a Receptor for Neuropeptide VGF and Its Role in Neuropathic Pain. *Journal of Biological Chemistry,* 288**,** 34638-34646.

CHESLER, E. J., WILSON, S. G., LARIVIERE, W. R., RODRIGUEZ-ZAS, S. L. & MOGIL, J. S. 2002. Influences of laboratory environment on behavior. *Nat Neurosci,* 5**,** 1101-2.

CHO, K. O., SKARNES, W. C., MINSK, B., PALMIERI, S., JACKSONGRUSBY, L. & WAGNER, J. A. 1989. NERVE GROWTH-FACTOR REGULATES GENE-EXPRESSION BY SEVERAL DISTINCT MECHANISMS. *Molecular and Cellular Biology,* 9**,** 135-143.

CLARK, J. D. 2016. Preclinical Pain Research: Can We Do Better? *Anesthesiology,* 125**,** 846-849.

CLARKE, M. 2004. Doing new research? Don't forget the old. *PLoS Med,* 1**,** e35.

CLARKE, M., ALDERSON, P. & CHALMERS, I. 2002. Discussion sections in reports of controlled trials published in general medical journals. *Jama,* 287**,** 2799-801.

CLAYTON, J. A. & COLLINS, F. S. 2014. Policy: NIH to balance sex in cell and animal studies. *Nature,* 509**,** 282-3.
COCHRAN, W. G. 1954. Some Methods for Strengthening the Common χ 2 Tests. *Biometrics,* 10**,** 417.
COHEN, A. M., AMBERT, K. & MCDONAGH, M. 2012. Studying the potential impact of automated document classification on scheduling a systematic review update. *Bmc Medical Informatics and Decision Making,* 12.
COHEN, A. M., HERSH, W. R., PETERSON, K. & YEN, P. Y. 2006. Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association,* 13**,** 206-219.
COHEN, J. 1962. The Statistical Power of Abnormal-Social Psychological Research: A Review. *The Journal of Abnormal and Social Psychology,* 65**,** 145-153.
COLLINS, A. 2014. *Citizen Science in the Classroom: Assessing the Impact of an Urban Field Ecology Program on Learning Gains and Attitudes toward Science.* M.A. Master's Thesis, Columbia University.
COLLINS, J. A. & FAUSER, B. C. J. M. 2005. Balancing the strengths of systematic and narrative reviews. *Human Reproduction Update,* 11**,** 103-104.
COLQUHOUN, D. 2014. An investigation of the false discovery rate and the misinterpretation of p-values. *R Soc Open Sci,* 1**,** 140216.
COOK, D. J., MULROW, C. D. & HAYNES, R. B. 1997. Systematic reviews: synthesis of best evidence for clinical decisions. *Ann Intern Med,* 126**,** 376-80.
COOPER, S., KHATIB, F., TREUILLE, A., BARBERO, J., LEE, J., BEENEN, M., LEAVER-FAY, A., BAKER, D., POPOVIĆ, Z. & PLAYERS, F. 2010. Predicting protein structures with a multiplayer online game. *Nature,* 466**,** 756-760.
COSTIGAN, M., BEFORT, K., KARCHEWSKI, L., GRIFFIN, R. S., D'URSO, D., ALLCHORNE, A., SITARSKI, J., MANNION, J. W., PRATT, R. E. & WOOLF, C. J. 2002. Replicate high-density rat genome oligonucleotide microarrays reveal hundreds of regulated genes in the dorsal root ganglion after peripheral nerve injury. *Bmc Neuroscience,* 3.
COULL, J. A. M., BEGGS, S., BOUDREAU, D., BOIVIN, D., TSUDA, M., INOUE, K., GRAVEL, C., SALTER, M. W. & DE KONINCK, Y. 2005. BDNF from microglia causes the shift in neuronal anion gradient underlying neuropathic pain. *Nature,* 438**,** 1017-1021.
CRAFT, R. M., MOGIL, J. S. & ALOISI, A. M. 2004. Sex differences in pain and analgesia: the role of gonadal hormones. *European Journal of Pain,* 8**,** 397-411.
CRAMOND, F., O'MARA-EVES, A., DORAN-CONSTANT, L., RICE, A. S., MACLEOD, M. & THOMAS, J. 2019. The development and evaluation of an online application to assist in the extraction of data from graphs for use in systematic reviews. *Wellcome open research,* 3**,** 157-157.
CROSSLEY, N. A., SENA, E., GOEHLER, J., HORN, J., VAN DER WORP, B., BATH, P. M. W., MACLEOD, M. & DIRNAGL, U. 2008. Empirical evidence of bias in the design of experimental stroke studies - A metaepidemiologic approach. *Stroke,* 39**,** 929-934.
CROW, M. & DENK, F. 2019. RNA-seq data in pain research-an illustrated guide. *Pain,* 160**,** 1502-1504.
CUMMINGS, J. N. & KIESLER, S. 2014. Organization Theory and the Changing Nature of Science. *Journal of Organization Design,* 3**,** 1-16.
CURRIE, G. L., ANGEL-SCOTT, H. N., COLVIN, L., CRAMOND, F., HAIR, K., KHANDOKER, L., LIAO, J., MACLEOD, M., MCCANN, S. K., MORLAND, R., SHERRATT, N., STEWART, R., TANRIVER-AYDER, E., THOMAS, J., WANG, Q. Y., WODARSKI, R., XIONG, R., RICE, A. S. C. & SENA, E. S. 2019. Animal models of chemotherapy-induced peripheral neuropathy: A machine-assisted systematic review and meta-analysis. *Plos Biology,* 17.
DA SILVA, J. T. & SEMINOWICZ, D. A. 2019. Neuroimaging of pain in animal models: a review of recent literature. *Pain reports,* 4**,** e732-e732.
DAHAN, A., KEST, B., WAXMAN, A. R. & SARTON, E. 2008. Sex-Specific Responses to Opiates: Animal and Human Studies. *Anesthesia & Analgesia,* 107.

DAHLHAMER, J., LUCAS, J., ZELAYA, C., NAHIN, R., MACKEY, S., DEBAR, L., KERNS, R., VON KORFF, M., PORTER, L. & HELMICK, C. 2018. Prevalence of Chronic Pain and High-Impact Chronic Pain Among Adults - United States, 2016. *MMWR Morb Mortal Wkly Rep,* 67**,** 1001-1006.

DAVIS, K. D., AGHAEEPOUR, N., AHN, A. H., ANGST, M. S., BORSOOK, D., BRENTON, A., BURCZYNSKI, M. E., CREAN, C., EDWARDS, R., GAUDILLIERE, B., HERGENROEDER, G. W., IADAROLA, M. J., IYENGAR, S., JIANG, Y., KONG, J.-T., MACKEY, S., SAAB, C. Y., SANG, C. N., SCHOLZ, J., SEGERDAHL, M., TRACEY, I., VEASLEY, C., WANG, J., WAGER, T. D., WASAN, A. D. & PELLEYMOUNTER, M. A. 2020. Discovery and validation of biomarkers to aid the development of safe and effective pain therapeutics: challenges and opportunities. *Nature Reviews Neurology,* 16**,** 381-400.

DAWES, J. M., ANTUNES-MARTINS, A., PERKINS, J. R., PATERSON, K. J., SISIGNANO, M., SCHMID, R., RUST, W., HILDEBRANDT, T., GEISSLINGER, G., ORENGO, C., BENNETT, D. L. & MCMAHON, S. B. 2014. Genome-Wide Transcriptional Profiling of Skin and Dorsal Root Ganglia after Ultraviolet-B-Induced Inflammation. *Plos One,* 9.

DE VRIES, R. B. M., HOOIJMANS, C. R., LANGENDAM, M. W., VAN LUIJK, J., LEENAARS, M., RITSKES-HOITINGA, M. & WEVER, K. E. 2015a. A protocol format for the preparation, registration and publication of systematic reviews of animal intervention studies. *Evidence-based Preclinical Medicine,* 2**,** e00007.

DE VRIES, R. B. M., HOOIJMANS, C. R., TILLEMA, A., LEENAARS, M. & RITSKES-HOITINGA, M. 2014. Updated version of the Embase search filter for animal studies. *Laboratory Animals,* 48**,** 88-88.

DE VRIES, R. B. M., LEENAARS, M., TRA, J., HUIJBREGTSE, R., BONGERS, E., JANSEN, J. A., GORDIJN, B. & RITSKES-HOITINGA, M. 2015b. The potential of tissue engineering for developing alternatives to animal experiments: a systematic review. *Journal of Tissue Engineering and Regenerative Medicine,* 9**,** 771-778.

DE VRIES, Y. A., ROEST, A. M., DE JONGE, P., CUIJPERS, P., MUNAFÒ, M. R. & BASTIAANSEN, J. A. 2018. The cumulative effect of reporting and citation biases on the apparent efficacy of treatments: the case of depression. *Psychological Medicine,* 48**,** 2453-2455.

DENK, F., BENNETT, D. L. & MCMAHON, S. B. 2017. Nerve Growth Factor and Pain Mechanisms. *Annu Rev Neurosci,* 40**,** 307-325.

DENK, F. & MCMAHON, S. B. 2012. Chronic pain: emerging evidence for the involvement of epigenetics. *Neuron,* 73**,** 435-444.

DOOLEN, S., COOK, J., RIEDL, M., KITTO, K., KOHSAKA, S., HONDA, C. N., FAIRBANKS, C. A., TAYLOR, B. K. & VULCHANOVA, L. 2017. Complement 3a receptor in dorsal horn microglia mediates pronociceptive neuropeptide signaling. *Glia,* 65**,** 1976-1989.

DU SERT, N. P., BAMSEY, I., BATE, S. T., BERDOY, M., CLARK, R. A., CUTHILL, I. C., FRY, D., KARP, N. A., MACLEOD, M., MOON, L., STANFORD, S. C. & LINGS, B. 2017. The Experimental Design Assistant. *Nature Methods,* 14**,** 1024-1025.

DU SERT, N. P. & RICE, A. S. C. 2014. Improving the translation of analgesic drugs to the clinic: animal models of neuropathic pain. *British Journal of Pharmacology,* 171**,** 2951-2963.

DUVAL, S. & TWEEDIE, R. 2000. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics,* 56**,** 455-463.

DUYX, B., URLINGS, M. J. E., SWAEN, G. M. H., BOUTER, L. M. & ZEEGERS, M. P. 2017. Scientific citations favor positive results: a systematic review and meta-analysis. *J Clin Epidemiol,* 88**,** 92-101.

DWAN, K., GAMBLE, C., WILLIAMSON, P. R. & KIRKHAM, J. J. 2013. Systematic review of the empirical evidence of study publication bias and outcome reporting bias - an updated review. *PLoS One,* 8**,** e66844.

DWORKIN, R. H., TURK, D. C., FARRAR, J. T., HAYTHORNTHWAITE, J. A., JENSEN, M. P., KATZ, N. P., KERNS, R. D., STUCKI, G., ALLEN, R. R., BELLAMY, N., CARR, D. B., CHANDLER, J., COWAN, P., DIONNE, R., GALER, B. S., HERTZ, S., JADAD, A. R., KRAMER, L. D., MANNING, D. C., MARTIN,

S., MCCORMICK, C. G., MCDERMOTT, M. P., MCGRATH, P., QUESSY, S., RAPPAPORT, B. A., ROBBINS, W., ROBINSON, J. P., ROTHMAN, M., ROYAL, M. A., SIMON, L., STAUFFER, J. W., STEIN, W., TOLLETT, J., WERNICKE, J. & WITTER, J. 2005. Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. *PAIN,* 113**,** 9-19.

EGAN, K. J., VESTERINEN, H. M., BEGLOPOULOS, V., SENA, E. S. & MACLEOD, M. R. 2016. From a mouse: systematic analysis reveals limitations of experiments testing interventions in Alzheimer's disease mouse models. *Evidence-based preclinical medicine,* 3**,** e00015-e00015.

EGGER, M., SMITH, G. D., SCHNEIDER, M. & MINDER, C. 1997. Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal,* 315**,** 629-634.

EGGER, M., SMITH, G. D. & STERNE, J. A. C. 2001. Uses and abuses of meta-analysis. *Clinical Medicine,* 1**,** 478-484.

ELLIOTT, J. H., SYNNOT, A., TURNER, T., SIMMONDS, M., AKL, E. A., MCDONALD, S., SALANTI, G., MEERPOHL, J., MACLEHOSE, H., HILTON, J., TOVEY, D., SHEMILT, I., THOMAS, J. & LIVING SYSTEMATIC REVIEW, N. 2017. Living systematic review: 1. Introduction-the why, what, when, and how. *Journal of Clinical Epidemiology,* 91**,** 23-30.

ENGLAND, S., BEVAN, S. & DOCHERTY, R. J. 1996. PGE(2) modulates the tetrodotoxin-resistant sodium current in neonatal rat dorsal root ganglion neurones via the cyclic AMP-protein kinase A cascade. *Journal of Physiology-London,* 495**,** 429-440.

ERDOGAN, B. R., VOLLERT, J. & MICHEL, M. C. 2020. Choice of y-axis can mislead readers. *Naunyn-Schmiedeberg's Archives of Pharmacology,* 393**,** 1769-1772.

FAGGION, C. M., JR., BAKAS, N. P. & WASIAK, J. 2017. A survey of prevalence of narrative and systematic reviews in five major medical journals. *BMC medical research methodology,* 17**,** 176-176.

FAIRBANKS, C. A., PETERSON, C. D., SPELTZ, R. H., RIEDL, M. S., KITTO, K. F., DYKSTRA, J. A., BRAUN, P. D., SADAHIRO, M., SALTON, S. R. & VULCHANOVA, L. 2014. The VGF-derived peptide TLQP-21 contributes to inflammatory and nerve injury-induced hypersensitivity. *Pain,* 155**,** 1229-1237.

FARRELL, K. E., KEELY, S., GRAHAM, B. A., CALLISTER, R. & CALLISTER, R. J. 2014. A Systematic Review of the Evidence for Central Nervous System Plasticity in Animal Models of Inflammatory-mediated Gastrointestinal Pain. *Inflammatory Bowel Diseases,* 20**,** 176-195.

FAYAZ, A., CROFT, P., LANGFORD, R. M., DONALDSON, L. J. & JONES, G. T. 2016. Prevalence of chronic pain in the UK: a systematic review and meta-analysis of population studies. *BMJ Open,* 6**,** e010364.

FEDERICO, C. A., MOGIL, J. S., RAMSAY, T., FERGUSSON, D. A. & KIMMELMAN, J. 2020. A systematic review and meta-analysis of pregabalin preclinical studies. *PAIN,* 161**,** 684-693.

FELIZARDO, K. R., SALLEH, N., MARTINS, R. M., MENDES, E., MACDONELL, S. G., MALDONADO, J. C. & SOC, I. C. 2011. Using Visual Text Mining to Support the Study Selection Activity in Systematic Literature Reviews. *2011 Fifth International Symposium on Empirical Software Engineering and Measurement (Esem 2011)***,** 77-86.

FERNANDES, R. S., FERREIRA, D. D., FERREIRAC, C. D., GIAMMARILE, F., RUBELLO, D. & DE BARROS, A. L. B. 2016. Development of imaging probes for bone cancer in animal models. A systematic review. *Biomedicine & Pharmacotherapy,* 83**,** 1253-1264.

FERRI, G. L., LEVI, A. & POSSENTI, R. 1992. A NOVEL NEUROENDOCRINE GENE-PRODUCT - SELECTIVE VGF8A GENE-EXPRESSION AND IMMUNO-LOCALIZATION OF THE VGF PROTEIN IN ENDOCRINE AND NEURONAL POPULATIONS. *Molecular Brain Research,* 13**,** 139-143.

FERRI, G. L., NOLI, B., BRANCIA, C., D'AMATO, F. & COCCO, C. 2011. VGF: An inducible gene product, precursor of a diverse array of neuro-endocrine peptides and tissue-specific disease biomarkers. *Journal of Chemical Neuroanatomy,* 42**,** 249-261.

FERRI, G. L. & POSSENTI, R. 1992. *NEUROENDOCRINE EXPRESSION OF THE NERVE GROWTH FACTOR-INDUCED VGF PROTEIN*.

FILLINGIM, R. B., KING, C. D., RIBEIRO-DASILVA, M. C., RAHIM-WILLIAMS, B. & RILEY, J. L. 2009. Sex, Gender, and Pain: A Review of Recent Clinical and Experimental Findings. *Journal of Pain,* 10**,** 447-485.

FINN, D. P., HAROUTOUNIAN, S., HOHMANN, A. G., KRANE, E., SOLIMAN, N. & RICE, A. S. C. 2021. Cannabinoids, the endocannabinoid system, and pain: a review of preclinical studies. *PAIN,* Articles in Press.

FINNERUP, N. B., ATTAL, N., HAROUTOUNIAN, S., MCNICOL, E., BARON, R., DWORKIN, R. H., GILRON, I., HAANPÄÄ, M., HANSSON, P., JENSEN, T. S., KAMERMAN, P. R., LUND, K., MOORE, A., RAJA, S. N., RICE, A. S., ROWBOTHAM, M., SENA, E., SIDDALL, P., SMITH, B. H. & WALLACE, M. 2015. Pharmacotherapy for neuropathic pain in adults: a systematic review and meta-analysis. *Lancet Neurol,* 14**,** 162-73.

FINNERUP, N. B., OTTO, M., MCQUAY, H. J., JENSEN, T. S. & SINDRUP, S. H. 2005. Algorithm for neuropathic pain treatment: An evidence based proposal. *Pain,* 118**,** 289-305.

FINNERUP, N. B., SINDRUP, S. H. & JENSEN, T. S. 2010. The evidence for pharmacological treatment of neuropathic pain. *Pain,* 150**,** 573-581.

FISHER, E., MOORE, R. A., FOGARTY, A. E., FINN, D. P., FINNERUP, N. B., GILRON, I., HAROUTOUNIAN, S., KRANE, E., RICE, A. S. C., ROWBOTHAM, M., WALLACE, M. & ECCLESTON, C. 2020. Cannabinoids, cannabis, and cannabis-based medicine for pain management: a systematic review of randomised controlled trials. *PAIN,* Articles in Press.

FLEETWOOD-WALKER, S. M., QUINN, J. P., WALLACE, C., BLACKBURN-MUNRO, G., KELLY, B. G., FISKERSTRAND, C. E., NASH, A. A. & DALZIEL, R. G. 1999. Behavioural changes in the rat following infection with varicella-zoster virus. *Journal of General Virology,* 80**,** 2433-2436.

FORRESTER, T. D., BAKER, M., COSTELLO, R., KAYS, R., PARSONS, A. W. & MCSHEA, W. J. 2017. Creating advocates for mammal conservation through citizen science. *Biological Conservation,* 208**,** 98-105.

FOUNTOULAKIS, K. N., SAMARA, M. T. & SIAMOULI, M. 2014. Burning issues in the meta-analysis of pharmaceutical trials for depression. *Journal of Psychopharmacology,* 28**,** 106-117.

FRANZONI, C. & SAUERMANN, H. 2014. Crowd science: The organization of scientific research in open collaborative projects. *Research Policy,* 43**,** 1-20.

FREEDMAN, L. & GIBSON, M. 2015. The Impact of Preclinical Irreproducibility on Drug Development. *Clinical Pharmacology & Therapeutics,* 97**,** 16-18.

FREEDMAN, L. P., COCKBURN, I. M. & SIMCOE, T. S. 2015. The Economics of Reproducibility in Preclinical Research. *PLoS Biol,* 13**,** e1002165.

FRUNZA, O., INKPEN, D., MATWIN, S., KLEMENT, W. & O'BLENIS, P. 2011. Exploiting the systematic review protocol for classification of medical abstracts. *Artificial Intelligence in Medicine,* 51**,** 17-25.

FUKUOKA, T., KONDO, E., DAI, Y., HASHIMOTO, N. & NOGUCHI, K. 2001. Brain-derived neurotrophic factor increases in the uninjured dorsal root ganglion neurons in selective spinal nerve ligation model. *Journal of Neuroscience,* 21**,** 4891-4900.

FUKUOKA, T., TOKUNAGA, A., KONDO, E., MIKI, K., TACHIBANA, T. & NOGUCHI, K. 1998. Change in mRNAs for neuropeptides and the GABA(A) receptor in dorsal root ganglion neurons in a rat experimental neuropathic pain model. *Pain,* 78**,** 13-26.

GARRY, E. M., DELANEY, A., ANDERSON, H. A., SIRINATHSINGHJI, E. C., CLAPP, R. H., MARTIN, W. J., KINCHINGTON, P. R., KRAH, D. L., ABBADIE, C. & FLEETWOOD-WALKER, S. M. 2005. Varicella zoster virus induces neuropathic changes in rat dorsal root ganglia and behavioral reflex sensitisation that is attenuated by gabapentin or sodium channel blocking drugs. *Pain,* 118**,** 97-111.

GERDIN, A. K., IGOSHEVA, N., ROBERSON, L. A., ISMAIL, O., KARP, N., SANDERSON, M., CAMBRIDGE, E., SHANNON, C., SUNTER, D., RAMIREZ-SOLIS, R., BUSSELL, J. & WHITE, J. K. 2012. Experimental and husbandry procedures as potential modifiers of the results of phenotyping tests. *Physiol Behav,* 106**,** 602-11.

GETDATA. 2013. *GetData* [Online]. Available: http://getdata-graph-digitizer.com/download.php [Accessed 20/05/21].
GOLDBERG, D. S. & MCGEE, S. J. 2011. Pain as a global public health priority. *BMC Public Health,* 11**,** 770.
GRANT, M. J. & BOOTH, A. 2009. A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal,* 26**,** 91-108.
GREEN, B. N., JOHNSON, C. D. & ADAMS, A. 2006. Writing narrative literature reviews for peer-reviewed journals: secrets of the trade. *Journal of chiropractic medicine,* 5**,** 101-117.
GREENHALGH, T., THORNE, S. & MALTERUD, K. 2018. Time to challenge the spurious hierarchy of systematic over narrative reviews? *European Journal of Clinical Investigation,* 48**,** e12931.
GUINDON, J. & HOHMANN, A. G. 2009. The endocannabinoid system and pain. *CNS & neurological disorders drug targets,* 8**,** 403-421.
GULIN, J. E. N., ROCCO, D. M. & GARCÍA-BOURNISSEN, F. 2015. Quality of Reporting and Adherence to ARRIVE Guidelines in Animal Studies for Chagas Disease Preclinical Drug Research: A Systematic Review. *PLOS Neglected Tropical Diseases,* 9**,** e0004194.
GUO, J. B., CHEN, B. L., WANG, Y., ZHU, Y., SONG, G., YANG, Z., ZHENG, Y. L., WANG, X. Q. & CHEN, P. J. 2019. Meta-Analysis of the Effect of Exercise on Neuropathic Pain Induced by Peripheral Nerve Injury in Rat Models. *Frontiers in Neurology,* 10.
HAEUSSLER, C. & SAUERMANN, H. 2020. Division of labor in collaborative knowledge production: The role of team size and interdisciplinarity. *Research Policy,* 49.
HAIR, K. 2021. Building a Systematic Online Living Evidence Summary of COVID-19 Research. *Journal of EAHIL,* 17**,** 21-26.
HAIR, K., BAHOR, Z., MACLEOD, M., LIAO, J. & SENA, E. S. 2021. The Automated Systematic Search Deduplicator (ASySD): a rapid, open-source, interoperable tool to remove duplicate citations in biomedical systematic reviews. *bioRxiv***,** 2021.05.04.442412.
HAIR, K., MACLEOD, M. R., SENA, E. S., SENA, E. S., HAIR, K., MACLEOD, M. R., HOWELLS, D., BATH, P., IRVINE, C., MACCALLUM, C., MORRISON, G., CLARK, A., ALVINO, G., DOHM, M., LIAO, J., SENA, C., MORELAND, R., CRAMOND, F., CURRIE, G. L., BAHOR, Z., GRILL, P., BANNACH-BROWN, A., MARCU, D.-C., ANTAR, S., BLAZEK, K., KONOLD, T., DINGWALL, M., HOHENDORF, V., HOSH, M., GERLEI, K. Z., WEVER, K. E., JONES, V., QUINN, T. J., KARP, N. A., FREYMANN, J., SHEK, A., GREGORC, T., RINALDI, A., JHEETA, P., NAZZAL, A., HENSHALL, D. E., STOREY, J., BAGINSKAITE, J., DE OLIVEIRA, C. L., LABAN, K., CHARBONNEY, E., LYNN, S. A., CASCELLA, M., WHEATER, E., BAKER, D., CHEYNE, R., CHRISTOPHER, E., RONCON, P., DE-SOUZA, E. A., WARDA, M., CORKE, S., AMMAR, Z., O'CONNOR, L., DEVONSHIRE, I. M., MCCANN, S. K., GRAY, L. J., TANRIVER-AYDER, E. & ON BEHALF OF THE, I. C. 2019. A randomised controlled trial of an Intervention to Improve Compliance with the ARRIVE guidelines (IICARus). *Research Integrity and Peer Review,* 4**,** 12.
HALLADAY, C. W., TRIKALINOS, T. A., SCHMID, I. T., SCHMID, C. H. & DAHABREH, I. J. 2015. Using data sources beyond PubMed has a modest impact on the results of systematic reviews of therapeutic interventions. *Journal of Clinical Epidemiology,* 68**,** 1076-1084.
HALLBERG, M. 2015. Neuropeptides: Metabolism to Bioactive Fragments and the Pharmacology of Their Receptors. *Medicinal Research Reviews,* 35**,** 464-519.
HANNEDOUCHE, S., BECK, V., LEIGHTON-DAVIES, J., BEIBEL, M., ROMA, G., OAKELEY, E. J., LANNOY, V., BERNARD, J., HAMON, J., BARBIERI, S., PREUSS, I., LASBENNES, M. C., SAILER, A. W., SUPLY, T., SEUWEN, K., PARKER, C. N. & BASSILANA, F. 2013. Identification of the C3a Receptor (C3AR1) as the Target of the VGF-derived Peptide TLQP-21 in Rodent Cells. *Journal of Biological Chemistry,* 288**,** 27434-27443.
HAROUTOUNIAN, S., ARENDT-NIELSEN, L., BELTON, J., BLYTH, F. M., DEGENHARDT, L., DI FORTI, M., ECCLESTON, C., FINN, D. P., FINNERUP, N. B., FISHER, E., FOGARTY, A. E., GILRON, I., HOHMANN, A. G., KALSO, E., KRANE, E., MOHIUDDIN, M., MOORE, R. A., ROWBOTHAM, M., SOLIMAN, N., WALLACE, M., ZINBOONYAHGOON, N. & RICE, A. S. 2021. IASP Presidential

Task Force on Cannabis and Cannabinoid Analgesia: research agenda on the use of cannabinoids, cannabis, and cannabis-based medicines for pain management. *PAIN,* Articles in Press.

HAROUTOUNIAN, S., GILRON, I., BELTON, J., DEGENHARDT, L., DIFORTI, M., FINN, D. P., FOGARTY, A., KALSO, E., KRANE, E., MOORE, R. A., ROWBOTHAM, M., WALLACE, M. & RICE, A. S. C. 2020. Societal issues and policy implications related to the use of cannabinoids, cannabis, and cannabis-based medicines for pain management. *PAIN,* Articles in Press.

HARTLING, L., FEATHERSTONE, R., NUSPL, M., SHAVE, K., DRYDEN, D. M. & VANDERMEER, B. 2016. The contribution of databases to the results of systematic reviews: a cross-sectional study. *BMC Med Res Methodol,* 16**,** 127.

HASNIE, F. S., BREUER, J., PARKER, S., WALLACE, V., BLACKBEARD, J., LEVER, I., KINCHINGTON, P. R., DICKENSON, A. H., PHEBY, T. & RICE, A. S. C. 2007. Further characterization of a rat model of varicella zoster virus-associated pain: Relationship between mechanical hypersensitivity and anxiety-related behavior, and the influence of analgesic drugs. *Neuroscience,* 144**,** 1495-1508.

HASRIADI, WASANA, P. W. D., VAJRAGUPTA, O., ROJSITTHISAK, P. & TOWIWAT, P. 2021. Automated home-cage for the evaluation of innate non-reflexive pain behaviors in a mouse model of inflammatory pain. *Scientific Reports,* 11**,** 12240.

HIGGINS, J., THOMAS, J., CHANDLER, J., CUMPSTON, M., LI, T., PAGE, M. & WELCH, V. E. Cochrane, 2021. *Cochrane Handbook for Systematic Reviews of Interventions version 6.2 (updated February 2021)* [Online]. Available: www.training.cochrane.org/handbook [Accessed May 2021].

HIGGINS, J. P. T., ALTMAN, D. G., GOTZSCHE, P. C., JUNI, P., MOHER, D., OXMAN, A. D., SAVOVIC, J., SCHULZ, K. F., WEEKS, L., STERNE, J. A. C., COCHRANE BIAS METHODS, G. & COCHRANE STAT METHODS, G. 2011. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *Bmj-British Medical Journal,* 343.

HIGGINS, J. P. T., THOMPSON, S. G., DEEKS, J. J. & ALTMAN, D. G. 2003. Measuring inconsistency in meta-analyses. *BMJ (Clinical research ed.),* 327**,** 557-560.

HIRST, J. A., HOWICK, J., ARONSON, J. K., ROBERTS, N., PERERA, R., KOSHIARIS, C. & HENEGHAN, C. 2014. The Need for Randomization in Animal Trials: An Overview of Systematic Reviews. *Plos One,* 9.

HOOIJMANS, C. R., DE VRIES, R. B. M., RITSKES-HOITINGA, M., ROVERS, M. M., LEEFLANG, M. M., INTHOUT, J., WEVER, K. E., HOOFT, L., DE BEER, H., KUIJPERS, T., MACLEOD, M. R., SENA, E. S., TER RIET, G., MORGAN, R. L., THAYER, K. A., ROONEY, A. A., GUYATT, G. H., SCHÜNEMANN, H. J., LANGENDAM, M. W. & ON BEHALF OF THE, G. W. G. 2018. Facilitating healthcare decisions by assessing the certainty in the evidence from preclinical animal studies. *PLOS ONE,* 13**,** e0187271.

HOOIJMANS, C. R., GEESSINK, F. J., RITSKES-HOITINGA, M. & SCHEFFER, G. J. 2015. A systematic review and meta-analysis of the ability of analgesic drugs to reduce metastasis in experimental cancer models. *Pain,* 156**,** 1835-1844.

HOOIJMANS, C. R., ROVERS, M., DE VRIES, R. B., LEENAARS, M. & RITSKES-HOITINGA, M. 2012. An initiative to facilitate well-informed decision-making in laboratory animal research: report of the First International Symposium on Systematic Reviews in Laboratory Animal Science. *Laboratory Animals,* 46**,** 356-357.

HOOIJMANS, C. R., ROVERS, M. M., DE VRIES, R. B. M., LEENAARS, M., RITSKES-HOITINGA, M. & LANGENDAM, M. W. 2014. SYRCLE's risk of bias tool for animal studies. *BMC Medical Research Methodology,* 14**,** 43.

HOOIJMANS, C. R., TILLEMA, A., LEENAARS, M. & RITSKES-HOITINGA, M. 2010. Enhancing search efficiency by means of a search filter for finding all studies on animal experimentation in PubMed. *Laboratory Animals,* 44**,** 170-175.

HOWARD, B. E., PHILLIPS, J., MILLER, K., TANDON, A., MAV, D., SHAH, M. R., HOLMGREN, S., PELCH, K. E., WALKER, V., ROONEY, A. A., MACLEOD, M., SHAH, R. R. & THAYER, K. 2016a. SWIFT-Review: a text-mining workbench for systematic review. *Syst Rev,* 5**,** 87.

HOWARD, B. E., PHILLIPS, J., MILLER, K., TANDON, A., MAV, D., SHAH, M. R., HOLMGREN, S., PELCH, K. E., WALKER, V., ROONEY, A. A., MACLEOD, M., SHAH, R. R. & THAYER, K. 2016b. SWIFT-Review: a text-mining workbench for systematic review. *Systematic Reviews,* 5**,** 87.

HUANG, W., PERCIE DU SERT, N., VOLLERT, J. & RICE, A. S. C. 2020. General Principles of Preclinical Study Design. *In:* BESPALOV, A., MICHEL, M. C. & STECKLER, T. (eds.) *Good Research Practice in Non-Clinical Pharmacology and Biomedicine.* Cham: Springer International Publishing.

HUGGINS, J. P., SMART, T. S., LANGMAN, S., TAYLOR, L. & YOUNG, T. 2012. An efficient randomised, placebo-controlled clinical trial with the irreversible fatty acid amide hydrolase-1 inhibitor PF-04457845, which modulates endocannabinoids but fails to induce effective analgesia in patients with pain due to osteoarthritis of the knee. *Pain,* 153**,** 1837-1846.

IMPERIAL COLLEGE LONDON. 2020. *Payscale for Student Demonstrators* [Online]. Available: https://www.imperial.ac.uk/media/imperial-college/administration-and-support-services/hr/public/salaries/job-families/Student-Demonstrators-rates-2020-21.pdf [Accessed 20/04/2021].

INTHOUT, J., IOANNIDIS, J. P. A. & BORM, G. F. 2014. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Medical Research Methodology,* 14**,** 25.

IOANNIDIS, J. P. A. 2005. Why most published research findings are false. *Plos Medicine,* 2**,** 696-701.

IOANNIDIS, J. P. A., GREENLAND, S., HLATKY, M. A., KHOURY, M. J., MACLEOD, M. R., MOHER, D., SCHULZ, K. F. & TIBSHIRANI, R. 2014. Increasing value and reducing waste in research design, conduct, and analysis. *Lancet,* 383**,** 166-175.

JACKSON, C. B., ØSTERLUND, C., MUGAR, G., HASSMAN, K. D. & CROWSTON, K. Motivations for Sustained Participation in Crowdsourcing: Case Studies of Citizen Science on the Role of Talk. 2015 48th Hawaii International Conference on System Sciences, 5-8 Jan. 2015 2015. 1624-1634.

JENSEN, T. S., BARON, R., HAANPAA, M., KALSO, E., LOESER, J. D., RICE, A. S. C. & TREEDE, R. D. 2011. A new definition of neuropathic pain. *Pain,* 152**,** 2204-2205.

JÜNI, P., ALTMAN, D. G. & EGGER, M. 2001. Assessing the quality of controlled clinical trials. *BMJ,* 323**,** 42-46.

KAUR, I., BEHL, T., BUNGAU, S., ZENGIN, G., KUMAR, A., EL-ESAWI, M. A., KHULLAR, G., VENKATACHALAM, T. & ARORA, S. 2020. The endocannabinoid signaling pathway as an emerging target in pharmacotherapy, earmarking mitigation of destructive events in rheumatoid arthritis. *Life Sciences,* 257**,** 118109.

KAWRYKOW, A., ROUMANIS, G., KAM, A., KWAK, D., LEUNG, C., WU, C., ZAROUR, E., PHYLO, P., SARMENTA, L., BLANCHETTE, M. & WALDISPÜHL, J. 2012. Phylo: A Citizen Science Approach for Improving Multiple Sequence Alignment. *PLOS ONE,* 7**,** e31362.

KEENAN, C., NOONE, C., MCCONNELL, K. & CHENG, S. H. 2021. A rapid response to the COVID-19 outbreak: the meta-evidence project. *Journal of EAHIL,* 17**,** 16-20.

KILKENNY, C., BROWNE, W., CUTHILL, I. C., EMERSON, M. & ALTMAN, D. G. 2010a. Animal research: Reporting in vivo experiments: The ARRIVE guidelines. *British Journal of Pharmacology,* 160**,** 1577-1579.

KILKENNY, C., BROWNE, W. J., CUTHILL, I. C., EMERSON, M. & ALTMAN, D. G. 2010b. Improving Bioscience Research Reporting: The ARRIVE Guidelines for Reporting Animal Research. *Plos Biology,* 8.

KILKENNY, C., PARSONS, N., KADYSZEWSKI, E., FESTING, M. F. W., CUTHILL, I. C., FRY, D., HUTTON, J. & ALTMAN, D. G. 2009. Survey of the Quality of Experimental Design, Statistical Analysis and Reporting of Research Using Animals. *PLOS ONE,* 4**,** e7824.

KIMMELMAN, J., MOGIL, J. S. & DIRNAGL, U. 2014. Distinguishing between Exploratory and Confirmatory Preclinical Research Will Improve Translation. *Plos Biology,* 12.

KNAPP, G. & HARTUNG, J. 2003. Improved tests for a random effects meta-regression with a single covariate. *Stat Med,* 22**,** 2693-710.

LANDIS, S. C., AMARA, S. G., ASADULLAH, K., AUSTIN, C. P., BLUMENSTEIN, R., BRADLEY, E. W., CRYSTAL, R. G., DARNELL, R. B., FERRANTE, R. J., FILLIT, H., FINKELSTEIN, R., FISHER, M., GENDELMAN, H. E., GOLUB, R. M., GOUDREAU, J. L., GROSS, R. A., GUBITZ, A. K., HESTERLEE, S. E., HOWELLS, D. W., HUGUENARD, J., KELNER, K., KOROSHETZ, W., KRAINC, D., LAZIC, S. E., LEVINE, M. S., MACLEOD, M. R., MCCALL, J. M., MOXLEY III, R. T., NARASIMHAN, K., NOBLE, L. J., PERRIN, S., PORTER, J. D., STEWARD, O., UNGER, E., UTZ, U. & SILBERBERG, S. D. 2012. A call for transparent reporting to optimize the predictive value of preclinical research. *Nature,* 490**,** 187-191.

LANGAN, D., HIGGINS, J. P. T., JACKSON, D., BOWDEN, J., VERONIKI, A. A., KONTOPANTELIS, E., VIECHTBAUER, W. & SIMMONDS, M. 2019. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Res Synth Methods,* 10**,** 83-98.

LEE, H., LAMB, S. E., BAGG, M. K., TOOMEY, E., CASHIN, A. G. & MOSELEY, G. L. 2018. Reproducible and replicable pain research: a critical review. *Pain,* 159**,** 1683-1689.

LEENAARS, M., HOOIJMANS, C. R., VAN VEGGEL, N., TER RIET, G., LEEFLANG, M., HOOFT, L., VAN DER WILT, G. J., TILLEMA, A. & RITSKES-HOITINGA, M. 2012. A step-by-step guide to systematically identify all relevant animal studies. *Laboratory Animals,* 46**,** 24-31.

LEFEBVRE, C., GLANVILLE, J., WIELAND, L. S., COLES, B. & WEIGHTMAN, A. L. 2013. Methodological developments in searching for studies for systematic reviews: past, present and future? *Systematic Reviews,* 2**,** 78.

LEFEBVRE, C., MANHEIMER, E. & GLANVILLE, J. 2008. Searching for Studies. *Cochrane Handbook for Systematic Reviews of Interventions.*

LENG, G., LENG, R. I. & MACLEAN, S. 2019. The vasopressin–memory hypothesis: a citation network analysis of a debate. *Annals of the New York Academy of Sciences,* 1455**,** 126-140.

LEVI, A., ELDRIDGE, J. D. & PATERSON, B. M. 1985. MOLECULAR-CLONING OF A GENE SEQUENCE REGULATED BY NERVE GROWTH-FACTOR. *Science,* 229**,** 393-395.

LEVI, A., FERRI, G. L., WATSON, E., POSSENTI, R. & SALTON, S. R. J. 2004. Processing, distribution, and function of VGF, a neuronal and endocrine peptide precursor. *Cellular and Molecular Neurobiology,* 24**,** 517-533.

LIAO, J., ANANIADOU, S., CURRIE, G. L., HOWARD, B. E., RICE, A., SENA, E. S., THOMAS, J., VARGHESE, A. & MACLEOD, M. R. 2018. Automation of citation screening in pre-clinical systematic reviews. *bioRxiv*.

LICHTMAN, A. H. & CHAPMAN, V. 2011. A FAAH-fetched approach to treat osteoarthritis pain. *PAIN,* 152**,** 959-960.

LIN, W. J., JIANG, C., SADAHIRO, M., BOZDAGI, O., VULCHANOVA, L., ALBERINI, C. M. & SALTON, S. R. 2015. VGF and Its C-Terminal Peptide TLQP-62 Regulate Memory Formation in Hippocampus via a BDNF-TrkB-Dependent Mechanism. *Journal of Neuroscience,* 35**,** 10343-10356.

LIND, A. L., KHOONSARI, P. E., SJODIN, M., KATILA, L., WETTERHALL, M., GORDH, T. & KULTIMA, K. 2016. Spinal Cord Stimulation Alters Protein Levels in the Cerebrospinal Fluid of Neuropathic Pain Patients: A Proteomic Mass Spectrometric Analysis. *Neuromodulation,* 19**,** 549-562.

MACLEOD, M., O'COLLINS, T., HOWELLS DAVID, W. & DONNAN GEOFFREY, A. 2004. Pooling of Animal Experimental Data Reveals Influence of Study Design and Publication Bias. *Stroke,* 35**,** 1203-1208.

MACLEOD, M. R., LAWSON MCLEAN, A., KYRIAKOPOULOU, A., SERGHIOU, S., DE WILDE, A., SHERRATT, N., HIRST, T., HEMBLADE, R., BAHOR, Z., NUNES-FONSECA, C., POTLURU, A., THOMSON, A., BAGINSKITAE, J., EGAN, K., VESTERINEN, H., CURRIE, G. L., CHURILOV, L., HOWELLS, D. W. & SENA, E. S. 2015. Risk of Bias in Reports of In Vivo Research: A Focus for Improvement. *PLOS Biology,* 13**,** e1002273.

MACLEOD, M. R., MICHIE, S., ROBERTS, I., DIRNAGL, U., CHALMERS, I., IOANNIDIS, J. P. A., SALMAN, R. A., CHAN, A. W. & GLASZIOU, P. 2014. Biomedical research: increasing value, reducing waste. *Lancet,* 383**,** 101-104.

MACLEOD, M. R., TANRIVER-AYDER, E., HAIR, K. & SENA, E. 2020. Design of Meta-Analysis Studies. *In:* BESPALOV, A., MICHEL, M. C. & STECKLER, T. (eds.) *Good Research Practice in Non-Clinical Pharmacology and Biomedicine.* Cham: Springer International Publishing.

MAGNUSSON, K. 2020. Interpreting Cohen's d effect size: An interactive visualization (Version 2.4.2).

MAIER, C., BARON, R., TOLLE, T. R., BINDER, A., BIRBAUMER, N., BIRKLEIN, F., GIERTHMUHLEN, J., FLOR, H., GEBER, C., HUGE, V., KRUMOVA, E. K., LANDWEHRMEYER, G. B., MAGERL, W., MAIHOFNER, C., RICHTER, H., ROLKE, R., SCHERENS, A., SCHWARZ, A., SOMMER, C., TRONNIER, V., UCEYLER, N., VALET, M., WASNER, G. & TREEDE, R. D. 2010. Quantitative sensory testing in the German Research Network on Neuropathic Pain (DFNS): Somatosensory abnormalities in 1236 patients with different neuropathic pain syndromes. *Pain,* 150**,** 439-450.

MALCANGIO, M., FERNANDES, K. & TOMLINSON, D. R. 1998. NMDA receptor activation modulates evoked release of substance P from rat spinal cord. *British Journal of Pharmacology,* 125**,** 1625-1626.

MANTYH, P. W., KOLTZENBURG, M., MENDELL, L. M., TIVE, L. & SHELTON, D. L. 2011. Antagonism of Nerve Growth Factor-TrkA Signaling and the Relief of Pain. *Anesthesiology,* 115**,** 189-204.

MAO, J. 2012. Current challenges in translational pain research. *Trends in Pharmacological Sciences,* 33**,** 568-573.

MARATOU, K., WALLACE, V. C. J., HASNIE, F. S., OKUSE, K., HOSSEINI, R., JINA, N., BLACKBEARD, J., PHEBY, T., ORENGO, C., DICKENSON, A. H., MCMAHON, S. B. & RICE, A. S. C. 2009. Comparison of dorsal root ganglion gene expression in rat models of traumatic and HIV-associated neuropathic pain. *European Journal of Pain,* 13**,** 387-398.

MASTERS, K., YOUNG, E. O., COX, J., SIMMONS, B., LINTOTT, C., GRAHAM, G., GREENHILL, A. & HOLMES, K. 2016. Science learning via participation in online citizen science. *Journal of Science Communication,* 15.

MATHES, T., KLASSEN, P. & PIEPER, D. 2017. Frequency of data extraction errors and methods to increase data extraction quality: a methodological review. *Bmc Medical Research Methodology,* 17.

MATSON, D. J., BROOM, D. C., CARSON, S. R., BALDASSARI, J., KEHNE, J. & CORTRIGHT, D. N. 2007. Inflammation-induced reduction of spontaneous activity by adjuvant: A novel model to study the effect of analgesics in rats. *Journal of Pharmacology and Experimental Therapeutics,* 320**,** 194-201.

MCGRATH, J. C., DRUMMOND, G. B., MCLACHLAN, E. M., KILKENNY, C. & WAINWRIGHT, C. L. 2010. Guidelines for reporting experiments involving animals: the ARRIVE guidelines. *British Journal of Pharmacology,* 160**,** 1573-1576.

MCGUINNESS, L., SCHMIDT, L., SONKKILA, T. & JAHN, N. 2021. *medrxivr: Access and Search MedRxiv and BioRxiv Preprint Data* [Online]. Available: https://CRAN.R-project.org/package=medrxivr [Accessed 31/03/21].

MCNUTT, M. K., BRADFORD, M., DRAZEN, J. M., HANSON, B., HOWARD, B., JAMIESON, K. H., KIERMER, V., MARCUS, E., POPE, B. K., SCHEKMAN, R., SWAMINATHAN, S., STANG, P. J. & VERMA, I. M. 2018. Transparency in authors' contributions and responsibilities to promote integrity in scientific publication. *Proceedings of the National Academy of Sciences of the United States of America,* 115**,** 2557-2560.

MENON, J. 2020. Case study for the MKMD funding program: conducting syntheses of evidence – the impact on research and researchers.

MERSKEY, H. 1994. LOGIC, TRUTH AND LANGUAGE IN CONCEPTS OF PAIN. *Quality of Life Research,* 3**,** S69-S76.

MESH. 2019a. *Medical Subject Headings* [Online]. Available: https://www.ncbi.nlm.nih.gov/mesh/2028178 [Accessed 07/05/2021].
MESH. 2019b. *Medical Subject Headings* [Online]. Available: https://www.ncbi.nlm.nih.gov/mesh?cmd=Retrieve&dopt=Full&list_uids=68012196 [Accessed 07/05/2021].
MEYER, M., MATSUOKA, I., WETMORE, C., OLSON, L. & THOENEN, H. 1992. ENHANCED SYNTHESIS OF BRAIN-DERIVED NEUROTROPHIC FACTOR IN THE LESIONED PERIPHERAL-NERVE - DIFFERENT MECHANISMS ARE RESPONSIBLE FOR THE REGULATION OF BDNF AND NGF MESSENGER-RNA. *Journal of Cell Biology,* 119**,** 45-54.
MILLAN, M. J. 1999. The induction of pain: An integrative review. *Progress in Neurobiology,* 57**,** 1-164.
MIROWSKI, P. 2018. The future(s) of open science. *Social Studies of Science,* 48**,** 171-203.
MIWA, M., THOMAS, J., O'MARA-EVES, A. & ANANIADOU, S. 2014. Reducing systematic review workload through certainty-based screening. *Journal of Biomedical Informatics,* 51**,** 242-253.
MOGIL, J. S. 2009. Animal models of pain: progress and challenges. *Nat Rev Neurosci,* 10**,** 283-94.
MOGIL, J. S. 2012a. Pain genetics: past, present and future. *Trends in Genetics,* 28**,** 258-266.
MOGIL, J. S. 2012b. Sex differences in pain and pain inhibition: multiple explanations of a controversial phenomenon. *Nature Reviews Neuroscience,* 13**,** 859-866.
MOGIL, J. S. 2017. Laboratory environmental factors and pain behavior: the relevance of unknown unknowns to reproducibility and translation. *Lab Animal,* 46**,** 136-141.
MOGIL, J. S. & CHANDA, M. L. 2005. The case for the inclusion of female subjects in basic science studies of pain. *Pain,* 117**,** 1-5.
MOGIL, J. S., CHESLER, E. J., WILSON, S. G., JURASKA, J. M. & STERNBERG, W. F. 2000. Sex differences in thermal nociception and morphine antinociception in rodents depend on genotype. *Neuroscience & Biobehavioral Reviews,* 24**,** 375-389.
MOGIL, J. S. & CRAGER, S. E. 2004. What should we be measuring in behavioral studies of chronic pain in animals? *Pain,* 112**,** 12-15.
MOGIL, J. S., DAVIS, K. D. & DERBYSHIRE, S. W. 2010. The necessity of animal models in pain research. *PAIN,* 151**,** 12-17.
MOGIL, J. S., SMITH, S. B., O'REILLY, M. K. & PLOURDE, G. 2005. Influence of nociception and stress-induced antinociception on genetic variation in isoflurane anesthetic potency among mouse strains. *Anesthesiology,* 103**,** 751-758.
MOHAMMAD, F., HO, J., WOO, J. H., LIM, C. L., POON, D. J. J., LAMBA, B. & CLARIDGE-CHANG, A. 2016. Concordance and incongruence in preclinical anxiety models: Systematic review and meta-analyses. *Neuroscience and Biobehavioral Reviews,* 68**,** 504-529.
MOHER, D., GLASZIOU, P., CHALMERS, I., NASSER, M., BOSSUYT, P. M. M., KOREVAAR, D. A., GRAHAM, I. D., RAVAUD, P. & BOUTRON, I. 2016. Increasing value and reducing waste in biomedical research: who's listening? *Lancet,* 387**,** 1573-1586.
MOHIUDDIN, M., BLYTH, F. M., DEGENHARDT, L., DI FORTI, M., ECCLESTON, C., HAROUTOUNIAN, S., MOORE, A., RICE, A. S. C., WALLACE, M., PARK, R. & GILRON, I. 2020. General risks of harm with cannabinoids, cannabis, and cannabis-based medicine possibly relevant to patients receiving these for pain management: an overview of systematic reviews. *PAIN,* Articles in Press.
MOLTENI, L., RIZZI, L., BRESCIANI, E., POSSENTI, R., PASSERI, P. P., GHE, C., MUCCIOLI, G., FEHRENTZ, J. A., VERDIE, P., MARTINEZ, J., OMELJANIUK, R. J., BIAGINI, G., BINDA, A., RIVOLTA, I., LOCATELLI, V. & TORSELLO, A. 2017. Pharmacological and Biochemical Characterization of TLQP-21 Activation of a Binding Site on CHO Cells. *Frontiers in Pharmacology,* 8.
MOORE, R. A., FISHER, E., FINN, D. P., FINNERUP, N. B., GILRON, I., HAROUTOUNIAN, S., KRANE, E., RICE, A. S. C., ROWBOTHAM, M., WALLACE, M. & ECCLESTON, C. 2020. Cannabinoids, cannabis, and cannabis-based medicines for pain management: an overview of systematic reviews. *PAIN,* Articles in Press.

MOSS, A., INGRAM, R., KOCH, S., THEODOROU, A., LOW, L., BACCEI, M., HATHWAY, G. J., COSTIGAN, M., SALTON, S. R. & FITZGERALD, M. 2008. Origins, actions and dynamic expression patterns of the neuropeptide VGF in rat peripheral and central sensory neurones following peripheral nerve injury. *Molecular Pain,* 4.

MOURAUX, A., BANNISTER, K., BECKER, S., FINN, D. P., PICKERING, G., POGATZKI-ZAHN, E. & GRAVEN-NIELSEN, T. 2021. Challenges and opportunities in translational pain research – An opinion paper of the working group on translational pain research of the European pain federation (EFIC). *European Journal of Pain,* 25**,** 731-756.

NEGUS, S. S. 2019. Core Outcome Measures in Preclinical Assessment of Candidate Analgesics. *Pharmacological Reviews,* 71**,** 225.

NESTOR, M. S., FISCHER, D. L., ARNOLD, D., BERMAN, B. & DEL ROSSO, J. Q. 2020. Rethinking the Journal Impact Factor and Publishing in the Digital Age. *The Journal of clinical and aesthetic dermatology,* 13**,** 12-17.

NEUGEBAUER, V., SCHAIBLE, H. G. & SCHMIDT, R. F. 1989. SENSITIZATION OF ARTICULAR AFFERENTS TO MECHANICAL STIMULI BY BRADYKININ. *Pflugers Archiv-European Journal of Physiology,* 415**,** 330-335.

NIESTERS, M., DAHAN, A., KEST, B., ZACNY, J., STIJNEN, T., AARTS, L. & SARTON, E. 2010. Do sex differences exist in opioid analgesia? A systematic review and meta-analysis of human experimental and clinical studies. *PAIN,* 151.

NOEL-STORR, A., DOOLEY, G., AFFENGRUBER, L. & GARTLEHNER, G. 2021a. Citation screening using crowdsourcing and machine learning produced accurate results: Evaluation of Cochrane's modified Screen4Me service. *Journal of Clinical Epidemiology,* 130**,** 23-31.

NOEL-STORR, A., DOOLEY, G., ELLIOTT, J., STEELE, E., SHEMILT, I., MAVERGAMES, C., WISNIEWSKI, S., MCDONALD, S., MURANO, M., GLANVILLE, J., FOXLEE, R., BEECHER, D., WARE, J. & THOMAS, J. 2021b. An evaluation of Cochrane Crowd found that crowdsourcing produced accurate results in identifying randomized trials. *Journal of Clinical Epidemiology,* 133**,** 130-139.

NOEL-STORR, A., DOOLEY, G., FEATHERSTONE, R., WISNIEWSKI, S., SHEMILT, I., THOMAS, J., GARTLEHNER, G., NUßBAUMER-STEIT, B. & MAVERGAMES, C. 2021c. Crowdsourcing and COVID-19: a case study of Cochrane Crowd. *Journal of EAHIL,* 17**,** 27-31.

NOORDZIJ, M., HOOFT, L., DEKKER, F. W., ZOCCALI, C. & JAGER, K. J. 2009. Systematic reviews and meta-analyses: when they are useful and when to be careful. *Kidney Int,* 76**,** 1130-6.

NOSEK, B. A., HARDWICKE, T. E., MOSHONTZ, H., ALLARD, A., COKER, K. S., DERBER, A., FIDLER, F., HILGARD, J., STRUHL, M. K., NUIJTEN, M., ROHRER, J., ROMERO, F., SHEEL, A., SHERER, L., SCHONBRODT, F. & VAZIRE, S. 2021. Replicability, Robustness, and Reproducibility in Psychological Science. *PsyArXiv*.

O'CONNOR, A. M., TSAFNAT, G., THOMAS, J., GLASZIOU, P., GILBERT, S. B. & HUTTON, B. 2019. A question of trust: can we build an evidence base to gain trust in systematic review automation technologies? *Syst Rev,* 8**,** 143.

O'CONNOR, T. M., O'CONNELL, J., O'BRIEN, D. I., GOODE, T., BREDIN, C. P. & SHANAHAN, F. 2004. The role of substance P in inflammatory disease. *Journal of Cellular Physiology,* 201**,** 167-180.

O'MARA-EVES, A., THOMAS, J., MCNAUGHT, J., MIWA, M. & ANANIADOU, S. 2015. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic Reviews,* 4**,** 5.

ONIE, S. 2020. Redesign open science for Asia, Africa and Latin America. *Nature*.

OPEN AND COLLABORATIVE SCIENCE IN DEVELOPMENT NETWORK, O. *Open Science Manifesto* [Online]. Available: https://ocsdnet.org/manifesto/open-science-manifesto/ [Accessed 26/04/2021].

ORTEGA, F., GONZALEZ-BARAHONA, J. M. & ROBLES, G. 2008. On the Inequality of Contributions to Wikipedia. *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)***,** 304-304.

PAWSON, R., GREENHALGH, T., HARVEY, G. & WALSHE, K. 2005. Realist review--a new method of systematic review designed for complex policy interventions. *J Health Serv Res Policy,* 10 Suppl 1**,** 21-34.

PAYAN, D. G. 1989. NEUROPEPTIDES AND INFLAMMATION - THE ROLE OF SUBSTANCE-P. *Annual Review of Medicine,* 40**,** 341-352.

PEARSON, H. 2021. How COVID broke the evidence pipeline. *Nature,* 593**,** 182-185.

PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M. & DUCHESNAY, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research,* 12**,** 2825-2830.

PERCIE DU SERT, N., AHLUWALIA, A., ALAM, S., AVEY, M. T., BAKER, M., BROWNE, W. J., CLARK, A., CUTHILL, I. C., DIRNAGL, U., EMERSON, M., GARNER, P., HOLGATE, S. T., HOWELLS, D. W., HURST, V., KARP, N. A., LAZIC, S. E., LIDSTER, K., MACCALLUM, C. J., MACLEOD, M., PEARL, E. J., PETERSEN, O. H., RAWLE, F., REYNOLDS, P., ROONEY, K., SENA, E. S., SILBERBERG, S. D., STECKLER, T. & WÜRBEL, H. 2020. Reporting animal research: Explanation and elaboration for the ARRIVE guidelines 2.0. *PLOS Biology,* 18**,** e3000411.

PERCIE DU SERT, N., HURST, V., AHLUWALIA, A., ALAM, S., AVEY, M. T., BAKER, M., BROWNE, W. J., CLARK, A., CUTHILL, I. C., DIRNAGL, U., EMERSON, M., GARNER, P., HOLGATE, S. T., HOWELLS, D. W., KARP, N. A., LIDSTER, K., MACCALLUM, C. J., MACLEOD, M., PETERSEN, O., RAWLE, F., REYNOLDS, P., ROONEY, K., SENA, E. S., SILBERBERG, S. D., STECKLER, T. & WÜRBEL, H. 2019. The ARRIVE guidelines 2019: updated guidelines for reporting animal research. *bioRxiv***,** 703181.

PEREL, P., ROBERTS, I., SENA, E., WHEBLE, P., BRISCOE, C., SANDERCOCK, P., MACLEOD, M., MIGNINI, L. E., JAYARAM, P. & KHAN, K. S. 2007. Comparison of treatment effects between animal experiments and clinical trials: systematic review. *British Medical Journal,* 334**,** 197-200.

PETERS, J. L., SUTTON, A. J., JONES, D. R., ABRAMS, K. R. & RUSHTON, L. 2007. Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. *Statistics in Medicine,* 26**,** 4544-4562.

PEZET, S. & MCMAHON, S. B. 2006. Neurotrophins: Mediators and modulators of pain. *Annual Review of Neuroscience,* 29**,** 507-538.

PHILLIPS, T. J. C., BROWN, M., RAMIREZ, J. D., PERKINS, J., WOLDEAMANUEL, Y. W., WILLIAMS, A. C. D., ORENGO, C., BENNETT, D. L. H., BODI, I., COX, S., MAIER, C., KRUMOVA, E. K. & RICE, A. S. C. 2014. Sensory, psychological, and metabolic dysfunction in HIV-associated peripheral neuropathy: A cross-sectional deep profiling study. *Pain,* 155**,** 1846-1860.

PICKERING, G., JOURDAN, D., MILLECAMPS, M., CHAPUY, E., ALLIOT, J. & ESCHALIER, A. 2006. Age-related impact of neuropathic pain on animal behaviour. *European Journal of Pain,* 10**,** 749-755.

PITZER, C., KUNER, R. & TAPPE-THEODOR, A. 2016. Voluntary and evoked behavioral correlates in inflammatory pain conditions under different social housing conditions. *Pain reports* [Online], 1.

POGATZKI-ZAHN, E., SEGELCKE, D. & ZAHN, P. 2018. Mechanisms of acute and chronic pain after surgery: update from findings in experimental animal models. *Current Opinion in Anesthesiology,* 31**,** 575-585.

POGATZKI-ZAHN, E. M., SEGELCKE, D. & SCHUG, S. A. 2017. Postoperative pain-from mechanisms to treatment. *Pain Rep,* 2**,** e588.

POL, A. N. V. D., BINA, K., DECAVEL, C. & GHOSH, P. 1994. VGF expression in the brain. *Journal of Comparative Neurology,* 347**,** 455-469.

PORRO, C. A. & CAVAZZUTI, M. 1993. SPATIAL AND TEMPORAL ASPECTS OF SPINAL-CORD AND BRAIN-STEM ACTIVATION IN THE FORMALIN PAIN MODEL. *Progress in Neurobiology,* 41**,** 565-+.

POUND, P. & RITSKES-HOITINGA, M. 2020. Can prospective systematic reviews of animal studies improve clinical translation? *Journal of translational medicine,* 18**,** 15-15.

RAJA, S. N., CARR, D. B., COHEN, M., FINNERUP, N. B., FLOR, H., GIBSON, S., KEEFE, F. J., MOGIL, J. S., RINGKAMP, M., SLUKA, K. A., SONG, X.-J., STEVENS, B., SULLIVAN, M. D., TUTELMAN, P. R., USHIDA, T. & VADER, K. 2020. The revised International Association for the Study of Pain definition of pain: concepts, challenges, and compromises. *PAIN,* 161**,** 1976-1982.

RAY, P., TORCK, A., QUIGLEY, L., WANGZHOU, A., NEIMAN, M., RAO, C., LAM, T., KIM, J.-Y., KIM, T. H., ZHANG, M. Q., DUSSOR, G. & PRICE, T. J. 2018. Comparative transcriptome profiling of the human and mouse dorsal root ganglia: an RNA-seq-based resource for pain and sensory neuroscience research. *Pain,* 159**,** 1325-1345.

REARDON, S. 2016. A mouse's house may ruin experiments. *Nature,* 530**,** 264-264.

REICHARDT, L. F. 2006. Neuratrophin-regulated signalling pathways. *Philosophical Transactions of the Royal Society B-Biological Sciences,* 361**,** 1545-1564.

RICE, A. S. C., CIMINO-BROWN, D., EISENACH, J. C., KONTINEN, V. K., LACROIX-FRALISH, M. L., MACHIN, I., MOGIL, J. S. & STÖHR, T. 2008. Animal models and the prediction of efficacy in clinical trials of analgesic drugs: A critical appraisal and call for uniform reporting standards. *PAIN,* 139**,** 243-247.

RICE, A. S. C., FINNERUP, N. B., KEMP, H. I., CURRIE, G. L. & BARON, R. 2018. Sensory profiling in animal models of neuropathic pain: a call for back-translation. *PAIN,* 159**,** 819-824.

RICE, A. S. C., SMITH, B. H. & BLYTH, F. M. 2016. Pain and the global burden of disease. *PAIN,* 157**,** 791-796.

RIEDL, M. S., BRAUN, P. D., KITTO, K. F., ROIKO, S. A., ANDERSON, L. B., HONDA, C. N., FAIRBANKS, C. A. & VULCHANOVA, L. 2009. Proteomic Analysis Uncovers Novel Actions of the Neurosecretory Protein VGF in Nociceptive Processing. *Journal of Neuroscience,* 29**,** 13377-13388.

RIZZI, R., BARTOLOMUCCI, A., MOLES, A., D'AMATO, F., SACERDOTE, P., LEVI, A., LA CORTE, G., CIOTTI, M. T., POSSENTI, R. & PAVONE, F. 2008. The VGF-derived peptide TLQP-21: A new modulatory peptide for inflammatory pain. *Neuroscience Letters,* 441**,** 129-133.

ROBINSON, K. A. & GOODMAN, S. N. 2011. A Systematic Examination of the Citation of Prior Research in Reports of Randomized, Controlled Trials. *Annals of Internal Medicine,* 154**,** 50-U187.

ROBINSON L.D., CAWTHRAY, J. L., WEST, S. E., BONN, A. & ANSINE, J. 2018. Ten principles of citizen science. *Citizen Science: Innovation in Open Science, Society and Policy.* London: UCL Press.

ROOKE, E. D. M., VESTERINEN, H. M., SENA, E. S., EGAN, K. J. & MACLEOD, M. R. 2011. Dopamine agonists in animal models of Parkinson's disease: A systematic review and meta-analysis. *Parkinsonism & Related Disorders,* 17**,** 313-320.

SAARIO, S. M. & LAITINEN, J. T. 2007. Therapeutic potential of endocannabinoid-hydrolysing enzyme inhibitors. *Basic Clin Pharmacol Toxicol,* 101**,** 287-93.

SALTON, S. R. J. 1991. NUCLEOTIDE-SEQUENCE AND REGULATORY STUDIES OF VGF, A NERVOUS SYSTEM-SPECIFIC MESSENGER-RNA THAT IS RAPIDLY AND RELATIVELY SELECTIVELY INDUCED BY NERVE GROWTH-FACTOR. *Journal of Neurochemistry,* 57**,** 991-996.

SALTON, S. R. J., FERRI, G. L., HAHM, S., SNYDER, S. E., WILSON, A. J., POSSENTI, R. & LEVI, A. 2000. VGF: A novel role for this neuronal and neuroendocrine polypeptide in the regulation of energy balance. *Frontiers in Neuroendocrinology,* 21**,** 199-219.

SAMUEL, A. L. 1959. Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development,* 3**,** 210-229.

SAUERMANN, H. & FRANZONI, C. 2013. Participation Dynamics in Crowd-Based Knowledge Production: The Scope and Sustainability of Interest-Based Motivation. *SSRN*.

SAUERMANN, H. & FRANZONI, C. 2015. Crowd science user contribution patterns and their implications. *Proceedings of the National Academy of Sciences,* 112**,** 679-684.

SCHLESSELMAN, J. J. & COLLINS, J. A. 2003. Evaluating systematic reviews and meta-analyses. *Semin Reprod Med,* 21**,** 95-105.
SCHLOSBURG, J. E., KINSEY, S. G. & LICHTMAN, A. H. 2009. Targeting fatty acid amide hydrolase (FAAH) to treat pain and inflammation. *The AAPS journal,* 11**,** 39-44.
SEIDLER, A. L., HUNTER, K. E., CHEYNE, S., GHERSI, D., BERLIN, J. A. & ASKIE, L. 2019. A guide to prospective meta-analysis. *BMJ,* 367**,** l5342.
SENA, E., VAN DER WORP, H. B., HOWELLS, D. & MACLEOD, M. 2007. How can we improve the pre-clinical development of drugs for stroke? *Trends in Neurosciences,* 30**,** 433-439.
SENA, E. S., CURRIE, G. L., MCCANN, S. K., MACLEOD, M. R. & HOWELLS, D. W. 2014. Systematic reviews and meta-analysis of preclinical studies: why perform them and how to appraise them critically. *Journal of Cerebral Blood Flow and Metabolism,* 34**,** 737-742.
SENA, E. S. & MACLEOD, M. R. 2007. Concordance between laboratory and clinical drug efficacy: Lessons from systematic review and meta-analysis. *Stroke,* 38**,** 502-502.
SENA, E. S., VAN DER WORP, H. B., BATH, P. M. W., HOWELLS, D. W. & MACLEOD, M. R. 2010. Publication Bias in Reports of Animal Stroke Studies Leads to Major Overstatement of Efficacy. *Plos Biology,* 8.
SHEA, B. J., REEVES, B. C., WELLS, G., THUKU, M., HAMEL, C., MORAN, J., MOHER, D., TUGWELL, P., WELCH, V., KRISTJANSSON, E. & HENRY, D. A. 2017. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ,* 358**,** j4008.
SHEMILT, I., ARNO, A., THOMAS, J., LORENC, T., KHOUJA, C., RAINER, G., SUTCLIFFE, K., P., D. S., WRIGHT, K. & SOWDEN, A. 2021. Using automation to produce a 'living map' of the COVID-19 research literature. *Journal of EAHIL,* 17**,** 11-15.
SHEMILT, I., SIMON, A., HOLLANDS, G. J., MARTEAU, T. M., OGILVIE, D., O'MARA-EVES, A., KELLY, M. P. & THOMAS, J. 2014. Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. *Res Synth Methods,* 5**,** 31-49.
SHOJANIA, K. G., SAMPSON, M., ANSARI, M. T., JI, J., DOUCETTE, S. & MOHER, D. 2007. How quickly do systematic reviews go out of date? A survival analysis. *Annals of Internal Medicine,* 147**,** 224-233.
SHORTLAND, P. J., BAYTUG, B., KRZYZANOWSKA, A., MCMAHON, S. B., PRIESTLEY, J. V. & AVERILL, S. 2006. ATF3 expression in L4 dorsal root ganglion neurons after L5 spinal nerve transection. *European Journal of Neuroscience,* 23**,** 365-373.
SIDELI, L., TROTTA, G., SPINAZZOLA, E., LA CASCIA, C. & DI FORTI, M. 2020. Adverse effects of heavy cannabis use: even plants can harm the brain. *PAIN,* Articles in Press.
SIQUEIRA-LIMA, P. S., SILVA, J. C., QUINTANS, J. S. S., ANTONIOLLI, A. R., SHANMUGAM, S., BARRETO, R. S. S., SANTOS, M. R. V., ALMEIDA, J., BONJARDIM, L. R., MENEZES, I. R. A. & QUINTANS, L. J. 2017. Natural products assessed in animal models for orofacial pain - a systematic review. *Revista Brasileira De Farmacognosia-Brazilian Journal of Pharmacognosy,* 27**,** 124-134.
SKORPUT, A. G. J., ZHANG, X., WAATAJA, J. J., PETERSON, C. D., RIEDL, M. S., KITTO, K. F., TRUONG, H., HUFFMAN, C., SALTON, S. R., FAIRBANKS, C. A., HONDA, C. N. & VULCHANOVA, L. 2018. Involvement of the VGF-derived peptide TLQP-62 in nerve injury–induced hypersensitivity and spinal neuroplasticity. *PAIN,* 159**,** 1802-1813.
SMITH, B. H., HÉBERT, H. L. & VELUCHAMY, A. 2020. Neuropathic pain in the community: prevalence, impact, and risk factors. *PAIN,* 161**,** S127-S137.
SNYDER, S. E., PINTAR, J. E. & SALTON, S. R. J. 1998. Developmental expression of VGF mRNA in the prenatal and postnatal rat. *Journal of Comparative Neurology,* 394**,** 64-90.
SNYDER, S. E. & SALTON, S. R. J. 1998. Expression of VGF mRNA in the adult rat central nervous system. *Journal of Comparative Neurology,* 394**,** 91-105.

SOBOCZENSKI, F., TRIKALINOS, T. A., KUIPER, J., BIAS, R. G., WALLACE, B. C. & MARSHALL, I. J. 2019. Machine learning to help researchers evaluate biases in clinical trials: a prospective, randomized user study. *BMC Medical Informatics and Decision Making,* 19**,** 96.

SOLIMAN, N., HAROUTOUNIAN, S., HOHMANN, A. G., KRANE, E., LIAO, J., MACLEOD, M., SEGELCKE, D., SENA, C., THOMAS, J., VOLLERT, J., WEVER, K., ALAVERDYAN, H., BARAKAT, A., BARTHLOW, T., HARRIS BOZER, A. L., DAVIDSON, A., DIAZ-DELCASTILLO, M., DOLGORUKOVA, A., FERDOUSI, M. I., HEALY, C., HONG, S., HOPKINS, M., JAMES, A., LEAKE, H. B., MALEWICZ, N. M., MANSFIELD, M., MARDON, A. K., MATTIMOE, D., MCLOONE, D. P., NOES-HOLT, G., POGATZKI-ZAHN, E. M., POWER, E., PRADIER, B., ROMANOS-SIRAKIS, E., SEGELCKE, A., VINAGRE, R., YANES, J. A., ZHANG, J., ZHANG, X. Y., FINN, D. P. & RICE, A. S. C. 2021. A systematic review and meta-analysis of cannabis-based medicines, cannabinoids and endocannabinoid system modulators tested for antinociceptive effects in animal models of injury-related or pathological persistent pain. *PAIN,* Articles in Press.

SOLIMAN, N., HOHMANN, A. G., HAROUTOUNIAN, S., WEVER, K., RICE, A. S. C. & FINN, D. P. 2019a. A protocol for the systematic review and meta-analysis of studies in which cannabinoids were tested for antinociceptive effects in animal models of pathological or injury-related persistent pain. *Pain reports,* 4**,** e766-e766.

SOLIMAN, N., OKUSE, K. & RICE, A. S. C. 2019b. VGF: a biomarker and potential target for the treatment of neuropathic pain? *PAIN Reports,* 4**,** e786.

SOLIMAN, N., RICE, A. S. C. & VOLLERT, J. 2020. A practical guide to preclinical systematic review and meta-analysis. *PAIN,* 161**,** 1949-1954.

SONG, F., PAREKH, S., HOOPER, L., LOKE, Y. K., RYDER, J., SUTTON, A. J., HING, C., KWOK, C. S., PANG, C. & HARVEY, I. 2010. Dissemination and publication of research findings: an updated review of related biases. *Health Technology Assessment,* 14**,** 1-+.

SPIERS, H., SWANSON, A., FORTSON, L., SIMMONS, B. D., TROUILLE, L., BLICKHAN, S. & LINTOTT, C. 2019. Everyone counts? Design considerations in online citizen science. *Journal of Science Communication,* 18**,** A04.

STEVENSON, G. W., BILSKY, E. J. & NEGUS, S. S. 2006. Targeting pain-suppressed behaviors in preclinical assays of pain and analgesia: Effects of morphine on acetic acid-suppressed feeding in C57BL/6J mice. *Journal of Pain,* 7**,** 408-416.

STRAUB, M. C. P. 2016. Giving Citizen Scientists a Chance: A Study of Volunteer-led Scientific Discovery. *Citizen Science: Theory and Practice,* 1**,** 5.

SUN, W., KOU, D., YU, Z., YANG, S., JIANG, C., XIONG, D., XIAO, L., DENG, Q., XIE, H. & HAO, Y. 2020. A Transcriptomic Analysis of Neuropathic Pain in Rat Dorsal Root Ganglia Following Peripheral Nerve Injury. *Neuromolecular Med,* 22**,** 250-263.

SUOKAS, A. K., SAGAR, D. R., MAPP, P. I., CHAPMAN, V. & WALSH, D. A. 2014. Design, study quality and evidence of analgesic efficacy in studies of drugs in models of OA pain: a systematic review and a meta-analysis. *Osteoarthritis and Cartilage,* 22**,** 1207-1223.

TANNENBAUM, C., GREAVES, L. & GRAHAM, I. D. 2016. Why sex and gender matter in implementation research. *BMC Medical Research Methodology,* 16**,** 145.

TANRIVER-AYDER, E., FAES, C., VAN DE CASTEELE, T., MCCANN, S. K. & MACLEOD, M. R. 2021. Comparison of commonly used methods in random effects meta-analysis: application to preclinical data in drug discovery research. *BMJ Open Science,* 5**,** e100074.

TER RIET, G., KOREVAAR, D. A., LEENAARS, M., STERK, P. J., VAN NOORDEN, C. J. F., BOUTER, L. M., LUTTER, R., ELFERINK, R. P. O. & HOOFT, L. 2012. Publication Bias in Laboratory Animal Research: A Survey on Magnitude, Drivers, Consequences and Potential Solutions. *Plos One,* 7.

TERRIN, N., SCHMID, C. H., LAU, J. & OLKIN, I. 2003. Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine,* 22**,** 2113-2126.

TETREAULT, L. A., ZHU, M. P., WILSON, J. R., KARADIMAS, S. K. & FEHLINGS, M. G. 2020. The Impact of Riluzole on Neurobehavioral Outcomes in Preclinical Models of Traumatic and

Nontraumatic Spinal Cord Injury: Results From a Systematic Review of the Literature. *Global Spine Journal,* 10**,** 216-229.
THAKKER-VARIA, S., BEHNKE, J., DOOBIN, D., DALAL, V., THAKKAR, K., KHADIM, F., WILSON, E., PALMIERI, A., ANTILA, H., RANTAMAKI, T. & ALDER, J. 2014. VGF (TLQP-62)-induced neurogenesis targets early phase neural progenitor cells in the adult hippocampus and requires glutamate and BDNF signaling. *Stem Cell Research,* 12**,** 762-777.
THOMAS, D., CHANCELLOR, D., MICKLUS, A., LAFEVER, S., HAY, M., CHAUDHURI, S., BOWDEN, R. & LO, A. W. 2021a. Clinical Development Success Rates and Contributing Factors 2011–2020.
THOMAS, J., BRUNTON, J. & GRAZIOSI, S. 2010. *EPPI-Reviewer 4.0: software for research synthesis* [Online]. EPPI-Centre Software: London: Social Science Research Unit, Institute of Education. Available: https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=2914 [Accessed].
THOMAS, J., MCDONALD, S., NOEL-STORR, A., SHEMILT, I., ELLIOTT, J., MAVERGAMES, C. & MARSHALL, I. J. 2021b. Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for Cochrane Reviews. *J Clin Epidemiol,* 133**,** 140-151.
THOMAS, J., NOEL-STORR, A., MARSHALL, F., WALLACE, B., MCDONALD, S., MAVERGAMES, C., GLASZIOU, P., SHEMILT, I., SYNNOT, A., TURNER, T., ELLIOTT, J. & LIVING SYSTEMATIC REVIEW, N. 2017. Living systematic reviews: 2. Combining human and machine effort. *Journal of Clinical Epidemiology,* 91**,** 31-37.
TOSHINAI, K. & NAKAZATO, M. 2009. Neuroendocrine regulatory peptide-1 and-2: Novel bioactive peptides processed from VGF. *Cellular and Molecular Life Sciences,* 66**,** 1939-1945.
TREEDE, R.-D., RIEF, W., BARKE, A., AZIZ, Q., BENNETT, M. I., BENOLIEL, R., COHEN, M., EVERS, S., FINNERUP, N. B., FIRST, M. B., GIAMBERARDINO, M. A., KAASA, S., KOSEK, E., LAVAND'HOMME, P., NICHOLAS, M., PERROT, S., SCHOLZ, J., SCHUG, S., SMITH, B. H., SVENSSON, P., VLAEYEN, J. W. S. & WANG, S.-J. 2015. A classification of chronic pain for ICD-11. *Pain,* 156**,** 1003-1007.
TREEDE, R. D., JENSEN, T. S., CAMPBELL, J. N., CRUCCU, G., DOSTROVSKY, J. O., GRIFFIN, J. W., HANSSON, P., HUGHES, R., NURMIKKO, T. & SERRA, J. 2008. Neuropathic pain - Redefinition and a grading system for clinical and research purposes. *Neurology,* 70**,** 1630-1635.
TRICCO, A. C., BREHAUT, J., CHEN, M. H. & MOHER, D. 2008. Following 411 Cochrane Protocols to Completion: A Retrospective Cohort Study. *Plos One,* 3.
TSAFNAT, G., GLASZIOU, P., CHOONG, M. K., DUNN, A., GALGANI, F. & COIERA, E. 2014. Systematic review automation technologies. *Systematic Reviews,* 3**,** 74.
TSOU, A. Y., TREADWELL, J. R., ERINOFF, E. & SCHOELLES, K. 2020. Machine learning for screening prioritization in systematic reviews: comparative performance of Abstrackr and EPPI-Reviewer. *Systematic Reviews,* 9**,** 73.
VALDER, C. R., LIU, J. J., SONG, Y. H. & LUO, Z. D. 2003. Coupling gene chip analyses and rat genetic variances in identifying potential target genes that may contribute to neuropathic allodynia development. *Journal of Neurochemistry,* 87**,** 560-573.
VAN DER GRAAF, A. M., WIEGMAN, M. J., PLOSCH, T., ZEEMAN, G. G., VAN BUITEN, A., HENNING, R. H., BUIKEMA, H. & FAAS, M. M. 2013. Endothelium-Dependent Relaxation and Angiotensin II Sensitivity in Experimental Preeclampsia. *Plos One,* 8.
VAN DER WORP, H. B., HOWELLS, D. W., SENA, E. S., PORRITT, M. J., REWELL, S., O'COLLINS, V. & MACLEOD, M. R. 2010. Can Animal Models of Disease Reliably Inform Human Studies? *Plos Medicine,* 7.
VERONIKI, A. A., JACKSON, D., VIECHTBAUER, W., BENDER, R., BOWDEN, J., KNAPP, G., KUSS, O., HIGGINS, J. P., LANGAN, D. & SALANTI, G. 2016. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Res Synth Methods,* 7**,** 55-79.
VESTERINEN, H. M., SENA, E. S., EGAN, K. J., HIRST, T. C., CHUROLOV, L., CURRIE, G. L., ANTONIC, A., HOWELLS, D. W. & MACLEOD, M. R. 2014. Meta-analysis of data from animal studies: A practical guide. *Journal of Neuroscience Methods,* 221**,** 92-102.

VIECHTBAUER, W. 2005. Bias and Efficiency of Meta-Analytic Variance Estimators in the Random-Effects Model. *Journal of Educational and Behavioral Statistics,* 30**,** 261-293.

VIERCK, C. J., HANSSON, P. T. & YEZIERSKI, R. P. 2008. Clinical and pre-clinical pain assessment: Are we measuring the same thing? *Pain,* 135**,** 7-10.

VOELKL, B., ALTMAN, N. S., FORSMAN, A., FORSTMEIER, W., GUREVITCH, J., JARIC, I., KARP, N. A., KAS, M. J., SCHIELZETH, H., VAN DE CASTEELE, T. & WÜRBEL, H. 2020. Reproducibility of animal research in light of biological variation. *Nature Reviews Neuroscience,* 21**,** 384-393.

VOELKL, B., VOGT, L., SENA, E. S. & WÜRBEL, H. 2018. Reproducibility of preclinical animal research improves with heterogeneity of study samples. *PLOS Biology,* 16**,** e2003693.

VOLLERT, J., MAIER, C., ATTAL, N., BENNETT, D. L. H., BOUHASSIRA, D., ENAX-KRUMOVA, E. K., FINNERUP, N. B., FREYNHAGEN, R., GIERTHMÜHLEN, J., HAANPÄÄ, M., HANSSON, P., HÜLLEMANN, P., JENSEN, T. S., MAGERL, W., RAMIREZ, J. D., RICE, A. S. C., SCHUH-HOFER, S., SEGERDAHL, M., SERRA, J., SHILLO, P. R., SINDRUP, S., TESFAYE, S., THEMISTOCLEOUS, A. C., TÖLLE, T. R., TREEDE, R.-D. & BARON, R. 2017. Stratifying patients with peripheral neuropathic pain based on sensory profiles: algorithm and sample size recommendations. *Pain,* 158**,** 1446-1455.

VOLLERT, J., SCHENKER, E., MACLEOD, M., BESPALOV, A., WUERBEL, H., MICHEL, M., DIRNAGL, U., POTSCHKA, H., WALDRON, A.-M., WEVER, K., STECKLER, T., VAN DE CASTEELE, T., ALTEVOGT, B., SIL, A. & RICE, A. S. C. 2020. Systematic review of guidelines for internal validity in the design, conduct and analysis of preclinical biomedical experiments involving laboratory animals. *BMJ Open Science,* 4**,** e100046.

VOORHEES, E., ALAM, T., BEDRICK, S., DEMNER-FUSHMAN, D., HERSH, W. R., LO, K., ROBERTS, K., SOBOROFF, I. & WANG, L. L. 2021. TREC-COVID: constructing a pandemic information retrieval test collection. *SIGIR Forum,* 54**,** Article 1.

WALLACE, B. C., NOEL-STORR, A., MARSHALL, I. J., COHEN, A. M., SMALHEISER, N. R. & THOMAS, J. 2017. Identifying reports of randomized controlled trials (RCTs) via a hybrid machine learning and crowdsourcing approach. *Journal of the American Medical Informatics Association,* 24**,** 1165-1168.

WALLACE, B. C., TRIKALINOS, T. A., LAU, J., BRODLEY, C. & SCHMID, C. H. 2010. Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioinformatics,* 11**,** 55.

WALLACE, V. C. J., BLACKBEARD, J., PHEBY, T., SEGERDAHL, A. R., DAVIES, M., HASNIE, F., HALL, S., MCMAHON, S. B. & RICE, A. S. C. 2007a. Pharmacological, behavioural and mechanistic analysis of HIV-1 gp120 induced painful neuropathy. *Pain,* 133**,** 47-63.

WALLACE, V. C. J., BLACKBEARD, J., SEGERDAHL, A. R., HASNIE, F., PHEBY, T., MCMAHON, S. B. & RICE, A. S. C. 2007b. Characterization of rodent models of HIV-gp120 and anti-retroviral-associated neuropathic pain. *Brain,* 130**,** 2688-2702.

WANG, H., SUN, H., DELLA PENNA, K., BENZ, R. J., XU, J., GERHOLD, D. L., HOLDER, D. J. & KOBLAN, K. S. 2002. Chronic neuropathic pain is accompanied by global changes in gene expression and shares pathobiology with neurodegenerative diseases. *Neuroscience,* 114**,** 529-546.

WANG, Q., LIAO, J., HAIR, K., BANNACH-BROWN, A., BAHOR, Z., CURRIE, G. L., MCCANN, S. K., HOWELLS, D. W., SENA, E. S. & MACLEOD, M. R. 2018. Estimating the statistical performance of different approaches to meta-analysis of data from animal studies in identifying the impact of aspects of study design. *bioRxiv***,** 256776.

WARDEN, M. K. & YOUNG, W. S. 1988. DISTRIBUTION OF CELLS CONTAINING MESSENGER-RNAS ENCODING SUBSTANCE-P AND NEUROKININ-B IN THE RAT CENTRAL NERVOUS-SYSTEM. *Journal of Comparative Neurology,* 272**,** 90-113.

WEVER, K. E., MENTING, T. P., ROVERS, M., VAN DER VLIET, J. A., RONGEN, G. A., MASEREEUW, R., RITSKES-HOITINGA, M., HOOIJMANS, C. R. & WARLE, M. 2012. Ischemic Preconditioning in the Animal Kidney, a Systematic Review and Meta-Analysis. *Plos One,* 7.

WHITE, F. A., JUNG, H. & MILLER, R. J. 2007. Chemokines and the pathophysiology of neuropathic pain. *Proceedings of the National Academy of Sciences of the United States of America,* 104**,** 20151-20158.

WHITING, P., SAVOVIĆ, J., HIGGINS, J. P., CALDWELL, D. M., REEVES, B. C., SHEA, B., DAVIES, P., KLEIJNEN, J. & CHURCHILL, R. 2016. ROBIS: A new tool to assess risk of bias in systematic reviews was developed. *J Clin Epidemiol,* 69**,** 225-34.

WILKINSON, M. D., DUMONTIER, M., AALBERSBERG, I. J., APPLETON, G., AXTON, M., BAAK, A., BLOMBERG, N., BOITEN, J.-W., DA SILVA SANTOS, L. B., BOURNE, P. E., BOUWMAN, J., BROOKES, A. J., CLARK, T., CROSAS, M., DILLO, I., DUMON, O., EDMUNDS, S., EVELO, C. T., FINKERS, R., GONZALEZ-BELTRAN, A., GRAY, A. J. G., GROTH, P., GOBLE, C., GRETHE, J. S., HERINGA, J., 'T HOEN, P. A. C., HOOFT, R., KUHN, T., KOK, R., KOK, J., LUSHER, S. J., MARTONE, M. E., MONS, A., PACKER, A. L., PERSSON, B., ROCCA-SERRA, P., ROOS, M., VAN SCHAIK, R., SANSONE, S.-A., SCHULTES, E., SENGSTAG, T., SLATER, T., STRAWN, G., SWERTZ, M. A., THOMPSON, M., VAN DER LEI, J., VAN MULLIGEN, E., VELTEROP, J., WAAGMEESTER, A., WITTENBURG, P., WOLSTENCROFT, K., ZHAO, J. & MONS, B. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data,* 3**,** 160018.

WODARSKI, R., DELANEY, A., ULTENIUS, C., MORLAND, R., ANDREWS, N., BAASTRUP, C., BRYDEN, L. A., CASPANI, O., CHRISTOPH, T., GARDINER, N. J., HUANG, W., KENNEDY, J. D., KOYAMA, S., LI, D., LIGOCKI, M., LINDSTEN, A., MACHIN, I., PEKCEC, A., ROBENS, A., ROTARIU, S. M., VOß, S., SEGERDAHL, M., STENFORS, C., SVENSSON, C. I., TREEDE, R.-D., UTO, K., YAMAMOTO, K., RUTTEN, K. & RICE, A. S. C. 2016. Cross-centre replication of suppressed burrowing behaviour as an ethologically relevant pain outcome measure in the rat: a prospective multicentre study. *PAIN,* 157**,** 2350-2365.

WOODHAMS, S. G., CHAPMAN, V., FINN, D. P., HOHMANN, A. G. & NEUGEBAUER, V. 2017. The cannabinoid system and pain. *Neuropharmacology,* 124**,** 105-120.

WOOLF, C. J. & COSTIGAN, M. 1999. Transcriptional and posttranslational plasticity and the generation of inflammatory pain. *Proceedings of the National Academy of Sciences of the United States of America,* 96**,** 7723-7730.

YEZIERSKI, R. P. & HANSSON, P. 2018. Inflammatory and Neuropathic Pain From Bench to Bedside: What Went Wrong? *J Pain,* 19**,** 571-588.

ZHANG, Q., SHI, T. J., JI, R. R., ZHANG, Y. Z., SUNDLER, F., HANNIBAL, J., FAHRENKRUG, J. & HOKFELT, T. 1997. Expression of pituitary adenylate cyclase-activating polypeptide in dorsal root ganglia following axotomy: Time course and coexistence (vol 705, pg 149, 1995). *Brain Research,* 745**,** 357-357.

ZWETSLOOT, P.-P., VAN DER NAALD, M., SENA, E. S., HOWELLS, D. W., INTHOUT, J., DE GROOT, J. A., CHAMULEAU, S. A., MACLEOD, M. R. & WEVER, K. E. 2017. Standardized mean differences cause funnel plot distortion in publication bias assessments. *eLife,* 6**,** e24260.