Edinburgh Research Explorer

Animal models of chemotherapy-induced peripheral neuropathy: A machine-assisted systematic review and meta-analysis
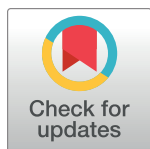
OPEN ACCESS

Animal models of chemotherapy-induced peripheral neuropathy: A machine-assisted systematic review and meta-analysis

Gillian L. Currie[1], Helena N. Angel-Scott[2], Lesley Colvin[3,4], Fala Cramond[2], Kaitlyn Hair[1], Laila Khandoker[2], Jing Liao[1], Malcolm Macleod[1], Sarah K. McCann[1], Rosie Morland[2], Nicki Sherratt[1], Robert Stewart[1], Ezgi Tanriver-Ayder[1], James Thomas[5], Qianying Wang[1], Rachel Wodarski[2], Ran Xiong[2], Andrew S. C. Rice[2☯], Emily S. Sena[1☯]*

1 Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, United Kingdom, 2 Pain Research, Department of Surgery and Cancer, Imperial College London, London, United Kingdom, 3 Department of Anaesthesia, Critical Care & Pain, University of Edinburgh, Edinburgh, United Kingdom, 4 Division of Population Health and Genomics, University of Dundee, Dundee, United Kingdom, 5 EPPI-Centre, University College London, London, United Kingdom

☯ These authors contributed equally to this work.
* emily.sena@ed.ac.uk

OPEN ACCESS

Abstract

We report a systematic review and meta-analysis of research using animal models of chemotherapy-induced peripheral neuropathy (CIPN). We systematically searched 5 online databases in September 2012 and updated the search in November 2015 using machine learning and text mining to reduce the screening for inclusion workload and improve accuracy. For each comparison, we calculated a standardised mean difference (SMD) effect size, and then combined effects in a random-effects meta-analysis. We assessed the impact of study design factors and reporting of measures to reduce risks of bias. We present power analyses for the most frequently reported behavioural tests; 337 publications were included. Most studies (84%) used male animals only. The most frequently reported outcome measure was evoked limb withdrawal in response to mechanical monofilaments. There was modest reporting of measures to reduce risks of bias. The number of animals required to obtain 80% power with a significance level of 0.05 varied substantially across behavioural tests. In this comprehensive summary of the use of animal models of CIPN, we have identified areas in which the value of preclinical CIPN studies might be increased. Using both sexes of animals in the modelling of CIPN, ensuring that outcome measures align with those most relevant in the clinic, and the animal's pain contextualised ethology will likely improve external validity. Measures to reduce risk of bias should be employed to increase the internal validity of studies. Different outcome measures have different statistical power, and this can refine our approaches in the modelling of CIPN.

Author summary

Many frequently used and effective cancer chemotherapies can cause a disabling side effect that features pain, numbness, tingling, and sensitivity to cold and heat in the extremities known as chemotherapy-induced peripheral neuropathy (CIPN). There are currently no effective therapies to treat or prevent this condition, and animal models have been developed to address this. It is important that experiments using animal models of CIPN are robust and valid if they are to effectively help patients. We used a systematic approach to identify all 337 studies that have been published describing the use of animal models of CIPN. We were able to identify that many studies are imperfect in their experimental design, use only male animals, and assess outcomes with limited relevance to the human condition. Based on a meta-analysis, we provide guidance to the CIPN animal modelling community to guide future experiments that may improve their utility and validity.

Introduction

Chemotherapy-induced peripheral neuropathy (CIPN) is a disabling side effect of many frequently used and effective cancer chemotherapeutic agents and is known to impair daily function and diminish quality of life [1]. Frequently used chemotherapeutic agents reported to cause neurotoxic effects include platinum derivatives, taxanes [2], vinca alkaloids, epothilones, and also newer agents (e.g., thalidomide and bortezomib) [3]. The predominant sensory phenotype in patients exposed to oxaliplatin or docetaxel is distal symmetrical sensory loss affecting both upper and lower extremities. Symptoms of sensory disturbance reported by patients include paraesthesiae, numbness or tingling, and, less frequently, pain and cold allodynia [4]. CIPN can present clinically in 2 distinct forms: acute and chronic. The acute form is a chemotherapy dose-related, and often dose-limiting, polyneuropathy, which in many cases resolves in patients once the chemotherapy ceases. In some patients, this will persist, with other patients only developing symptoms after treatment has finished. A chronic, often painful, distal sensory neuropathy is still present in 33% of patients 1 year after completion of treatment [5]. No preventive or curative disease modifying treatments exist, and therefore there is a pressing need for more effective treatments [6].

Animal models of CIPN are used to investigate the pathophysiology of CIPN and to test potential therapies [7]. Frequently, chemotherapeutic agents are administered to induce a sensory neuropathy, and behavioural tests are used to assess induced sensory phenomena, such as evoked pain, and locomotor activity. Unfortunately, the conventional paradigm for drug development, in which findings are translated from preclinical animal research to clinical treatments, has been characterised by a lack of success [8,9]. Metaresearch from preclinical stroke research suggests that limitations in experimental design, conduct, analysis, and reporting—such as failure to carry out blinded assessment of outcome, randomisation and allocation concealment—may be impeding the development of effective therapies [10–13]. This led to the development of evidence-based guidelines for scientists [14]. These recommendations have been highly successful in transforming the reporting of measures to reduce risk of bias in the preclinical stroke field [15].

We have used a systematic review, in which we systematically identify and appraise all available evidence relevant to a predefined research question to provide a complete and unbiased summary of available evidence. We seek to establish the extent to which limited experimental biases influence the preclinical CIPN literature and to provide evidence to inform tactics to

increase the scientific validity of this research. Our aim is to provide a systematic overview of research in the field of in vivo animal modelling of CIPN, with a focus on the reporting of pain-related behavioural outcome measures, to provide useful information for preclinical researchers wishing to improve the design of experiments and refine the in vivo modelling of painful neuropathy.

Results

Identification of publications

Our initial systematic search (September 2012) identified 33,184 unique publications, of which 6,506 were identified as reporting in vivo models of painful neuropathy. This screening stage took 18 person months; 180 of these publications reported models of CIPN (Fig 1).

In the updated search (November 2015), we identified a further 11,880 publications. Using machine learning and text mining, we identified 6,108 publications as likely to report models of neuropathic pain, and 928 of these reported models of CIPN. In a random 10% sample of screened publications ($n = 1,188$), the classifier with the best fit—using stochastic gradient descent—had a screening performance of 97% sensitivity, 67% specificity, and 56% precision. Further details of the different machine-learning approaches applied are available [16]. Of the 928 studies identified to report animal models of CIPN, 157 met our inclusion criteria.

From both searches, a total of 337 unique publications are included in this review. The rate of new publications per year is shown in S1 Fig. Metadata from the 337 publications included in this study are available on the Open Science Framework (OSF; https://doi.org/10.17605/OSF.IO/ZJEHY). To address concerns that the systematic search is dated we performed a cumulative meta-analysis, in a post hoc analysis, of the effect sizes and tau^2 estimates (an estimate of between-study heterogeneity), ordered by year of publication. It appears that the data are mature and stable from around 250 studies onwards (S2 Fig).

To investigate sources of heterogeneity we divided the reporting of results by type of study (i.e., modelling experiments or intervention experiments), and by type of outcome measures reported (i.e., pain-related behaviours or other behaviours). Therefore, we have 4 datasets: (i) Data set 1a—modelling of CIPN and reporting pain-related behavioural outcome measures, (ii) Data set 1b—modelling of CIPN and reporting other behavioural outcome measures, (iii) Data set 2a –effects of interventions in animal models of CIPN and reporting pain-related behavioural outcome measures, and (iv) Data set 2b—effects of interventions in animal models of CIPN and reporting other behavioural outcome measures.

Outcome measures

Across the 337 publications included, we extracted all behavioural outcome measure data. Pain-related outcome measures included evoked limb withdrawal to stimuli (mechanical, heat, cold, and/or dynamic mechanical touch), evoked limb withdrawal and/or vocalisation to pressure stimuli, evoked tail withdrawal to stimuli (cold, heat, and/or pressure), and complex behaviours, e.g., burrowing activity. Other outcome measures included assessment of locomotor function, memory, reward, and attention. Pain-related and other outcome measures for both modelling and intervention experiments were analysed separately (Fig 2). The full list of behavioural outcome measures and behavioural tests is given in Tables 1 and 2. The most frequently reported pain-related outcome measure was evoked limb withdrawal to mechanical stimuli, most frequently assessed using monofilaments (Table 1). The most frequently reported other behavioural outcome measure was locomotor function, with the rotarod apparatus used in most cases (Table 2).
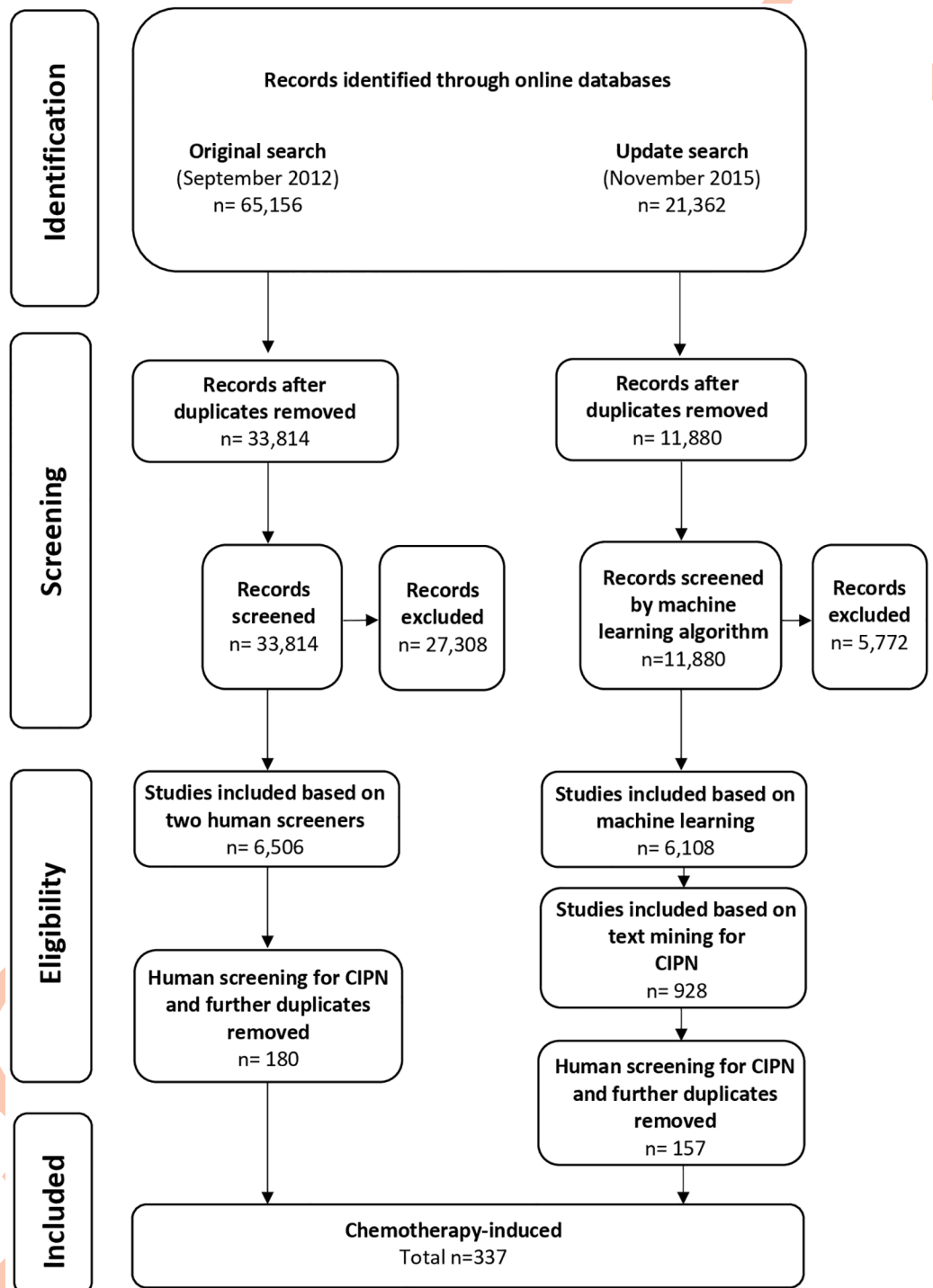
Fig 1. Flow diagram of included studies. CIPN, chemotherapy-induced peripheral neuropathy.

https://doi.org/10.1371/journal.pbio.3000243.g001

Interventions

A total of 306 different interventions were tested (S3 Fig). Most (80%) were only tested in 1 publication, and the most frequently reported interventions were gabapentin, morphine, and pregabalin, which were reported in 26, 22, and 11 publications, respectively.
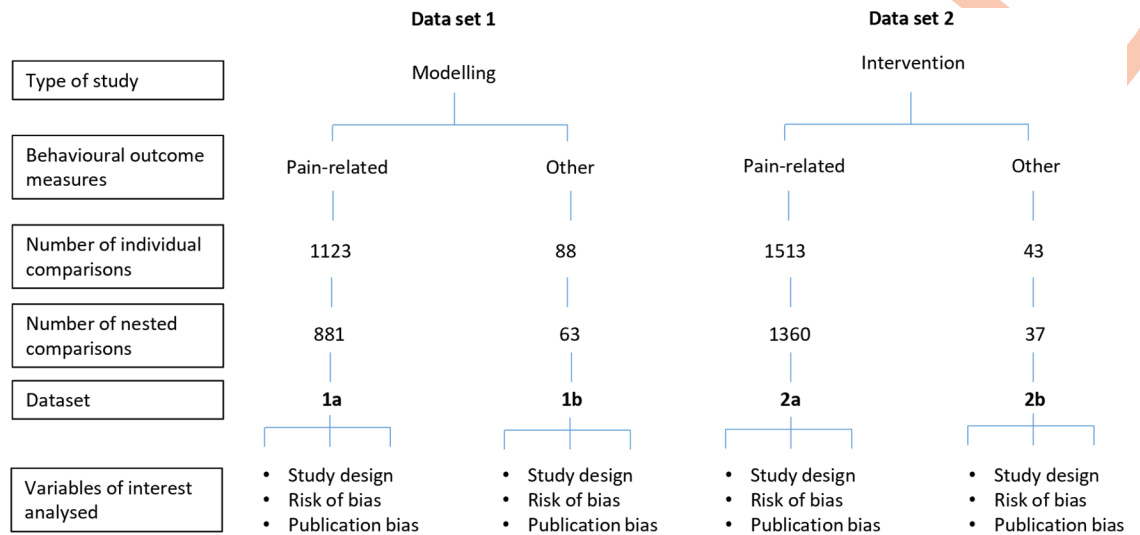
Fig 2. Description and summary of data sets included in this review.

Risk of bias

The reporting of measures to reduce risk of bias was 'moderate' across included studies ($n = 337$): 51.3% ($n = 173$) reported blinded assessment of outcome, 28.5% ($n = 96$) reported randomisation to group, 17.8% ($n = 60$) reported animal exclusions, 2.1% ($n = 7$) reported the use of a sample size calculation, and 1.5% ($n = 5$) reported allocation concealment.

Across all included studies, 49.6% ($n = 167$) reported a conflict of interest statement, and 96.7% ($n = 326$) reported compliance with animal welfare regulations (Table 3).

The methods used to implement randomisation and blinding, and the methods and assumptions for sample size calculations, were rarely reported: 6 publications reported that animals were randomly allocated to experimental groups using randomly generated number sequences, and 2 publications reported that this was done by block randomisation (8.3% of those that reported randomisation; 8 out of 96). One publication reported that randomisation was performed by 'picking animals randomly from a cage', which we do not consider a valid method of randomisation [17]. Nine publications reported that blinded outcome assessment was achieved by using a different experimenter to perform assessments, and 2 publications reported that a group code was used (6.4% of those that reported blinding; 11 out of 173). One study reported that allocation concealment was achieved using a coded system (20% of those that reported allocation concealment; 1 out of 5). Methods of sample size calculation were reported by 5 publications (71.4% of those that reported a sample size calculation; 5 out of 7): 3 used published or previous results from the group, and 2 had performed a pilot study to inform sample size calculations.

Modelling experiments

Animal studies modelling CIPN: Pain-related behavioural outcome measures (data set 1a). In modelling experiments using pain-related behavioural outcome measures, administration of a chemotherapeutic agent led to increased pain-related behaviour compared to sham controls (-2.56 standard deviation [SD] [95% CI -2.71 to -2.41], $n = 881$ comparisons). Species did not account for a significant proportion of the heterogeneity, and therefore mouse and rat experiments were analysed together (mice: -2.63 SD [95% CI -2.86 to -2.39], $n = 337$

Table 1. Pain-related behavioural outcome measures across intervention and modelling experiments. Numbers indicate the number of individual comparisons.

Pain-related behavioural outcome	Modelling	Intervention
	Number of comparisons	
ELW: Mechanical	**534**	**750**
Dynamic plantar aesthesiometer	1	2
Electronic "von Frey"	66	143
Mechanical monofilament	458	593
Normal non-noxious palpation of the tibia	1	0
Pin prick	8	12
Pinch test	2	2
ELW: Cold	191	264
Acetone/ethylchloride/menthol	110	140
Cold plate	73	117
Cold probe	1	0
Cold tolerance	0	1
ELW: cold water	6	6
Limb immersion: cold	1	0
ELW or vocalisation: Pressure	**149**	240
ELW or vocalisation: pressure	149	240
ELW: Heat	135	146
ELW: radiant heat	101	84
Hot plate	34	61
Paw immersion: heat	0	1
ETW: Cold	**43**	**51**
Tail immersion: cold	43	51
ETW: Heat	**54**	**44**
Tail flick: heat	33	32
Tail immersion: heat	21	12
Complex behaviour	**8**	**12**
AITC-evoked nocifensive behaviour	1	2
Burrowing activity	2	0
Capsaicin-evoked nocifensive behaviour	1	0
Conditioned place preference: pain-related	0	2
Orofacial operant assessment: cold	2	4
Thermal place preference: cold	2	4
ELW: Dynamic mechanical touch	**5**	**4**
Rough brush	2	2
Soft brush/cotton tip/ball	3	2
ETW: Pressure	**2**	**0**
Tail pressure test	2	0

Frequent descriptions for ELW or vocalisations: pressure (with the Randall-Selitto/paw pressure test). Frequent description for ELW: radiant heat (with Hargreave's).

Abbreviations: AITC, TRPA1 agonist allyl isothiocyanate; ELW, evoked limb withdrawal; ETW, evoked tail withdrawal.

https://doi.org/10.1371/journal.pbio.3000243.t001

Table 2. Other behavioural outcome measures across intervention and modelling experiments. Number of individual comparisons.

Other behavioural outcome	Modelling	Intervention
	Number of comparisons	
Locomotor function	**73**	**37**
Rotarod: locomotor function	34	23
Grip test: motor strength	14	1
Locomotor activity	10	6
Open field: distance travelled	4	0
Open field: exploratory behaviour	2	2
Open field: loading on hind limbs	2	2
Open field: immobility	2	0
Open field: locomotor activity	2	0
Catwalk: gait alterations	2	0
Balance beam: motor coordination	1	1
Cannabinoid tetrad test	0	1
Cannabinoid tetrad test: motility	0	1
Memory	**8**	**0**
Novel location preference: preference score	4	0
Novel object preference: preference score	4	0
Reward	**4**	**6**
Conditioned place preference: rewarding effects	4	6
Attention	**3**	**0**
Prepulse inhibition: attention	3	0

https://doi.org/10.1371/journal.pbio.3000243.t002

comparisons; rats: −2.52 SD [95% CI −2.71 to −2.32], n = 544 comparisons; Q = 1.16, df = 1, p = 0.28).

Study design. The type of pain-related outcome measure accounted for a significant proportion of the heterogeneity (Q = 307.27, df = 8, $p \leq 0.01$; Fig 3A).

We identified 12 different chemotherapeutic agents used to model CIPN in animals (Table 4). The chemotherapeutic agent used accounted for a significant proportion of the heterogeneity observed (Q = 174.26, df = 11, $p < 0.01$; Fig 3B). Sex accounted for a significant proportion of the heterogeneity (Q = 137.11, df = 3, $p < 0.01$; Fig 4A).

The time to assessment did not account for a significant proportion of heterogeneity (τ^2 = 2.55, I^2 = 85.98%, p = 0.999).

In a post hoc analysis, we found that the strain of animal accounted for a significant proportion of the heterogeneity (Q = 269.58, df = 22, $p < 0.01$; Fig 4B). The most frequently reported strain was Sprague Dawley rats (−2.43 SD [95% CI −2.61 to −2.26], n = 437 comparisons).

Statistical power of different outcome measures. The number of animals required to achieve 80% power with a significance level of 0.05 varied substantially across the behavioural tests. For the most frequently reported behavioural tests—mechanical monofilaments, Randall-Selitto paw pressure test, electronic 'von Frey', acetone test/ethyl chloride spray, cold plate, and Plantar Test (Hargreave's method)—we calculated the number of animals required in model and sham groups.

When both standardised mean difference (SMD) effect sizes and pooled SD were at the 50th percentile, the number of animals required ranged from 5 (electronic 'von Frey') to 75 per group (Randall-Sellito paw pressure test) (Fig 5). With an effect size at the 20th percentile and a variance at the 50th percentile, the number of animals required ranged from 13 (electronic 'von Frey') to 297 (Randall-Selitto paw pressure test) (Fig 5), demonstrating that some

Table 3. Reporting of measures to reduce the risk of bias and reporting.

	Risk of bias						Reporting	
	Blinded assessment of outcome	Allocation concealment	Random allocation to group	Sample size calculation	Animal exclusions		Conflict of interest statement	Compliance with animal welfare regulations
% Reporting (Number/337)	51.3 (173)	1.5 (5)	28.5 (96)	2.1 (7)	17.8 (60)		49.6 (167)	96.7 (326)

https://doi.org/10.1371/journal.pbio.3000243.t003

behavioural tests have less sensitivity to detect small effect sizes. The values for the 20th, 50th, and 80th percentiles of SMD effect sizes and SDs for each behavioural test are available on the OSF (https://doi.org/10.17605/OSF.IO/ZJEHY).

Risk of bias. Reporting of blinded assessment of outcome (Q = 33.62, df = 1, $p < 0.007$) and animal exclusions (Q = 28.99, df = 1, $p < 0.007$) accounted for a significant proportion of

Fig 3. Impact of study design in modelling experiments using pain-related behavioural outcomes (data set 1a). The size of the squares represents the number of nested comparisons that contribute to that data point and the value *N* represents the number of animals that contribute to that data point. (A) Outcome measure accounted for a significant proportion of the heterogeneity. ETW, ELW, and complex behaviours used to measure pain. (B) Chemotherapeutic agent accounted for a significant proportion of the heterogeneity. ELW, evoked limb withdrawal; ETW, evoked tail withdrawal; SMD, standardised mean difference.

https://doi.org/10.1371/journal.pbio.3000243.g003

Table 4. Model details including chemotherapeutic agents, route of administration, median cumulative dose, and upper and lower quartiles (data set 1a).

Animal	Type	Chemotherapeutic agent	Route of administration	Median	Q1	Q3
Mouse	Hydroxamate histone deacetylase inhibitors	JNJ-26481585	Subcutaneous	32.5	23.75	41.25
	Other	Salinomycin	Intraperitoneal	140	140	140
	Platinum compounds	Cisplatin	Hindpaw	0.004	0.002	0.022
			Intraperitoneal	20.5	8.25	23
			Subcutaneous	0.15	0.15	0.15
		Oxaliplatin	Hindpaw	0.04	0.04	0.04
			Intraperitoneal	10	3	30
			Intravenous	23	13.5	31
			Subcutaneous	0.04	0.04	0.04
	Proteasome inhibitors	Bortezomib	Intraperitoneal	0.75	0.425	1.95
			Intravenous	6.4	6.4	6.4
			Subcutaneous	12	12	18
	Taxanes	Paclitaxel	Intraperitoneal	10	5	16
			Intravenous	190	165	230
			Subcutaneous	10	10	10
			Tail vein	75	75	75
	Vinca alkaloids	Vincristine	Intraperitoneal	0.7	0.2	3.9
Rat	Binding of growth factor inhibitors	Suramin	Intravenous	25	17.5	37.5
	Other	Methotrexate	Subcutaneous	3.75	3.75	3.75
	Platinum compounds	Carboplatin	Intraperitoneal	112.5	101.25	123.75
		Cisplatin	Intraperitoneal	10	5	15
			Intravenous	2	2	2
			Subcutaneous	8	8	8
		Oxaliplatin	Intraperitoneal	16	8	32
			Intravenous	2	2	8
	Proteasome inhibitors	Bortezomib	Intraperitoneal	0.7	0.6	1
			Intravenous	4.8	4.2	4.8
	Taxanes	Docetaxel	Intravenous	25	17.5	32.5
			Tail vein	10	10	10
		Paclitaxel	Intraperitoneal	8	7.25	8.5
			Intravenous	24.5	9.75	33.25
			Tail vein	8	8	8
			Intrathecal[a]	6	3.1	13
		Nanoparticle albumin-bound paclitaxel	Intravenous	26.8	25.4	28.2
	Vinca alkaloids	Vincristine	Intraperitoneal	0.95	0.5	1
			Intravenous	0.435	0.2125	0.75
			Subcutaneous	0.42	0.42	0.42
			Tail	0.625	0.5625	0.6875

[a] Drug doses are shown in mg/kg, except intrathecal paclitaxel, shown in ng.

the observed heterogeneity, although effect sizes for blinding are very similar between strata (in which strata refers to the subgroups of comparisons, i.e., blinded versus not blinded). Reporting of randomisation, allocation concealment, and sample size calculation did not account for a significant proportion of the observed heterogeneity (Fig 6); data table available on the OSF (https://doi.org/10.17605/OSF.IO/ZJEHY).
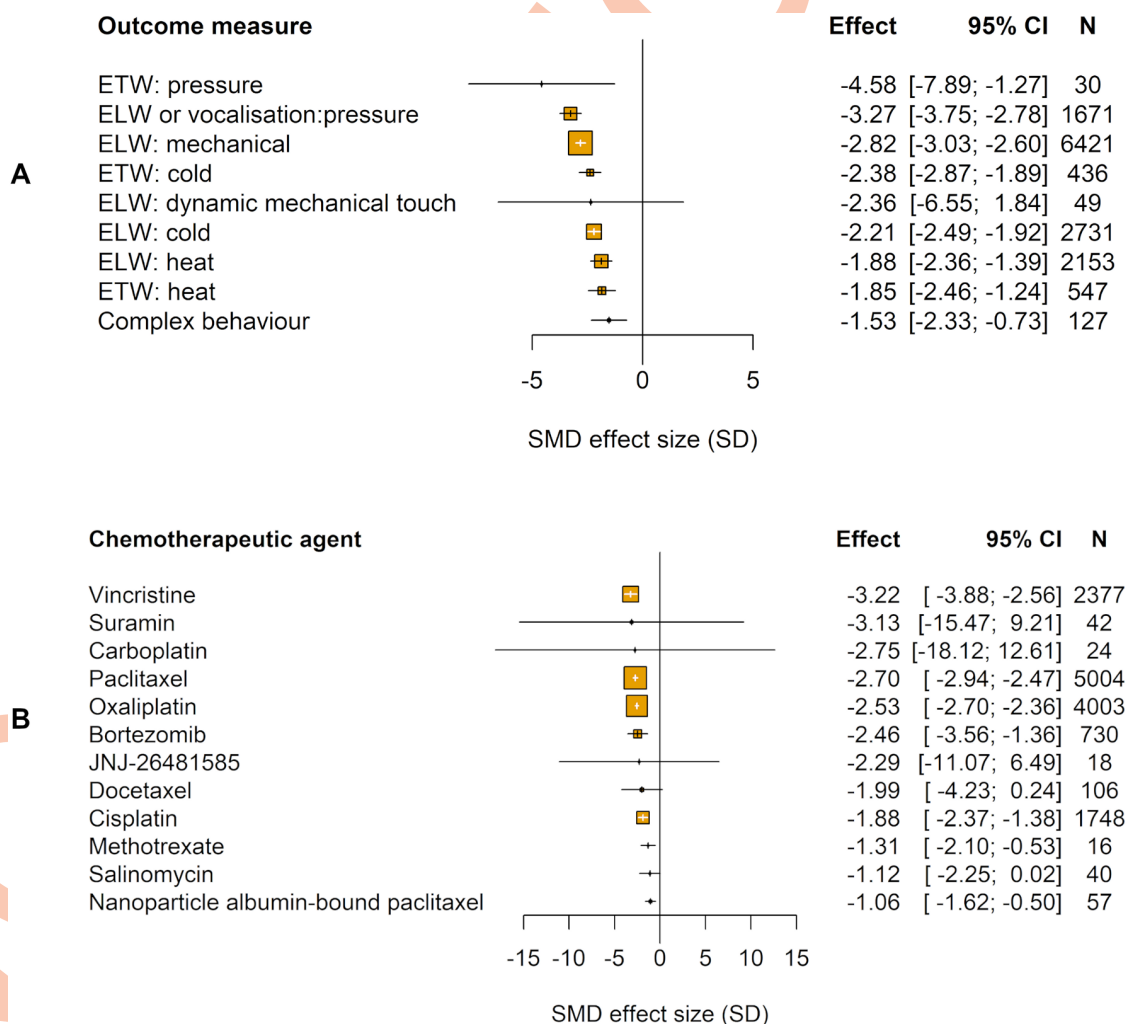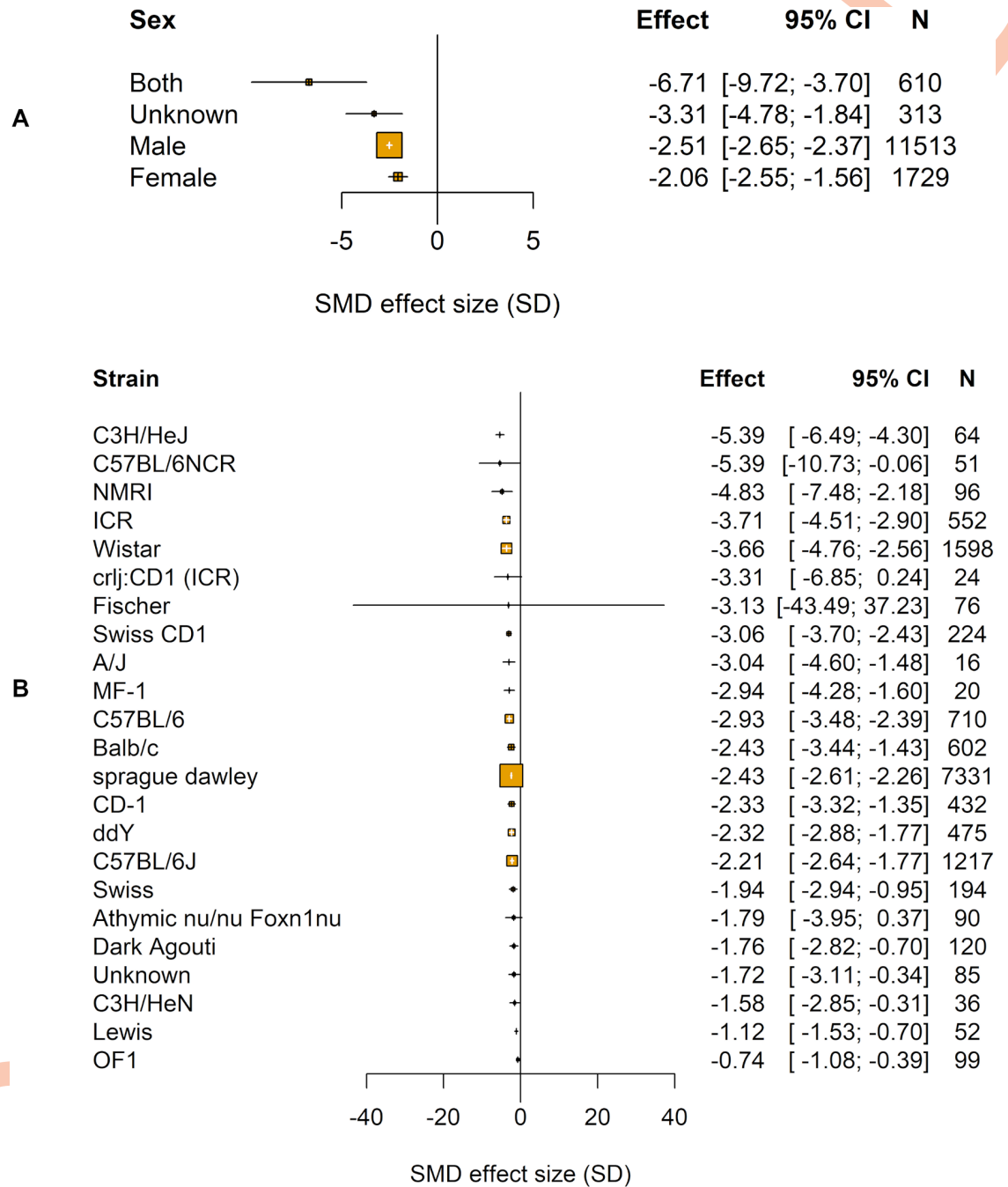
Fig 4. Impact of study design in modelling experiments using pain-related behavioural outcomes (data set 1a). The size of the squares represents the number of nested comparisons that contribute to that data point, and the value *N* represents the number of animals that contribute to that data point. (A) Sex accounted for a significant proportion of the heterogeneity. (B) Strain accounted for a significant proportion of the heterogeneity. SMD, standardised mean difference.

Compliance with animal welfare regulations accounted for a significant proportion of observed heterogeneity (Q = 19.44, df = 1, $p < 0.007$), although effect sizes are very similar between strata (Fig 7). Reporting of a conflict of interest statement did not account for a significant proportion of the heterogeneity (Fig 7).
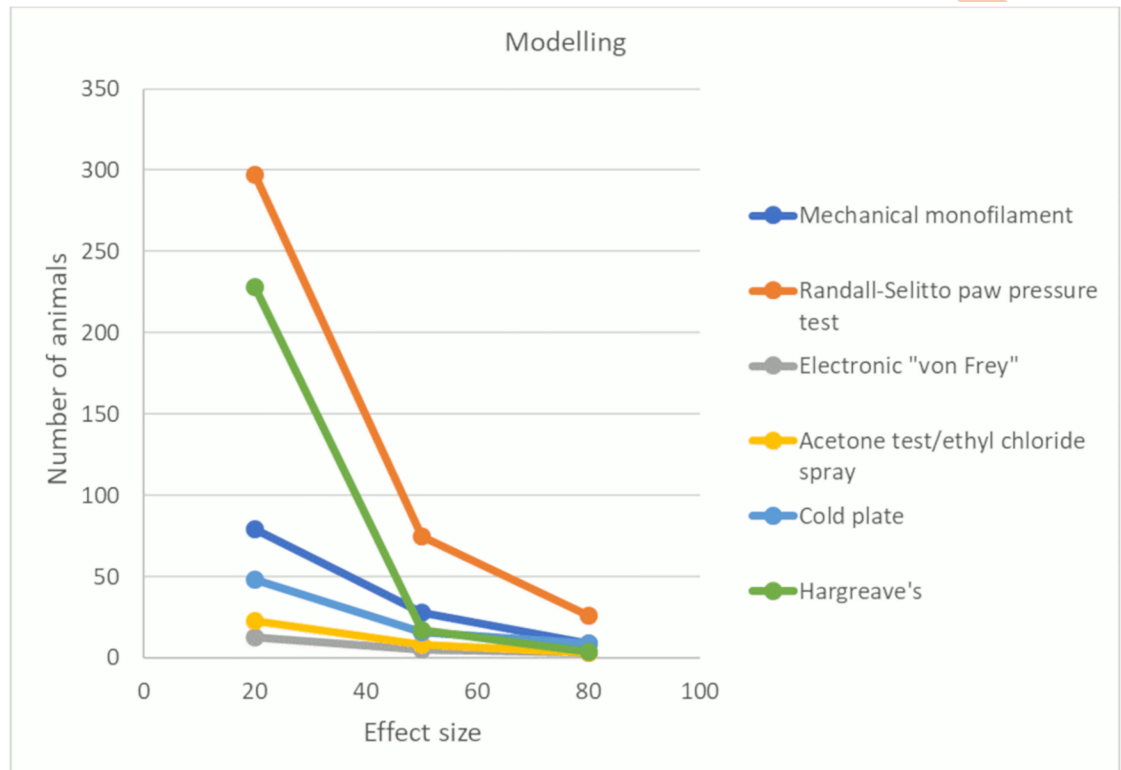
Fig 5. Power analysis for modelling experiments (data set 1a). Number of animals required per group to obtain 80% power with a significance level of 0.05 using mechanical monofilaments, Randall-Selitto paw pressure test, electronic 'von Frey', acetone test/ethyl chloride spray, cold plate, and Hargreave's. Effect sizes calculated by SMD. SMD, standardised mean difference.

Publication bias. There were 1,123 individual comparisons (−2.58 SD [95% CI −2.72 to −2.45]). Visual inspection of funnel plots indicated asymmetry, suggesting missing studies (Fig 8A). Trim and fill analysis imputed 316 theoretical missing studies on the right-hand side of the funnel plot (Fig 8B). Inclusion of these theoretical missing studies decreased the estimate of modelling-induced pain-related behaviour by 30% to −1.82 SD (95% CI −1.97 to −1.68). Furthermore, Egger's regression line and 95% CIs did not pass through the origin ($p = 6.85 \times 10^{-7}$), consistent with small study effects and again consistent with publication bias (Fig 8C).

Animal studies modelling CIPN: Other behavioural outcomes (data set 1b)

In addition, as a secondary outcome, we abstracted data from modelling experiments using other behavioural outcomes (locomotor function, memory, reward behaviours, and attention). Administration of chemotherapeutic agents led to increased pain-related behaviours compared to sham controls (−0.75 [95% CI −1.04 to −0.47], $n = 63$ comparisons). Species did not account for a significant proportion of the heterogeneity (Q = 3.29, df = 1, $p = 0.070$), and therefore rats and mice were analysed together.

Study design. Type of outcome measure accounted for a significant proportion of the heterogeneity (Q = 25.44, df = 3, $p < 0.01$; all figures related to data set 1b are available in S4A Fig).

Chemotherapeutic agent accounted for a significant proportion of the heterogeneity (Q = 28.90, df = 7, $p < 0.01$; S4B Fig).
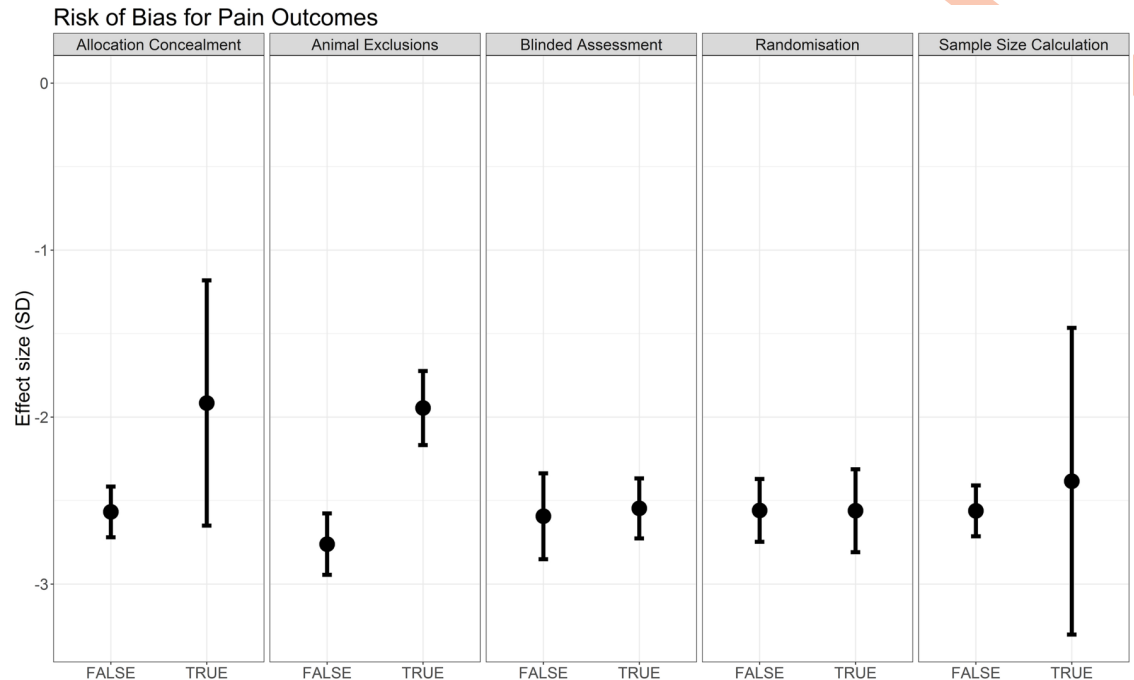
Fig 6. Effect sizes associated with measures to reduce risk of bias in modelling experiments using pain-related behavioural outcomes (data set 1a).

https://doi.org/10.1371/journal.pbio.3000243.g006

Sex ($Q = 3.29$, df = 1, $p = 0.70$) and time to assessment ($\tau^2 = 0.63$, $I^2 = 73.45\%$, $p = 0.05$) did not account for a significant proportion of the heterogeneity in modelling experiments using other behavioural outcomes.

A post hoc analysis found that strain accounted for a significant proportion of the heterogeneity ($Q = 23.98$, df = 6, $p < 0.01$; S4C Fig).

Risk of bias. Reporting of randomisation, blinding, allocation concealment, sample size calculation, or animal exclusions did not account for significant proportions of the heterogeneity (S5 Fig) nor did reporting of compliance with animal welfare regulations or a conflict of interest statement (S6 Fig).

Publication bias. There were 88 individual comparisons (−0.71 SD [95% CI −0.96 to −0.47]). Visual inspection of funnel plots indicated asymmetry, suggesting missing studies (S7A Fig). Trim and fill analysis imputed 21 theoretical missing studies on the right-hand side of the funnel plot (S7B Fig). Inclusion of these theoretical missing studies decreased the estimate of modelling-induced pain-related behaviour by 56% to −0.31 SD (95% CI −0.58 to −0.05). However, Egger's regression was not consistent with small study effects ($p = 0.293$) (S7C Fig).

Intervention experiments

Drug interventions in animal models of CIPN: Pain-related behavioural outcome measures (Data set 2a). In CIPN intervention studies using pain-related behavioural outcome measures, administration of an intervention led to a 1.53 SD (95% CI 1.45–1.61) attenuation of pain-related behaviour compared to control ($n = 1,360$ comparisons, $p < 0.007$). Species did not account for a significant proportion of the heterogeneity ($Q = 4.57$, df = 1, $p = 0.03$), and so mouse and rat experiments were analysed together.

Study design. The type of intervention accounted for a significant proportion of the heterogeneity ($Q = 1,418.27$, df = 304, $p < 0.007$). The most frequently tested interventions were

Reporting Criteria for Pain Outcomes
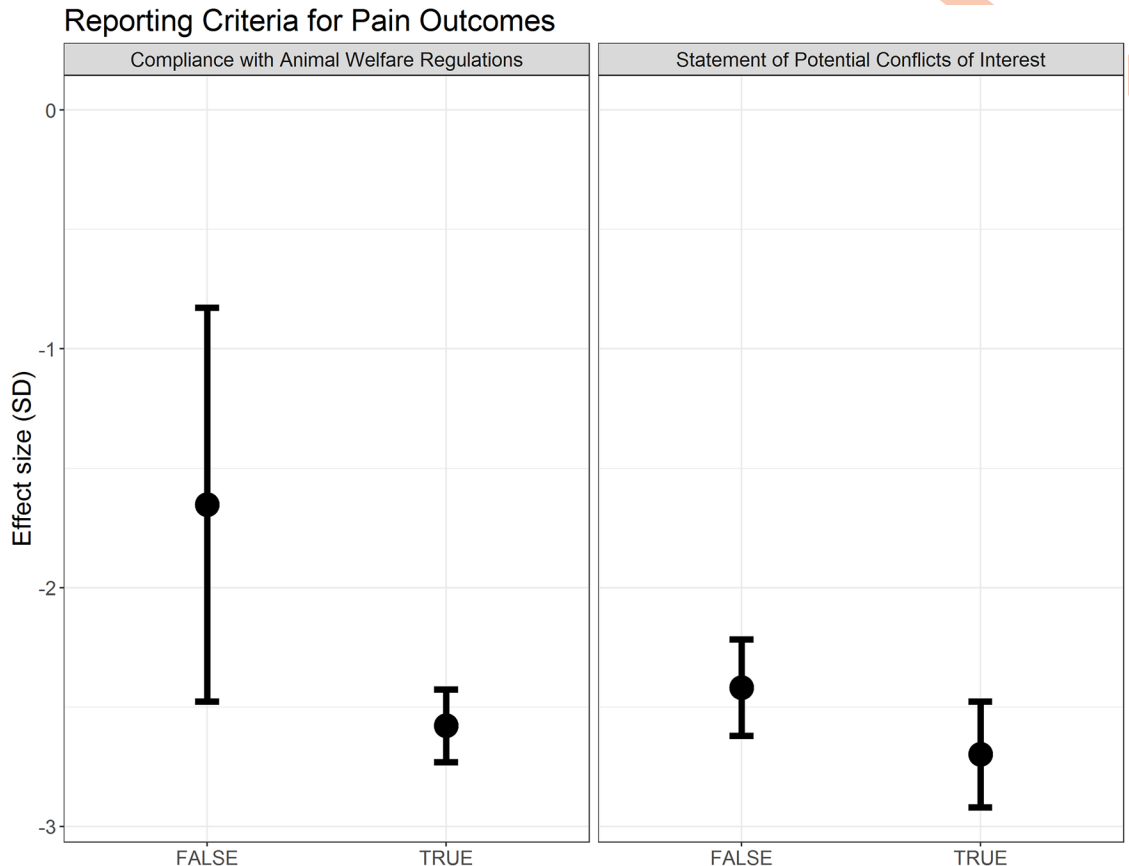
Fig 7. Effect sizes associated with reporting of compliance with animal welfare regulations and a statement of potential conflict of interests in modelling experiments using pain-related behavioural outcomes (data set 1a).

morphine ($n = 53$ comparisons), gabapentin ($n = 51$ comparisons), and pregabalin ($n = 35$ comparisons) (Fig 9). No clear dose-response relationship was observed for any of these drugs, investigated by calculating the cumulative dose.

The type of pain-related outcome measure accounted for a significant proportion of the heterogeneity (Q = 24.36, df = 7, $p < 0.007$; Fig 10A).

In intervention studies, the chemotherapeutic agent used to induce the pain model accounted for a significant proportion of the heterogeneity (Q = 22.51, df = 6, $p < 0.007$; Fig 10B). The most frequently reported chemotherapeutic agents were paclitaxel ($n = 520$) and oxaliplatin ($n = 480$).

Sex of animal did not account for a significant proportion of the heterogeneity (Q = 3.27, df = 3, $p = 0.35$).

Time to assessment accounted for a significant proportion of the heterogeneity, with a longer interval associated with greater attenuation of pain-related behaviour ($p < 0.007$; Fig 10C). However, time of intervention administration did not account for a significant proportion of the heterogeneity ($\tau^2 = 0.81$, $I^2 = 57.51\%$, $p = 0.5776$).

In a post hoc analysis, we found that the strain of animal accounted for a significant proportion of the heterogeneity (Q = 120.25, df = 19, $p < 0.007$; Fig 10D). The most frequently reported were Sprague Dawley rats ($n = 759$).

Ranking drug efficacy. We performed a post hoc analysis in which we compared the ranking of drugs common between a clinical systematic review [18] and our review. A
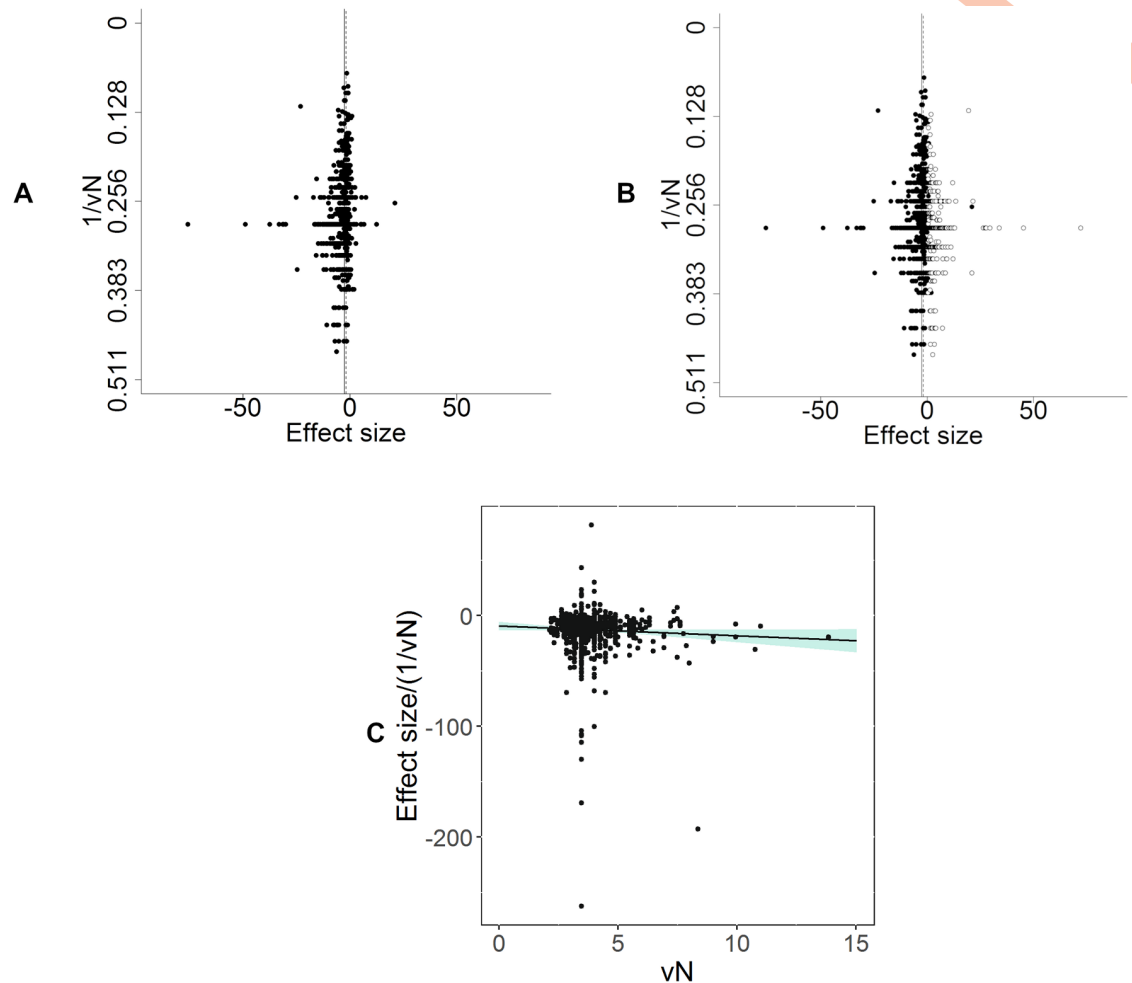
Fig 8. Assessment of publication bias in modelling experiments in which a pain-related outcome was used (data set 1a). (A) Visual inspection of the funnel plot suggests asymmetry. Filled circles represent reported experiments. Solid line represents global effect size, and dashed line represents adjusted global effect size. (B) Trim and fill analysis imputed theoretical missing studies (unfilled circles). Filled circles represent reported experiments. Solid line represents global effect size, and dashed line represents adjusted global effect size. (C) Egger's regression indicated small study effects.

https://doi.org/10.1371/journal.pbio.3000243.g008

Spearman's rank correlation coefficient found no correlation between clinical and preclinical rank ($r_s = -0.0099$, $p = 0.9699$; Fig 11).

Statistical power of different outcome measures. In intervention studies, the number of animals required to obtain 80% power with a significance level of 0.05 varied substantially across pain-related behavioural tests. For mechanical monofilaments, Randall-Selitto paw pressure test, electronic 'von Frey', acetone test/ethyl chloride spray, cold plate, and Plantar Test (Hargreave's method), we calculated the number of animals required in intervention and control groups.

When both the SMD effect size and pooled SD were the 50% percentile, the number of animals required ranged from 8 (acetone test/ethyl chloride spray) to 242 (Randall-Selitto paw pressure test) (Fig 12). With an effect size at the 20th percentile and a variance at the 50th percentile, the number of animals required increased substantially, ranging from 46 (Hargreave's) to 1,315 (Randall-Selitto paw pressure test). This again demonstrates that some behavioural tests have less sensitivity to detect small effect sizes. The values for the 20th, 50th, and 80th

Intervention	Effect	95% CI	N
CR4056	5.39	[2.39; 8.39]	147
AM1710	3.29	[2.12; 4.45]	138
10Panx	2.66	[1.75; 3.58]	112
PKC-epsilon-i	2.18	[0.67; 3.70]	91
Bulleyaconitine A	2.18	[1.18; 3.18]	110
Neurotropin	1.99	[1.36; 2.61]	136
Electroacupuncture	1.85	[1.46; 2.24]	145
WIN 55,212-2	1.71	[1.26; 2.17]	284
Mexiletine	1.65	[1.23; 2.06]	189
Gabapentin	1.64	[1.29; 1.99]	531
Morphine	1.58	[1.19; 1.97]	499
Lidocaine	1.57	[1.18; 1.97]	118
HC-030031	1.56	[0.91; 2.21]	318
Retigabine	1.52	[0.56; 2.48]	73
Capsazepine	1.51	[0.39; 2.63]	133
des-Arg9, [Leu8]-BK	1.51	[0.70; 2.31]	105
acetyl-L-carnitine	1.50	[1.27; 1.73]	273
Pregabalin	1.32	[0.68; 1.96]	324
Goshajinkigan	1.29	[0.85; 1.74]	177
Minocycline	1.21	[0.54; 1.87]	252
Fentanyl	1.17	[0.65; 1.68]	79
Lacosamide	1.13	[0.83; 1.43]	165
HOE-140	1.00	[0.33; 1.67]	95
Rotenone	1.00	[0.12; 1.88]	151
Amitriptyline	0.78	[0.08; 1.49]	131
milnacipran	0.72	[0.14; 1.30]	164
Pemirolast	0.68	[0.21; 1.15]	112

SMD effect size (SD)

Fig 9. Intervention accounted for a significant proportion of the heterogeneity in intervention experiments using pain-related behavioural outcomes (data set 2a). Plot shows interventions with 10 or more comparisons. The size of the squares represents the number of nested comparisons that contribute to that data point, and the value *N* represents the number of animals that contribute to that data point.

https://doi.org/10.1371/journal.pbio.3000243.g009

percentile of mean differences and pooled SDs for each behavioural test are provided on the OSF (https://doi.org/10.17605/OSF.IO/ZJEHY).

Risk of bias. Reporting of allocation concealment, animal exclusions, and sample size calculations accounted for a significant proportion of the heterogeneity (Fig 13; data table available on the OSF; https://doi.org/10.17605/OSF.IO/ZJEHY), with studies that did not report
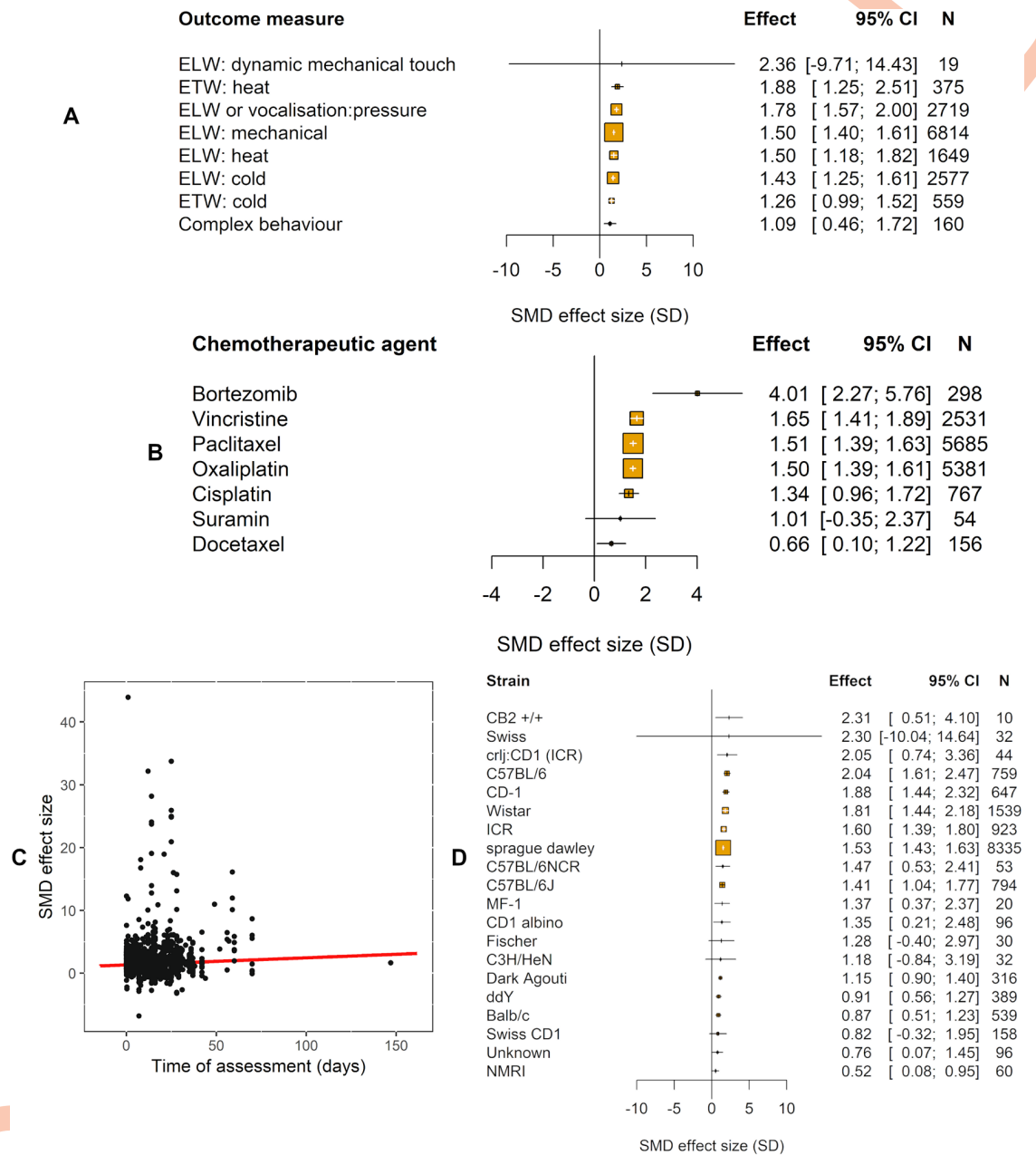
Fig 10. Impact of study design in intervention experiments using pain-related behavioural outcomes (data set 2a). The size of the squares represents the number of nested comparisons that contribute to that data point, and the value *N* represents the number of animals that contribute to that data point. (A) Outcome measure accounted for a significant proportion of the heterogeneity. (B) Chemotherapeutic agent accounted for a significant proportion of the heterogeneity. (C) Time of assessment accounted for a significant proportion of the heterogeneity. (D) Strain accounted for a significant proportion of the heterogeneity.

these items giving greater estimates of effect. Reporting of randomisation and blinded assessment of outcome did not account for a significant proportion of the heterogeneity.

Both reporting of compliance with animal welfare regulations (Q = 8.86, df = 1, $p < 0.007$) and reporting of a conflict of interest statement (Q = 8.28, df = 1, $p < 0.007$) accounted for a significant proportion of the heterogeneity (Fig 14). Failure to report this information was associated with smaller estimates of effect.

Fig 11. Rank order of clinical and preclinical drugs (data set 2a). A Spearman's correlation was run to assess the relationship between clinical and preclinical rank of 17 drugs. There was no correlation between clinical and preclinical rank; $r_s = -0.0099$, $p = 0.9699$.

https://doi.org/10.1371/journal.pbio.3000243.g011

Publication bias.   There were 1,513 individual comparisons (1.52 SD [95% CI 1.44–1.59]). Visual inspection of funnel plots indicated asymmetry, sugges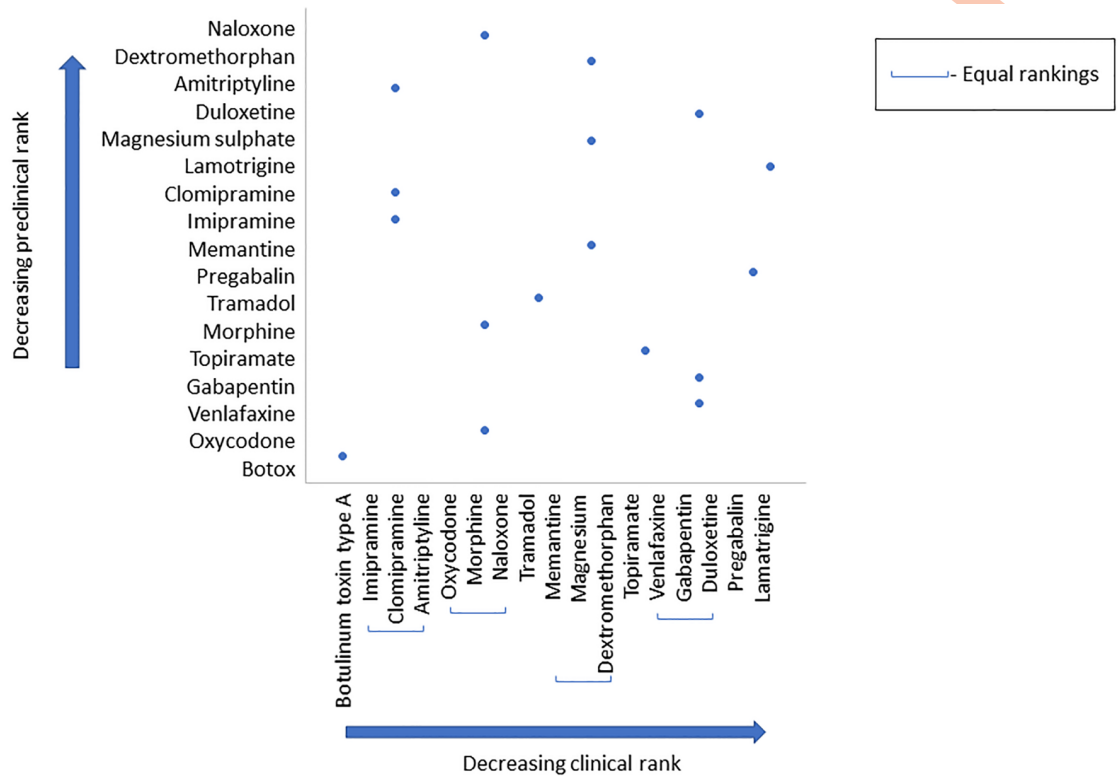ting missing studies (Fig 15A). Trim and fill analysis imputed 389 theoretical missing studies on the left-hand side of the funnel plot (Fig 15B). The inclusion of these theoretical missing studies decreased the estimate of intervention effects by 28% to 1.09 SD (95% CI 1.01–1.16). Furthermore, Egger's regression was consistent with small study effects ($p = 2.17 \times 10^{-6}$), suggesting funnel plot asymmetry (Fig 15C).

Drug interventions in animal models of CIPN: Other behavioural outcomes (data set 2b)

In intervention studies using other behavioural outcomes, administration of interventions led to improvement in other behaviours compared to controls (0.69 SD [95% CI 0.37–1.0], $n = 37$ comparisons). Species did not account for a significant proportion of the heterogeneity (Q = 0.75, df = 1, $p = 0.39$).

Study design.   Two outcome measures were used, and this accounted for a significant proportion of the heterogeneity. We observed greater improvement in reward-related behaviours compared with locomotor function (1.61 SD [95% CI 1.13–2.09], $n = 6$) comparisons versus (0.52 [95% CI 0.20–0.85], $n = 31$ comparisons [Q = 13.70, df = 1, $p < 0.007$]; S8A Fig).

The type of intervention accounted for a significant proportion of the heterogeneity (Q = 51.82, df = 19, $p < 0.007$; S9 Fig).
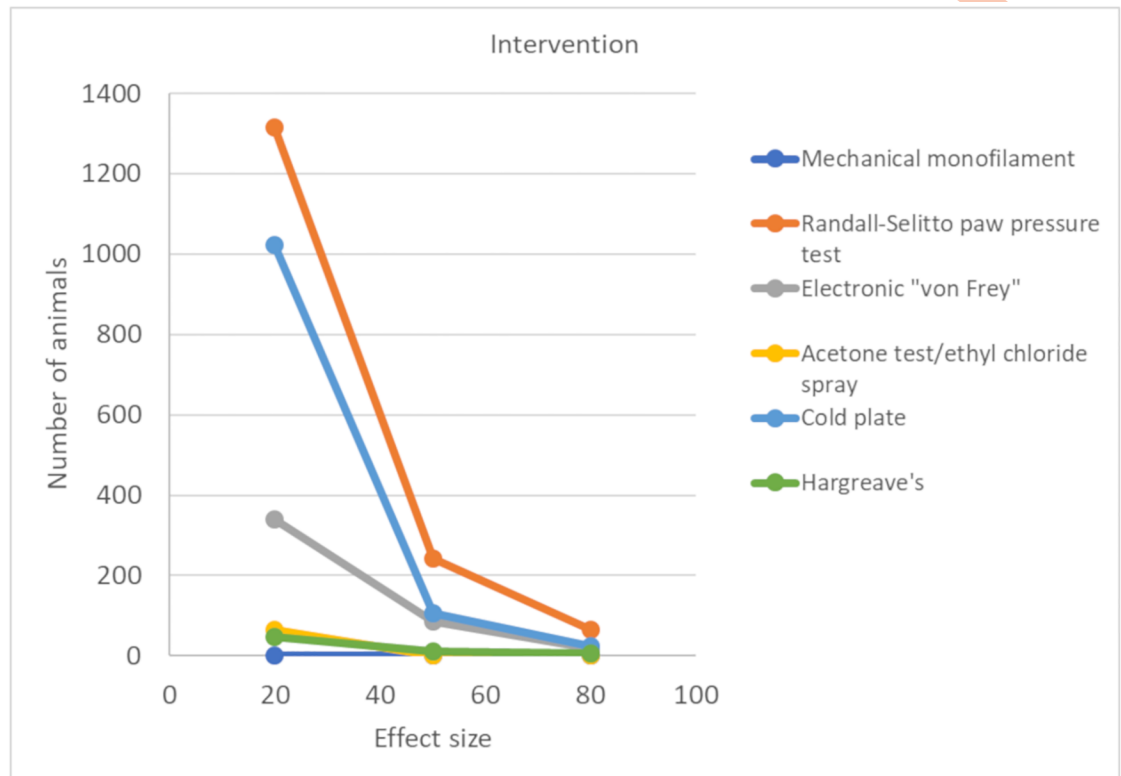
Fig 12. Power analysis for intervention experiments. Number of animals required to obtain 80% power with a significance level of 0.05 using mechanical monofilaments, Randall-Selitto paw pressure test, electronic 'von Frey', acetone test/ethyl chloride spray, cold plate, and Hargreave's (data set 2a). Effect sizes calculated by SMD. SMD, standardised mean difference.

https://doi.org/10.1371/journal.pbio.3000243.g012

Chemotherapeutic agent (Q = 2.21, df = 1, p = 0.137), sex (Q = 9.67, df = 2, p = 0.008), time to assessment (τ^2 = 0.37, I^2 = 54.21%, p = 0.398), and time of intervention administration (τ^2 = 0.32, I^2 = 52.37%, p = 0.331) did not account for a significant proportion of the heterogeneity.

In a post hoc analysis, we found that strain accounted for a significant proportion of the heterogeneity (Q = 16.18, df = 3, p < 0.007; S8B Fig).

Risk of bias. Blinded assessment of outcome accounted for a significant proportion of the heterogeneity (Q = 8.11, df = 1, p < 0.007) (S10 Fig). Reporting of randomisation or animal exclusions did not account for a significant proportion of the heterogeneity. No studies in this data set reported allocation concealment or the use of a sample size calculation. Reporting of a conflict of interest statement did not account for a significant proportion of the heterogeneity (S11 Fig). All studies in this data set reported compliance with animal welfare regulations.

Publication bias. There were 43 individual comparisons (0.70 SD [95% CI 0.41–0.99]). Visual inspection of funnel plots did not indicate asymmetry, suggesting no missing studies. Trim and fill analysis estimated no theoretical missing studies. Furthermore, Egger's regression was not consistent with small study effects (p = 0.352), suggesting funnel plot symmetry.

Animal husbandry

The reporting of details of animal husbandry was low across all included studies (S1 Table). No study reported whether different species were housed in the same room.
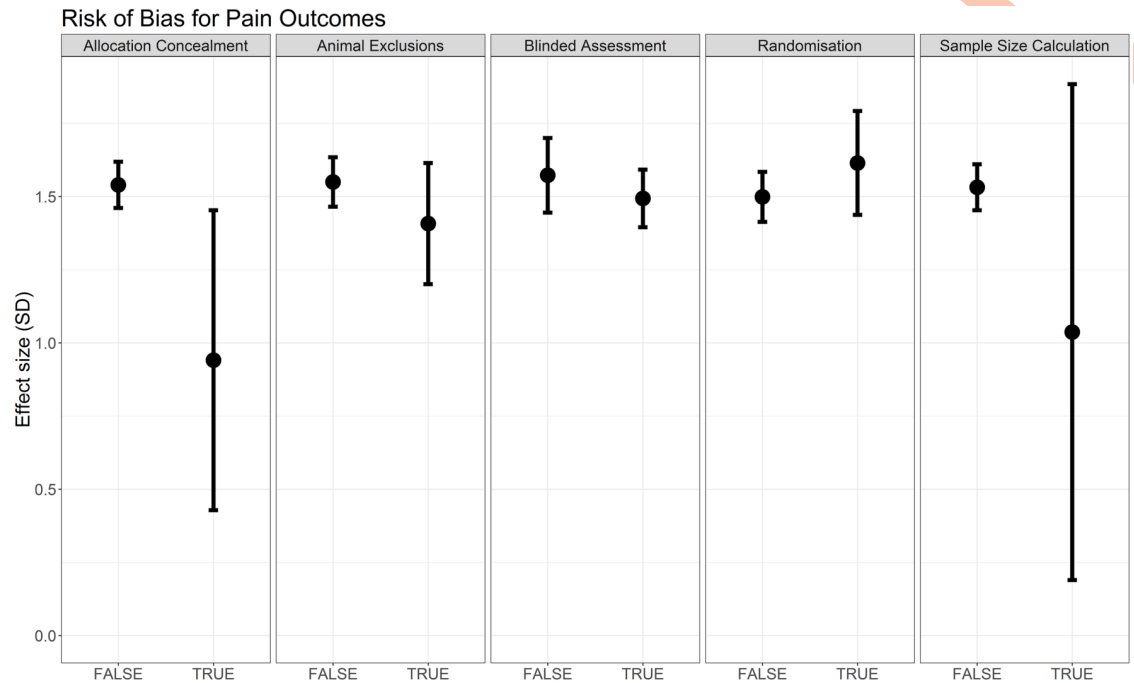
Fig 13. Effect sizes associated with measures to reduce risk of bias in intervention experiments using pain-related behavioural outcomes (data set 2a).

https://doi.org/10.1371/journal.pbio.3000243.g013

Fig 14. Effect sizes associated with reporting of compliance with animal welfare regulations and a statement of potential conflict of interests in intervention experiments using pain-related behavioural outcomes (data set 2a).
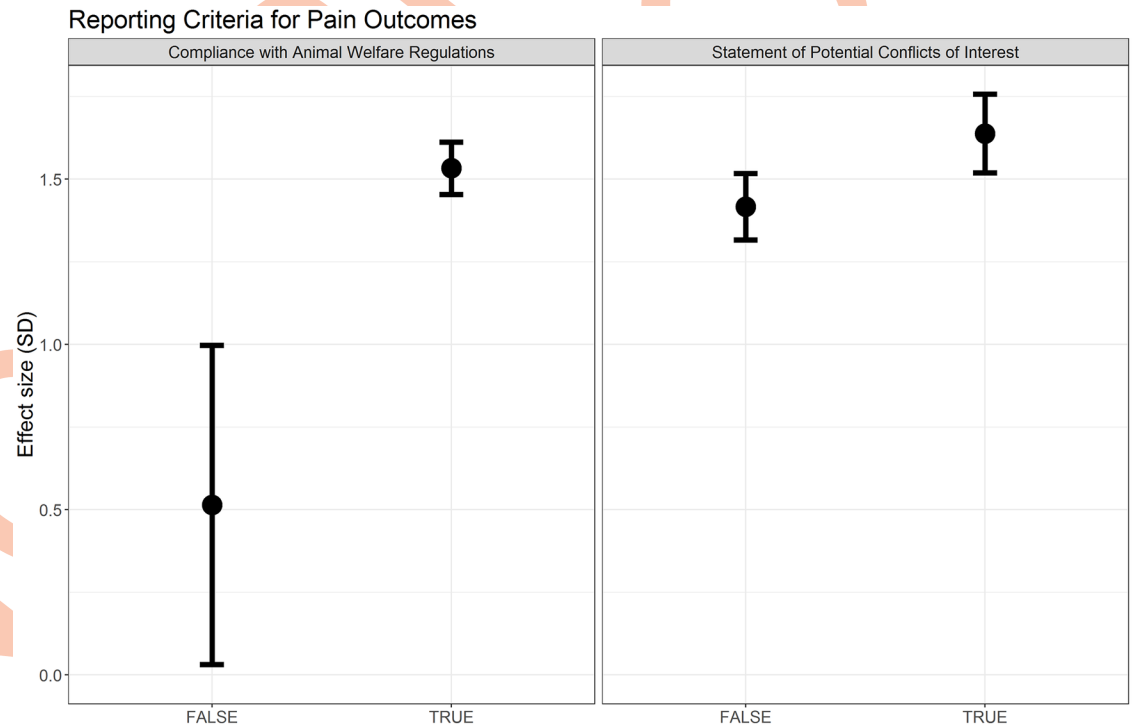
https://doi.org/10.1371/journal.pbio.3000243.g014
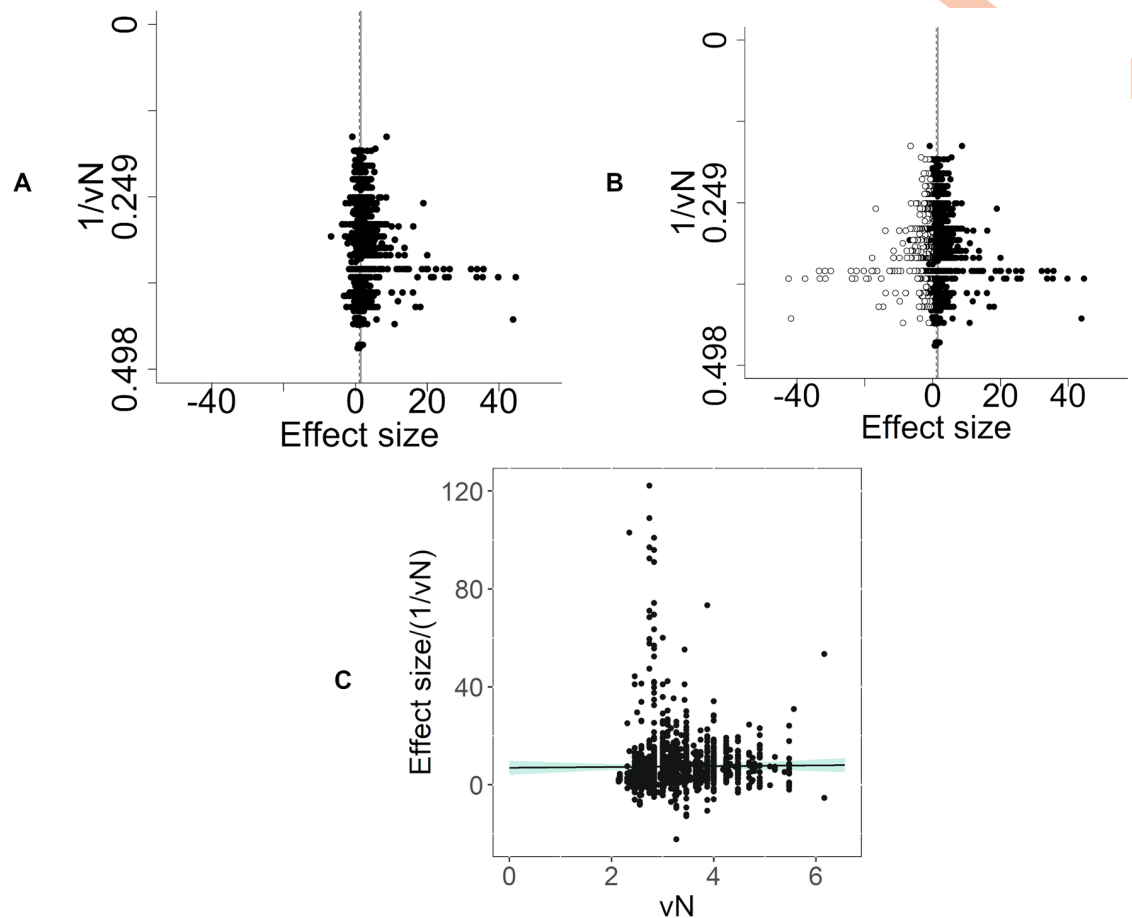
Fig 15. Intervention experiments in which a pain-related outcome was used (data set 2a). (A) Visual inspection of the funnel plot suggests asymmetry. Filled circles represent reported experiments. Solid line represents global effect size, and dashed line represents adjusted global effect size. (B) Trim and fill analysis imputed theoretical missing studies (unfilled circles). Filled circles represent reported experiments. Solid line represents global effect size, and dashed line represents adjusted global effect size. (C) Egger's regression indicated small study effects. vN, square root of N.

https://doi.org/10.1371/journal.pbio.3000243.g015

Discussion

Our systematic review and meta-analysis includes data from 337 publications describing animal models of CIPN. We demonstrate in modelling experiments that administration of a chemotherapeutic agent compared with sham controls leads to an increase in pain-related behaviours, and in intervention studies, drug administration attenuates pain-related behaviours.

Animal models of CIPN are used to elucidate the pathophysiology of the condition and to develop potential therapies. Our purpose here was to synthesise and summarise the entirety of the animal model CIPN literature primarily to make it accessible to scientists interested in the field and to provide them with data from which they can efficiently select optimal models to suit their experimental aims and to plan their experiments to a high level of rigour (e.g., suitably informed sample size calculations).

Here, we show—in 2 cohorts of primary studies, those modelling CIPN compared to sham controls and those testing the effect of intervention in animal models of CIPN—that there are some limitations in their experimental design. Our primary focus was on pain-related behaviours. Most studies used only male animals (84%) and evoked limb withdrawal to mechanical

stimuli. Reporting of measures to reduce risk of bias was moderate. Our indicative power calculations allow the ranking of the most commonly reported pain-related behavioural tests and suggest that the Randall-Sellito paw pressure test may be the least sensitive to detect small effect sizes. Our analyses also indicate likely publication bias and estimate an average of a 30% relative overestimation on reported results. This empirical evidence and our suggestions may generate discussion to guide the design of future studies and the importance of disseminating experimental findings irrespective of their direction of effect.

In undertaking this review, we observed increasing rates of publications describing primary studies of animal models of CIPN and that the accrual rate of relevant publications increased by 89% in 3 years. This presents technical challenges in synthesising a large data set in a timely manner. We were able to demonstrate the feasibility of using machine learning to facilitate screening for inclusion in systematic reviews of preclinical studies.

External validity of studies using animal models of CIPN

Misalignment between animal models and the clinical population. Most identified studies (285 out of 341 [84%]) used only male animals to model CIPN. In the clinic, the chemotherapeutic agents included in this analysis are frequently used to treat female cancer patients (e.g., ovarian or breast carcinoma), thus reducing the generalisability of the findings from these models to the clinical population. Sex only accounted for a significant proportion of the heterogeneity in the modelling of painful neuropathy in which pain-related behaviours were measured. It is likely that the paucity of female animals limits our ability to ascertain with sufficient power the impact of sex on models of CIPN in other contexts. Karp and colleagues have demonstrated that a large number of mammalian phenotypic traits are sexually dimorphic [19], and in line with National Institutes of Health policy [20], we advocate for the use of female animals in addition to males.

We also have concerns about the clinical relevance of the time courses frequently studied in animals. Acute CIPN is estimated to affect 68% of patients within the first month of chemotherapy cessation, 60% at 3 months, and 30% of patients at 6 months [5]. However, chronic CIPN has been observed in four-fifths of patients exposed to taxane [21] or oxaliplatin [22] approximately 2 years after treatment. A long-term study showed that oxaliplatin treatment was associated with CIPN at 6-year follow-up [23]. Therefore, the short duration of these models of CIPN identified in this systematic review likely model the acute phase. Of those publications in which study duration (the time between the first administration of chemotherapeutic agent and the time when animals were euthanised) was reported (39 of 341 publications), the median duration was only 21 days (16–28 IQR). Furthermore, the median time to outcome assessment in our modelling data set (the time in which there was the biggest difference between CIPN model and sham animals) was 14 days (7–25 IQR). The median time to outcome assessment in our intervention data set was also 14 days (7–22 IQR), indicating that the time in which the drug interventions are most effective is when the models show the largest modelling effect.

Misalignment between preclinical and clinical outcome measures. The most frequent behaviours reported in animal models of CIPN are manifestations of gain in sensory function; hypersensitivity in paw withdrawal evoked by mechanical stimuli was the assay most often employed. A review of studies reporting preclinical models of pain published in the journal *Pain* between 2000 and 2004 also found that the most frequently reported pain-related behaviours were such reflex withdrawal responses [24]. This contrasts with chronic CIPN in clinical practice, in which the predominant clinical sensory phenotype of these patients is one of sensory loss [4,25], and this may compromise the clinical relevance of these models for chronic CIPN; however, they may have more relevance to acute CIPN.

One approach to addressing this misalignment between outcome measures used to assess pain in patients in clinical trials and those frequently reported in animal models of CIPN would be the development of sensory profiling for rodent models of neuropathy that better reflect the clinical picture [25].

Internal validity of studies using animal models of CIPN

There was moderate reporting of measures to reduce risk of bias. Our subgroup analyses did not consistently identify that the reporting of these measures had an impact on experimental findings. It may be that there was insufficient power to test for these associations because of the small number of studies reporting these factors or that there is indeed no association. We also are only able to test the reporting of these measures to reduce risk of bias, and these may differ according to the actual use of these measures in the design, conduct, and analysis of a study. The details of methods used to implement randomisation and blinding and the methods and assumptions for sample size calculation were rarely reported. Despite the inconsistency of our findings, there is substantial empirical evidence of numerous research domains that these details are important to understand the validity of the procedures used [17], noting that one of the included studies reported that randomisation was achieved by selecting animals at random from the cage. If methods and assumptions were reported, this would allow assessment of the quality of these procedures that report using tools such as those used in clinical systematic reviews [26] and allow for more robust assessments of their impact on research findings.

Statistical modelling and meta-analysis have demonstrated that the exclusion of animals can distort true effects; even random loss of samples decreased statistical power, but if the exclusion is not random, this can dramatically increase the probability of false positive results [27]. It has been shown in other research fields that treatment efficacy is lower in studies that report measures to reduce risk of bias [13,28–30].

Publication bias

Our assessment of publication bias finds evidence to suggest that global effect sizes are substantially overstated in all data sets except the smallest (and this is likely due to reduced power to detect publication bias with only 37 studies). We observed relative overstatements in effect sizes that ranged from 28% to 56%. Publication bias is a prevailing problem in preclinical research, in which neutral or negative studies are less likely to be published than positive studies [31]. One potential reason for this is the high competition for academic promotion and funding, and few incentives to publish findings from studies in which the null hypothesis was not disproved. Initiatives such as Registered Reports provide one mechanism to support the publication of well-designed, thoroughly executed, and well-reported studies asking important questions regardless of the results.

Optimising experimental design

Experimental design of in vivo CIPN studies could be optimised by adopting measures to reduce risk of bias, such as using sample size calculations to ensure that experiments are appropriately powered. It is also important to use a model that best represents the clinical population of interest, for example, using both female and male animals. To help further address the issue of publication bias, we suggest that researchers make available prespecified protocols for confirmatory preclinical studies and publish all results. Others have shown that external validity may be increased by using multicentre studies to create more heterogeneous study samples, for example, by introducing variations in the animal genetics and environmental conditions

(housing and husbandry) between laboratories, an approach that may be useful in pain modelling.

One approach that would help optimise experimental design is to use the Experimental Design Assistant (EDA; https://eda.nc3rs.org.uk/), a free resource developed by the NC3Rs, whereby researchers create a record of their experimental design [32]. The output from the EDA could then be uploaded to the OSF as a record for transparency.

Reduction

There are opportunities to reduce waste and maximise the information gained from in vivo models of pain studies. This would require open and transparent reporting of results. For example, for complex behaviours, the online dissemination of individual animal video files [33] would allow reanalysis for further behaviours not reported in the original publication. It is interesting to note that although the open field was used in studies included in this systematic review, none of the included studies reported thigmotaxis, an outcome measure reported in other preclinical pain research. Sharing open field video files would allow this outcome to be assessed from previously conducted experiments. Our exemplar power calculations of the most frequently reported behavioural outcome measures highlight the substantial variability in the statistical performance of different outcome measures. Using these results, it is possible to rank the different pain-related behavioural tests according to how many animals are required per group as effect size or SD increases or decreases. This allows researchers to evaluate the sensitivity of their estimates of numbers required compared with variations in the effect sizes or variance achieved. Along with other factors, such as clinical relevance, these results can inform the choice of outcome measure in study design by allowing researchers to select outcome measures that require fewer animals.

The results of our systematic review show increasing rates of publications of experiments using animal models of CIPN. Between the initial search in 2012 and the updated search in 2015, the number of relevant publications increased by 89%. The high publication accrual rate is not unique to this field but is the case across clinical [34] and preclinical research; this makes it challenging for researchers and consumers of research to keep up to date with the literature in their field. This systematic review of preclinical models of pain uses machine learning and text mining and demonstrates the usefulness of these automation tools in this field.

Limitations

Conducting a systematic review is time and resource intensive, and the rate of publication of new primary research means that systematic reviews rapidly become outdated. This review is limited because the most recent information included was identified in November 2015. We plan that the present systematic review form a 'baseline' systematic review, which can be updated and developed into a living systematic review, i.e., one that is continually updated as new evidence is published [35]. An important secondary output of this review is the advances made in the use of machine learning to facilitate the automation of systematic reviews of preclinical studies. As new online platforms and tools for machine learning and automation become available, preclinical living systematic reviews become more feasible [36]. Guidelines for living systematic reviews [36] and the use of automation tools [37] have recently been published, and Cochrane has also launched pilot living systematic reviews [38,39].

The machine-learning algorithm based on our initial screening had a high sensitivity (97%) and medium specificity (67%). High sensitivity has a low risk of missing relevant literature. An algorithm with lower specificity is more likely to falsely identify studies for inclusion (i.e., false positives). As a result, during data abstraction, the 2 independent human screeners excluded

many studies identified by the machine for inclusion. We believe that this balance between sensitivity and specificity was appropriate because this reduced the risk of missing relevant studies.

A further possible limitation of our study is that we chose to extract behavioural data at the time point at which there was the largest difference between model and sham control animals or treatment and control animals. This time point was chosen to capture information on intervention effects regardless of their half-life. This limits what we can infer regarding the mismatch between timings, but we did also capture information on the first administration of intervention (relative to induction of the model) and the last administration. Future studies may use area under the curve approaches to capture response to model induction or drug intervention, but this was not possible for this large data set. There are tools under development for automation of data extraction, which may assist progress in this area [40].

In our meta-analysis, we grouped together the behavioural outcome measures that measure the same underlying biology. For example, in the case of experiments that reported using the grip test, 5 studies reported that the test was used to measure grip strength, and 1 reported that the test was used to measure muscle hypersensitivity [41]. For this reason, in our analysis, we grouped all grip test outcome measures together as a non–pain-related behavioural outcome measures. It is possible that the same tests or similar tests could be used and the same measurements reported as different outcomes; one test may also measure multiple facets of underlying biology. This is one of the challenges when analysing published data, and principle components analyses of large data sets such as these may help identify latent domains of behavioural outcome.

We only included studies in which the intervention drug was administered after or at the same time as the chemotherapeutic agent. Future literature reviews may consider drug interventions given before chemotherapeutic agents to determine whether prophylaxis can effectively prevent CIPN.

Unfortunately, the reporting of measures to reduce risk of bias was moderate in the studies included in this systematic review, which limits what we can infer from the results. We hope this review will highlight this issue in in vivo modelling of CIPN. Systematic review of animal experiments in other research areas has revealed low reporting of these measures and the negative impact of failure to report these measures across in vivo domains as diverse as modelling of stroke, intracerebral haemorrhage, Parkinson disease, multiple sclerosis, and bone cancer pain [15,29,30,42–44]. This has driven change, influencing the development of reporting guidelines [45], pain modelling specific guidelines [46], and the editorial policy of Nature Publishing Group [47]. However, requesting that submitting authors complete a reporting guideline without any other intervention is not associated with improved reporting [48]. After an initial review on the efficacy of interleukin-1 receptor antagonist in animal models of stroke highlighted low reporting of measures to reduce risk of bias [49], a subsequent review identified increased reporting of these measures [15], increasing the validity and reliability of these results. We hope that there will be a similar improvement in studies reporting the use of animal models of CIPN. We propose that if more studies implement and report measures to reduce the risk of bias, it will be possible to use a GRADE-type analysis to rate the certainty of the evidence of animal studies [50]. At present, any such approach is likely to lead to the majority of evidence being downgraded to the extent that no firm conclusions can be drawn. The measures to reduce risk of bias that we have assessed are largely derived from what is known to be important in clinical trials, and the extent to which these measures are important in animal studies has yet to be fully elucidated. However, reporting of these measures allows users of research to make informed judgments about the fidelity of the findings presented. Equally, it may be that there are other measures that are important in animal studies that we have not considered.

A recent study from our group has suggested that using SMD estimates of effect sizes with stratified meta-analysis has a moderate statistical power to detect the effect of a variable of interest when there are 200 included studies but that the false positive rate is low. This means that although we may not have sufficient power to detect an effect, we can have confidence that any significant results observed are likely to be true [51].

Conclusions

This systematic review and meta-analysis provides a comprehensive summary of the in vivo modelling of CIPN. The data herein can be used to inform robust experimental design of future studies. We have identified some areas in which the internal and external validity of pre-clinical CIPN studies may be increased; using both sexes of animals in the modelling of CIPN and ensuring outcome measures align with those most relevant in the clinic will likely improve external validity. Measures to reduce risk of bias should be employed to increase the internal validity of studies. Power analysis calculations illustrate the variation in group size under different conditions and between different behavioural tests and can be used to inform outcome measure choice in study design.

Materials and methods

This review forms part of a larger review of all in vivo models of painful neuropathy, and the full protocol is available at www.dcn.ed.ac.uk/camarades/research.html#protocols. Our review protocol predates the opening of the PROSPERO registry to reviews of in vivo preclinical data. Methods used were prespecified in the study protocol.

Search strategy

In September 2012, we systematically searched 5 online databases (PubMed, Web of Science, Biosis Citation Index, Biosis Previews, and Embase) with no language restrictions to identify publications reporting in vivo modelling of CIPN that reported a pain-related behavioural outcome measure. The search terms used for each database are detailed in S1 File. Search results were limited to animal studies using search filters [52,53]. Because we anticipated a high accrual rate of new publications, we ran an updated search in November 2015 and used machine learning and text mining to reduce the screening for inclusion workload. This updated search included 4 online databases (PubMed, Web of Science, Biosis Citation Index, and Embase) and used an updated animal filter [54]. Biosis Previews was no longer available.

Machine learning and text mining

We used machine learning to facilitate the screening of publications reporting animal models of CIPN and improve accuracy of the screening process [55]. The screening stage of a systematic review involves 'including' or 'excluding' publications identified in the search based on their title and abstract, and this was performed by 2 independent reviewers. The publications from our initial search (with 'include'/'exclude' decisions based on initial dual screening and differences reconciled by a third reviewer; inter-reviewer agreement Kappa = 0.95, standard error [SE] = 0.002) were used as a training set for machine learning approaches applied to the updated search.

Five machine learning groups participated, and 13 classifiers were created and applied to the updated search (validation set) [16]. We manually screened 10% of the updated publications (n = 1,188) and used this to assess the performance of these classifiers using measures of sensitivity, specificity, and precision as described by O'Mara-Eves and colleagues (2015) [56].

The reconciled decision of the human reviewers was considered the gold standard. We chose cut-off points such that the sensitivity of each classifier reached 0.95 and measured the resulting specificity and precision to choose the classifier that performed best for our data set.

To test the performance of the classifiers in the validation set, we used a random number generator to select a 10% random sample, and 2 independent investigators checked these for inclusion or exclusion. From the included studies in the updated search, we used text mining to identify studies reporting animal models of CIPN by searching for specific chemotherapy terms within the title and abstract of the identified publications; the inclusion of these studies was then verified by 2 independent reviewers.

Inclusion and exclusion criteria

We included controlled studies using pain-related behavioural outcome measures that either characterised models of neuropathy induced by chemotherapeutic agents or tested the effect of a drug intervention in such models (Fig 2). We required that studies report the number of animals per group, the mean, and a measure of variance (either the standard error of the mean [SEM] or the SD).

We excluded studies that administered the drug intervention before model induction, administered co-treatments, used transgenic models, or used in vitro models.

Measures to reduce risk of bias

We assessed the risk of bias of included studies by recording the reporting of 5 measures to reduce risk of bias at the study level: blinded assessment of outcome, random allocation to group, allocation concealment, reporting of animal exclusions, and a sample size calculation [57].

We also assessed the reporting of a statement of potential conflicts of interest and of compliance with animal welfare regulations [57,58].

Data abstraction

Data were abstracted to the CAMARADES Data Manager (Microsoft Access, Redmond, WA). For all included studies, we included details of publication (Table 5), animal husbandry, model, intervention, and other experiment details (Table 6). Outcome data presented graphically were abstracted using digital ruler software (Universal Desktop Ruler, AVPSoft.com or Adobe ruler) to determine values. When multiple time points were presented, we abstracted the time point that showed the greatest difference between model and control groups, or the greatest difference between treatment and control groups. If the type of variance (e.g., SEM or SD) was not reported, we characterised the variance as SEM because this is a more conservative approach in meta-analysis, in which studies are weighted in part by the inverse of the observed variance. All data were abstracted by 2 independent reviewers.

Data reconciliation

Publication and outcome level data abstracted by 2 independent reviewers were compared, and any discrepancies were reconciled. For outcome data, SMD effect sizes of individual comparisons were calculated for each reviewer's extracted data, and when these differed by $\geq 10\%$, they were identified for reconciliation. When individual comparisons differed by $< 10\%$, we took a mean of the 2 effect sizes and of the variance measure.

Table 5. Publication level data abstracted from each publication.

Publication level		
Metadata	**Risk of bias**	**Reporting quality**
• First author • Corresponding author • Year	• Random allocation • Allocation concealment • Blinded assessment • Sample size calculation • Animal exclusions	• Compliance with animal welfare regulations • Statement of potential conflict of interests

https://doi.org/10.1371/journal.pbio.3000243.t005

Data analysis

We separated the data according to those reporting the modelling of CIPN only and those testing the effect of an intervention in a model of CIPN. We analysed all the behavioural outcome measures reported. Behavioural outcome measures were separately considered as 'pain-related' or 'other (non–pain related)' behavioural outcome measures (Fig 2). This resulted in 4 data sets: (1) animal studies modelling CIPN: pain-related behavioural outcome measures (data set 1a), (2) animal studies modelling CIPN: other behavioural outcomes (data set 1b), (3) drug interventions in animal models of CIPN: pain-related behavioural outcome measures (data set 2a), and (4) drug interventions in animal models of CIPN: other behavioural outcomes (data set 2b). Data from individual experiments were extracted from each publication, and these are reported as 'individual comparisons'.

For each individual comparison, we calculated an SMD effect size. When more than one relevant behaviour was reported in the same cohort of animals, these individual comparisons were aggregated ('nested comparisons'; Fig 2) by behavioural subtype, determined by the site of stimulus application (e.g., limb or tail) and the modality of the stimulus used (e.g., mechanical or heat). Fixed-effects meta-analysis was used to give a summary estimate of these effects in each cohort. Cohort-level effect sizes were then pooled using a random-effects meta-analysis with restricted maximum-likelihood estimation of heterogeneity, in which heterogeneity refers

Table 6. Experiment level data abstracted from each publication.

Experiment level				
Animal	**Husbandry**	**Model**	**Intervention**	**Outcome measure**
• Species • Strain • Sex • Supplier • Country of supplier	• Type of diet • Cage rack ventilation • Type of bedding • Enrichment • Habituation time prestudy • Cage cleaning frequency • Number per cage • Shared cages between CIPN and sham animals • Number of hours in light cycle • Room temperature • Humidity • Food availability • Different species housed in same room	• Chemotherapeutic agent • Dose • Number of doses • Route of administration • Model duration • Weight • Age	• Time of drug administration • Dose • Route of administration	• Outcome measure type • Units • Larger values indicates better/worse • Number of groups served by control group • Time of assessment • Number of animals per group • Mean outcome • Variance • Number of treatment groups per control • Number of animal exclusions and reason

Abbreviation: CIPN, chemotherapy-induced peripheral neuropathy.

https://doi.org/10.1371/journal.pbio.3000243.t006

to the variation in study outcomes between studies. When a single control group served multiple comparator (model or treatment) groups, their contribution was adjusted by dividing the number of animals in the control group by the number of comparator groups served. The Hartung and Knapp method was used to adjust test statistics and confidence intervals; this calculates the confidence intervals using the following formula: effect size + t(0.975,k − 1) × SE. Results are presented in SMD units along with the 95% confidence intervals.

To provide empirical evidence to inform experimental design and refine modelling of CIPN, we assessed the extent to which predefined study design and study risk of bias characteristics explained observed heterogeneity. We used stratified meta-analysis for categorical variables and metaregression for continuous variables. The purpose of these subgroup analyses is to observe whether studies grouped together describing a similar characteristic (e.g., all studies using male animals versus all studies using female animals) differ in their overall estimates of effects. Such analyses provide empirical evidence of the impact of study design choices and are useful to design future experiments. The study design factors assessed using stratified meta-analysis were animal sex and species, therapeutic intervention, therapeutic intervention dose, methods to induce the model including the chemotherapeutic agent, and type of outcome measure. Because drug dose and route of administration are largely important in the context of the intervention being used, we did not assess the impact of dose or route of administration across different chemotherapeutic agents or drug interventions. We specified a priori that if species accounted for a significant proportion of heterogeneity, we would analyse the effect of study design factors on each species separately. If not, then all data would be analysed together. We also assessed the impact of reporting of measures to reduce bias. We used metaregression to assess the impact of time to assessment (defined as the interval between first administration of chemotherapeutic agent and outcome measurement) and time to intervention administration (defined as the interval between first administration of chemotherapeutic agent and administration of intervention). We used a meta-analysis online platform (code available here: https://github.com/qianyingw/meta-analysis-app) to perform all meta-analyses.

We applied a Bonferroni-Holmes correction for multiple testing that resulted in critical thresholds for significance as follows: in modelling experiments, $p < 0.01$ for study design features and $p < 0.007$ for reporting of measures to reduce risk of bias and measures of reporting; in intervention experiments, $p < 0.007$ for study design features, and $p < 0.007$ for reporting of measures to reduce risk of bias and measures of reporting.

Power analysis of in vivo modelling

To guide sample size estimation for future studies, we performed power calculations for the 6 most frequently reported behavioural tests. To do this, we separately ranked the observed SMD effect size and the pooled SD and, for each, identified the 20th, 50th, and 80th percentile. We then used these values to calculate the number of animals required in 9 hypothetical treatment and control groups. Calculations were based on the two-sample two-sided t test, with 80% power and an alpha value of 0.05.

Publication bias

We assessed for potential publication bias by assessing the asymmetry of funnel plots using visual inspection and Egger's regression [59]. We assessed for the impact of publication bias using Duval and Tweedie's trim and fill analysis [60,61]. We performed these assessments in 4 data sets separately and used individual comparisons rather than summary estimates for each cohort (Fig 1).

Comparison of intervention efficacy with that observed in human studies

In a clinical systematic review of neuropathic pain [18], selected analgesic agents had been ranked according to their efficacy, as measured by Number Needed to Treat (NNT) for 50% pain relief. If preclinical studies included in this review reported use of these agents or their analogues, we ranked the interventions according to their SMD effect size for attenuation of pain-related behaviour. We then assessed the correlation between clinical and preclinical rank using Spearman's rank correlation coefficient.

Supporting information

S1 Table. Reporting of animal husbandry details. The median habituation time was 7 days (7–7 IQR). The median number of animals per cage was 4 (2.5–4.5). Reporting of mixed housing with shams was always 'Not mixed'. Room temperature 22 ˚C (22 ˚C–23 ˚C IQR). Humidity 55 (53.75–55 IQR).
(DOCX)

S1 Fig. Number of included publications published each year.
(TIF)

S2 Fig. Cumulative meta-analysis of (A) effect sizes and (B) tau^2 estimates, ordered by year of publication.
(TIF)

S3 Fig. Tree plot of prevalence of interventions. A total of 306 different interventions reported.
(TIF)

S4 Fig. Impact of study design in modelling experiments using other behavioural outcomes. The size of the squares represents the number of nested comparisons that contribute to that data point, and the value *N* represents the number of animals that contribute to that data point. (A) Outcome measure accounted for a significant proportion of the heterogeneity. (B) Chemotherapeutic agent accounted for a significant proportion of the heterogeneity. (C) Strain accounted for a significant proportion of the heterogeneity.
(TIF)

S5 Fig. Effect sizes associated with measures to reduce risk of bias in modelling experiments using other behavioural outcomes.
(TIF)

S6 Fig. Effect sizes associated with reporting of compliance with animal welfare regulations and a statement of potential conflict of interests in modelling experiments using other behavioural outcomes.
(TIF)

S7 Fig. Modelling experiments using other behavioural outcomes. (A) Visual inspection of the funnel plot suggests asymmetry. Filled circles represent reported experiments. Solid line represents global effect size, and dashed line represents adjusted global effect size. (B) Trim and fill analysis imputed theoretical missing studies (unfilled circles). Filled circles represent reported experiments. Solid line represents global effect size, and dashed line represents adjusted global effect size. (C) Egger's regression was not consistent with small study effects.
(TIF)

S8 Fig. Impact of study design in intervention experiments using other behavioural outcomes. The size of the squares represents the number of nested comparisons that contribute to that data point, and the value *N* represents the number of animals that contribute to that data point. (A) Type of outcome measure accounted for a significant proportion of the heterogeneity. (B) Strain accounted for a significant proportion of the heterogeneity.
(TIF)

S9 Fig. In intervention experiments using other behavioural outcomes, type of intervention accounted for a significant proportion of the heterogeneity. The size of the squares represents the number of nested comparisons that contribute to that data point, and the value *N* represents the number of animals that contribute to that data point.
(TIF)

S10 Fig. Effect sizes associated with measures to reduce risk of bias in intervention experiments using other behavioural outcomes.
(TIF)

S11 Fig. Effect sizes associated with reporting of compliance with animal welfare regulations and a statement of potential conflict of interests in intervention experiments using other behavioural outcomes.
(TIF)

S1 File. Terms used in each database for systematic search.
(DOCX)

Acknowledgments

We would like to thank Dr. Zsanett Bahor for her support during the resubmission process.

Author Contributions

Conceptualization: Gillian L. Currie, Lesley Colvin, Malcolm Macleod, Andrew S. C. Rice, Emily S. Sena.

Data curation: Gillian L. Currie, Fala Cramond.

Formal analysis: Gillian L. Currie, Fala Cramond, Robert Stewart, Ezgi Tanriver-Ayder.

Funding acquisition: Gillian L. Currie, Lesley Colvin, Malcolm Macleod, Andrew S. C. Rice, Emily S. Sena.

Investigation: Gillian L. Currie, Helena N. Angel-Scott, Fala Cramond, Kaitlyn Hair, Laila Khandoker, Rosie Morland, Nicki Sherratt, Robert Stewart, Rachel Wodarski, Ran Xiong.

Methodology: Gillian L. Currie, Jing Liao, Malcolm Macleod, Andrew S. C. Rice, Emily S. Sena.

Project administration: Gillian L. Currie, Andrew S. C. Rice, Emily S. Sena.

Resources: Malcolm Macleod, Emily S. Sena.

Software: Jing Liao, James Thomas, Qianying Wang.

Supervision: Gillian L. Currie, Malcolm Macleod, Sarah K. McCann, Emily S. Sena.

Validation: Gillian L. Currie, Fala Cramond, Sarah K. McCann, Robert Stewart.

Visualization: Gillian L. Currie, Kaitlyn Hair.

Writing – original draft: Gillian L. Currie, Malcolm Macleod, Andrew S. C. Rice, Emily S. Sena.

Writing – review & editing: Gillian L. Currie, Helena N. Angel-Scott, Lesley Colvin, Fala Cramond, Kaitlyn Hair, Laila Khandoker, Jing Liao, Malcolm Macleod, Sarah K. McCann, Rosie Morland, Nicki Sherratt, Robert Stewart, Ezgi Tanriver-Ayder, James Thomas, Qianying Wang, Rachel Wodarski, Ran Xiong, Andrew S. C. Rice, Emily S. Sena.

References

1. Mols F, Beijers T, Vreugdenhil G, van de Poll-Franse L. Chemotherapy-induced peripheral neuropathy and its association with quality of life: a systematic review. Supportive care in cancer: official journal of the Multinational Association of Supportive Care in Cancer. 2014; 22(8):2261–9. https://doi.org/10.1007/s00520-014-2255-7 PMID: 24789421.

2. Song SJ, Min J, Suh SY, Jung SH, Hahn HJ, Im SA, et al. Incidence of taxane-induced peripheral neuropathy receiving treatment and prescription patterns in patients with breast cancer. Supportive care in cancer: official journal of the Multinational Association of Supportive Care in Cancer. 2017; 25(7):2241–8. Epub 2017/02/17. https://doi.org/10.1007/s00520-017-3631-x PMID: 28204996.

3. Wolf S, Barton D, Kottschade L, Grothey A, Loprinzi C. Chemotherapy-induced peripheral neuropathy: Prevention and treatment strategies. European Journal of Cancer. 2008; 44(11):1507–15. https://doi.org/10.1016/j.ejca.2008.04.018 PMID: 18571399

4. Ventzel L, Madsen CS, Karlsson P, Tankisi H, Isak B, Fuglsang-Frederiksen A, et al. Chronic Pain and Neuropathy Following Adjuvant Chemotherapy. Pain medicine. 2017. https://doi.org/10.1093/pm/pnx231 PMID: 29036361.

5. Seretny M, Currie GL, Sena ES, Ramnarine S, Grant R, Macleod MR, et al. Incidence, prevalence, and predictors of chemotherapy-induced peripheral neuropathy: A systematic review and meta-analysis. Pain. 2014; 155(12):2461–70. https://doi.org/10.1016/j.pain.2014.09.020 PMID: 25261162

6. Hershman DL, Lacchetti C, Dworkin RH, Smith EML, Bleeker J, Cavaletti G, et al. Prevention and Management of Chemotherapy-Induced Peripheral Neuropathy in Survivors of Adult Cancers: American Society of Clinical Oncology Clinical Practice Guideline. Journal of Clinical Oncology. 2014; 32(18):1941–67. https://doi.org/10.1200/JCO.2013.54.0914 PMID: 24733808.

7. Hoke A, Ray M. Rodent models of chemotherapy-induced peripheral neuropathy. ILAR journal. 2014; 54(3):273–81. https://doi.org/10.1093/ilar/ilt053 PMID: 24615440.

8. O'Collins VE, Macleod MR, Donnan GA, Horky LL, van der Worp BH, Howells DW. 1,026 experimental treatments in acute stroke. Annals of neurology. 2006; 59(3):467–77. https://doi.org/10.1002/ana.20741 PMID: 16453316.

9. Ioannidis JP. Why most published research findings are false. PLoS Med. 2005; 2(8):e124. https://doi.org/10.1371/journal.pmed.0020124 PMID: 16060722.

10. Macleod MR, O'Collins T, Howells DW, Donnan GA. Pooling of animal experimental data reveals influence of study design and publication bias. Stroke. 2004; 35(5):1203–8. https://doi.org/10.1161/01.STR.0000125719.25853.20 PMID: 15060322.

11. van der Worp HB, de Haan P, Morrema E, Kalkman CJ. Methodological quality of animal studies on neuroprotection in focal cerebral ischaemia. Journal of neurology. 2005; 252(9):1108–14. https://doi.org/10.1007/s00415-005-0802-3 PMID: 16170651.

12. Macleod MR, van der Worp HB, Sena ES, Howells DW, Dirnagl U, Donnan GA. Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. Stroke. 2008; 39(10):2824–9. https://doi.org/10.1161/STROKEAHA.108.515957 PMID: 18635842.

13. Macleod MR, O'Collins T, Horky LL, Howells DW, Donnan GA. Systematic review and metaanalysis of the efficacy of FK506 in experimental stroke. Journal of Cerebral Blood Flow and Metabolism. 2005; 25(6):713–21. https://doi.org/10.1038/sj.jcbfm.9600064 PMID: 15703698

14. Macleod MR, Fisher M, O'Collins V, Sena ES, Dirnagl U, Bath PM, et al. Good laboratory practice: preventing introduction of bias at the bench. Stroke. 2009; 40(3):e50–2. https://doi.org/10.1161/STROKEAHA.108.525386 PMID: 18703798.

15. McCann SK, Cramond F, Macleod MR, Sena ES. Systematic Review and Meta-Analysis of the Efficacy of Interleukin-1 Receptor Antagonist in Animal Models of Stroke: an Update. Translational Stroke Research. 2016; 7(5):395–406. https://doi.org/10.1007/s12975-016-0489-z PMID: 27526101

16. Liao J, Ananiadou S, Currie GL, Howard BE, Rice A, Sena ES, et al. Automation of citation screening in pre-clinical systematic reviews. bioRxiv. 2018:280131. https://doi.org/10.1101/280131

17. van der Worp HB, Howells DW, Sena ES, Porritt MJ, Rewell S, O'Collins V, et al. Can Animal Models of Disease Reliably Inform Human Studies? PLoS Med. 2010; 7(3):e1000245. https://doi.org/10.1371/journal.pmed.1000245 PMID: 20361020

18. Finnerup NB, Attal N, Haroutounian S, McNicol E, Baron R, Dworkin RH, et al. Pharmacotherapy for neuropathic pain in adults: a systematic review and meta-analysis. Lancet Neurol. 2015.

19. Karp NA, Mason J, Beaudet AL, Benjamini Y, Bower L, Braun RE, et al. Prevalence of sexual dimorphism in mammalian phenotypic traits. Nature Communications. 2017; 8:15475. https://doi.org/10.1038/ncomms15475 PMID: 28650954

20. Clayton JA, Collins FS. Policy: NIH to balance sex in cell and animal studies. Nature. 2014; 509:282–3. PMID: 24834516

21. Hershman DL, Weimer LH, Wang A, Kranwinkel G, Brafman L, Fuentes D, et al. Association between patient reported outcomes and quantitative sensory tests for measuring long-term neurotoxicity in breast cancer survivors treated with adjuvant paclitaxel chemotherapy. Breast cancer research and treatment. 2011; 125(3):767–74. Epub 2010/12/04. https://doi.org/10.1007/s10549-010-1278-0 PMID: 21128110.

22. Park SB, Lin CS, Krishnan AV, Goldstein D, Friedlander ML, Kiernan MC. Long-term neuropathy after oxaliplatin treatment: challenging the dictum of reversibility. The oncologist. 2011; 16(5):708–16. Epub 2011/04/12. https://doi.org/10.1634/theoncologist.2010-0248 PMID: 21478275.

23. Kidwell KM, Yothers G, Ganz PA, Land SR, Ko CY, Cecchini RS, et al. Long-term neurotoxicity effects of oxaliplatin added to fluorouracil and leucovorin as adjuvant therapy for colon cancer: results from National Surgical Adjuvant Breast and Bowel Project trials C-07 and LTS-01. Cancer. 2012; 118 (22):5614–22. Epub 2012/05/10. https://doi.org/10.1002/cncr.27593 PMID: 22569841.

24. Mogil JS, Crager SE. What should we be measuring in behavioral studies of chronic pain in animals? Pain. 2004; 112(1–2):12–5. Epub 2004/10/21. https://doi.org/10.1016/j.pain.2004.09.028 PMID: 15494180.

25. Rice ASC, Finnerup NB, Kemp HI, Currie GL, Baron R. Sensory profiling in animal models of neuropathic pain: a call for back-translation. Pain. 2017. https://doi.org/10.1097/j.pain.0000000000001138 PMID: 29300280.

26. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Controlled clinical trials. 1996; 17(1):1–12. PMID: 8721797.

27. Holman C, Piper SK, Grittner U, Diamantaras AA, Kimmelman J, Siegerink B, et al. Where Have All the Rodents Gone? The Effects of Attrition in Experimental Research on Cancer and Stroke. PLoS Biol. 2016; 14(1):e1002331. https://doi.org/10.1371/journal.pbio.1002331 PMID: 26726833

28. Crossley NA, Sena E, Goehler J, Horn J, van der Worp B, Bath PM, et al. Empirical evidence of bias in the design of experimental stroke studies: a metaepidemiologic approach. Stroke. 2008; 39(3):929–34. Epub 2008/02/02. https://doi.org/10.1161/STROKEAHA.107.498725 PMID: 18239164.

29. Rooke ED, Vesterinen HM, Sena ES, Egan KJ, Macleod MR. Dopamine agonists in animal models of Parkinson's disease: A systematic review and meta-analysis. Parkinsonism and Related Disorders. 2011; 17(5):313–20. https://doi.org/10.1016/j.parkreldis.2011.02.010 PMID: 21376651

30. Vesterinen HM, Sena ES, Ffrench-Constant C, Williams A, Chandran S, Macleod MR. Improving the translational hit of experimental treatments in multiple sclerosis. Multiple Sclerosis. 2010; 16(9):1044–55. https://doi.org/10.1177/1352458510379612 PMID: 20685763

31. Sena ES, van der Worp HB, Bath PMW, Howells DW, Macleod MR. Publication Bias in Reports of Animal Stroke Studies Leads to Major Overstatement of Efficacy. PLoS Biol. 2010; 8(3):e1000344. https://doi.org/10.1371/journal.pbio.1000344 PMID: 20361022

32. Percie du Sert N, Bamsey I, Bate ST, Berdoy M, Clark RA, Cuthill IC, et al. The Experimental Design Assistant. Nature methods. 2017; 14(11):1024–5. https://doi.org/10.1038/nmeth.4462 PMID: 28960183.

33. Morland RH, Novejarque A, Huang W, Wodarski R, Denk F, Dawes JD, et al. Short-term effect of acute and repeated urinary bladder inflammation on thigmotactic behaviour in the laboratory rat. F1000Research. 2015; 4:109. Epub 2015/01/01. https://doi.org/10.12688/f1000research.6255.1 PMID: 27158443.

34. Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? PLoS Med. 2010; 7(9):e1000326. https://doi.org/10.1371/journal.pmed.1000326 PMID: 20877712.

35. Elliott JH, Turner T, Clavisi O, Thomas J, Higgins JPT, Mavergames C, et al. Living Systematic Reviews: An Emerging Opportunity to Narrow the Evidence-Practice Gap. PLoS Med. 2014; 11(2).

36. Elliott JH, Synnot A, Turner T, Simmonds M, Akl EA, McDonald S, et al. Living systematic review: 1. Introduction-the why, what, when, and how. Journal of clinical epidemiology. 2017; 91:23–30. https://doi.org/10.1016/j.jclinepi.2017.08.010 PMID: 28912002.

37. Beller E, Clark J, Tsafnat G, Adams C, Diehl H, Lund H, et al. Making progress with the automation of systematic reviews: principles of the International Collaboration for the Automation of Systematic Reviews (ICASR). Systematic reviews. 2018; 7(1):77. https://doi.org/10.1186/s13643-018-0740-7 PMID: 29778096.

38. Akl EA, Kahale LA, Hakoum MB, Matar CF, Sperati F, Barba M, et al. Parenteral anticoagulation in ambulatory patients with cancer. The Cochrane database of systematic reviews. 2017; 9:Cd006652. Epub 2017/09/12. https://doi.org/10.1002/14651858.CD006652.pub5 PMID: 28892556.

39. Spurling GK, Del Mar CB, Dooley L, Foxlee R, Farley R. Delayed antibiotic prescriptions for respiratory infections. The Cochrane database of systematic reviews. 2017; 9:Cd004417. Epub 2017/09/08. https://doi.org/10.1002/14651858.CD004417.pub5 PMID: 28881007.

40. Cramond F, O'Mara-Eves A, Doran-Constant L, Rice A, Macleod M, Thomas J. The development and evaluation of an online application to assist in the extraction of data from graphs for use in systematic reviews [version 2; referees: 2 approved, 1 not approved]. Wellcome Open Research. 2019; 3(157). https://doi.org/10.12688/wellcomeopenres.14738.2 PMID: 30809592

41. Hori K, Ozaki N, Suzuki S, Sugiura Y. Upregulations of P2X(3) and ASIC3 involve in hyperalgesia induced by cisplatin administration in rats. Pain. 2010; 149(2):393–405. https://doi.org/10.1016/j.pain.2010.03.005 PMID: 20378247.

42. Frantzias J, Sena ES, Macleod MR, Salman RA-S. Treatment of Intracerebral Hemorrhage in Animal Models: Meta-Analysis. Annals of neurology. 2011; 69(2):389–99. https://doi.org/10.1002/ana.22243 PMID: 21387381

43. Currie GL, Delaney A, Bennett MI, Dickenson AH, Egan KJ, Vesterinen HM, et al. Animal models of bone cancer pain: Systematic review and meta-analyses. Pain. 2013; 154(6):917–26. https://doi.org/10.1016/j.pain.2013.02.033 PMID: 23582155

44. Sena E, Wheble P, Sandercock P, Macleod M. Systematic review and meta-analysis of the efficacy of tirilazad in experimental stroke. Stroke. 2007; 38(2):388–94. https://doi.org/10.1161/01.STR.0000254462.75851.22 PMID: 17204689

45. Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. Improving Bioscience Research Reporting: The ARRIVE Guidelines for Reporting Animal Research. PLoS Biol. 2010; 8(6).

46. Andrews NA, Latremoliere A, Basbaum AI, Mogil JS, Porreca F, Rice AS, et al. Ensuring transparency and minimization of methodologic bias in preclinical pain research: Pprecise considerations. Pain. 2015. https://doi.org/10.1097/j.pain.0000000000000458 PMID: 26683237.

47. Announcement: Reducing our irreproducibility. Nature. 2013; 496. https://doi.org/10.1038/496398a

48. Hair K, Macleod MR, Sena ES. A randomised controlled trial of an Intervention to Improve Compliance with the ARRIVE guidelines (IICARus). bioRxiv. 2018:370874. https://doi.org/10.1101/370874

49. Banwell V, Sena ES, Macleod MR. Systematic Review and Stratified Meta-analysis of the Efficacy of Interleukin-1 Receptor Antagonist in Animal Models of Stroke. Journal of Stroke and Cerebrovascular Diseases. 2009; 18(4):269–76. https://doi.org/10.1016/j.jstrokecerebrovasdis.2008.11.009 PMID: 19560680

50. Hooijmans CR, de Vries RBM, Ritskes-Hoitinga M, Rovers MM, Leeflang MM, IntHout J, et al. Facilitating healthcare decisions by assessing the certainty in the evidence from preclinical animal studies. PLoS ONE. 2018; 13(1):e0187271. https://doi.org/10.1371/journal.pone.0187271 PMID: 29324741.

51. Wang Q, Liao J, Hair K, Bannach-Brown A, Bahor Z, Currie GL, et al. 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 [Preprint]. [posted 2018 Jan 30; cited 23 April 2019]. https://www.biorxiv.org/content/10.1101/256776v1.

52. de Vries RBM, Hooijmans CR, Tillema A, Leenaars M, Ritskes-Hoitinga M. A search filter for increasing the retrieval of animal studies in Embase. Laboratory Animals. 2011; 45(4):268–70. https://doi.org/10.1258/la.2011.011056 PMID: 21890653

53. Hooijmans CR, Tillema A, Leenaars M, Ritskes-Hoitinga M. Enhancing search efficiency by means of a search filter for finding all studies on animal experimentation in PubMed. Laboratory Animals. 2010; 44(3):170–5. https://doi.org/10.1258/la.2010.009117 PMID: 20551243

54. de Vries RB, Hooijmans CR, Tillema A, Leenaars M, Ritskes-Hoitinga M. Updated version of the Embase search filter for animal studies. Lab Anim. 2014; 48(1):88. Epub 2013/07/10. https://doi.org/10.1177/0023677213494374 PMID: 23836850.

55. Bannach-Brown A, Przybyła P, Thomas J, Rice ASC, Ananiadou S, Liao J, et al. The use of text-mining and machine learning algorithms in systematic reviews: reducing workload in preclinical biomedical sciences and reducing human screening error. bioRxiv. 2018.

56. O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Erratum to: Using text mining for study identification in systematic reviews: a systematic review of current approaches. Systematic reviews. 2015; 4(1):59. https://doi.org/10.1186/s13643-015-0031-5 PMID: 25927201

57. Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, Bradley EW, et al. A call for transparent reporting to optimize the predictive value of preclinical research. Nature. 2012; 490(7419):187–91. https://doi.org/10.1038/nature11556 PMID: 23060188.

58. Macleod MR, O'Collins T, Howells DW, Donnan GA. Pooling of animal experimental data reveals influence of study design and publication bias. Stroke. 2004; 35(5):1203–8. https://doi.org/10.1161/01.STR.0000125719.25853.20 PMID: 15060322

59. Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997; 315(7109):629–34. https://doi.org/10.1136/bmj.315.7109.629 PMID: 9310563

60. Duval S, Tweedie R. Trim and Fill: A Simple Funnel-Plot–Based Method of Testing and Adjusting for Publication Bias in Meta-Analysis. Biometrics. 2000; 56(2):455–63. https://doi.org/10.1111/j.0006-341X.2000.00455.x PMID: 10877304

61. Zwetsloot P-P, Van Der Naald M, Sena ES, Howells DW, IntHout J, De Groot JAH, et al. Standardized mean differences cause funnel plot distortion in publication bias assessments. eLife. 2017; 6:e24260. https://doi.org/10.7554/eLife.24260 PMID: 28884685