Outcome heterogeneity and bias in acute experimental spinal cord injury
A meta-analysis

Ralf Watzlawick, MD, Ana Antonic, PhD, Emily S. Sena, PhD, Marcel A. Kopp, MD, Julian Rind, MS, Ulrich Dirnagl, MD, Malcolm Macleod, MD, David W. Howells, PhD,* and Jan M. Schwab, MD, PhD*

Neurology® 2019;93:e40-e51. doi:10.1212/WNL.0000000000007718

Abstract

Objective
To determine whether and to what degree bias and underestimated variability undermine the predictive value of preclinical research for clinical translation.

Methods
We investigated experimental spinal cord injury (SCI) studies for outcome heterogeneity and the impact of bias. Data from 549 preclinical SCI studies including 9,535 animals were analyzed with meta-regression to assess the effect of various study characteristics and the quality of neurologic recovery.

Results
Overall, the included interventions reported a neurobehavioral outcome improvement of 26.3% (95% confidence interval 24.3–28.4). Response to treatment was dependent on experimental modeling paradigms (neurobehavioral score, site of injury, and animal species). Applying multiple outcome measures was consistently associated with smaller effect sizes compared with studies applying only 1 outcome measure. More than half of the studies (51.2%) did not report blinded assessment, constituting a likely source of evaluation bias, with an overstated effect size of 7.2%. Assessment of publication bias, which extrapolates to identify likely missing data, suggested that between 2% and 41% of experiments remain unpublished. Inclusion of these theoretical missing studies suggested an overestimation of efficacy, reducing the effect sizes by between 0.9% and 14.3%.

Conclusions
We provide empirical evidence of prevalent bias in the design and reporting of experimental SCI studies, resulting in overestimation of the effectiveness. Bias compromises the internal validity and jeopardizes the successful translation of SCI therapies from the bench to bedside.
Glossary

BBB = Basso, Beattie, Bresnahan; BMS = Basso Mouse Scale for Locomotion; CAMARADES = Collaborative Approach to Meta-Analysis and Review of Animal Data From Experimental Studies; CI = confidence interval; OEG = olfactory ensheathing glia; SCI = spinal cord injury.

Traumatic spinal cord injury (SCI) affects hundreds of thousands of patients worldwide every year, and it is estimated that up to 5 million patients are currently affected by the chronic consequences of SCI.1,2 Despite successful improvements in the prevention of traumatic spinal cord injuries, it has been predicted that the growing population will result in a stable or even an increased total number of patients with SCI.3 However, limited financial resources, heterogeneous lesions, and small numbers of clinical trials impair clinical translation. Finding new treatments for these individuals relies on animal models of SCI. These are believed to have good face validity and to be sufficiently like SCI in humans to make them highly relevant for translation.4 This holds true as long as known translational limitations are taken into account. These include differences in locomotion (quadrupedal locomotion in rodent models5), neural innervation relevant for locomotion or fine motor control,6,7 different injury paradigms (sustained cord compression,8 concomitant injury of nerve roots in human SCI9), inflammatory responses,9 and other aspects of the physiology and pathophysiology of experimental SCI in rodents. In addition to cross-species differences, methodologic shortcomings are also believed to contribute to translational failure (translational roadblock) of animal model–derived findings.10–12 To determine the extent to which avoidable methodologic limitations apply to experimental SCI research, we aimed to characterize the impact of study design, study quality, and bias. To do so, we have performed a meta-analysis of published data from SCI animal studies. These techniques allowed us to interrogate data from larger numbers of animals, reducing the effect of individual underpowered studies. Insufficient statistical power has been identified as a major independent shortcoming limiting the translational and predictive value of animal models.13,14

The medical need for developing new therapies emphasizes the necessity for preclinical animal studies to generate clinically valuable (translational) data. This requires the assessment of confounding variables, which might impair the signal-to-noise ratio and reduce the likelihood of detecting an effect.

Risk of experimental and publication bias limits the value of the published literature on drug testing because both lead to overestimation of effect size. Given that systematic reviews and meta-analyses have identified limited internal validity (e.g., due to bias) associated with inflated effect sizes in preclinical ischemic brain injury studies,12 we investigate whether and to what degree preventable bias also influences experimental SCI studies. The pragmatic Collaborative Approach to Meta-analysis and Review of Animal Data From Experimental Studies (CAMARADES) approach has value in determining the likelihood that a particular dataset is subject to these problems.

In this meta-analysis and meta-regression, we determine the impact of study characteristics, methodologic and reporting bias, and common meta-epidemiologic characteristics on neurobehavioral recovery after experimental SCI in a large cohort of 9,535 animals.

Methods

Literature search

This study includes data stored in the CAMARADES database. The CAMARADES database contains preclinical animal data for various diseases (e.g., stroke, SCI, myocardial infarction, neuropathic pain) hosted at the University of Edinburgh since 2004. Access to the CAMARADES database is internationally available on request. For this study, all available SCI animal data were selected from the database in September 2015, including 6 independently assessed literature analyses of different SCI interventions: decompression, hypothermia, pharmaceutical inhibitors of the Rho/ROCK pathway, cellular treatment (stem cells, olfactory ensheathing glia [OEG]), and physical exercise after the injury. All of the corresponding study protocols and a protocol of this meta-analysis were finalized in advance of any data collection and are accessible online (dcn.ed.ac.uk/camarades/research.html#protocols). Five of those 6 studies have already been published as systematic reviews and meta-analyses.15–20 Eligibility criteria differ moderately among the included studies, but each reported locomotor recovery in experimental SCI models. We excluded studies that did not report the number of animals in each group, the mean effect size, and its variance. The initial literature searches in 3 online databases (PubMed, EMBASE, and ISI Web of Science) were performed until October 2014 for studies published before then and were screened by 2 independent investigators. The searches within the databases were not restricted to specific years.

Data extraction

Study characteristics from each publication were extracted into the CAMARADES database as outlined in previous reviews.15–20 In brief, when a single publication reported >1 experiment, they were treated as independent experiments.
For experiments reporting the improvement of a functional outcome measured by a neurobehavioral score, only the final time points were extracted. Studies were excluded if they did not report the number of animals in each group, the mean effect size and its SD, or SEM. Where various outcome measurements were reported for the same group of animals, we pooled those containing a value for multiple scales. The quality of each study was assessed with a modified 9-point checklist21 (1) reporting of a sample size calculation, (2) control of the animals’ temperature, (3) use of anesthetics other than ketamine (because of its marked intrinsic neuroprotective activity), (4) randomized treatment allocation, (5) treatment allocation concealment, (6) blinded assessment of outcome, (7) publication in a peer-reviewed journal; (8) statement of compliance with regulatory requirements, and (9) statement of potential conflicts of interest.

**Analysis**

The normalized effect size (normalized mean difference) was calculated for each experimental comparison as the proportional improvement in the treated compared to the control group, along with a standard error of this estimate. When a single control group served multiple treatment groups, the size of the control group was adjusted to account for this. This corrected number was used for any further calculation and in the weighting of effect sizes.22

We stratified the analysis according to (1) the applied treatment, (2) the neurobehavioral score used (Basso, Beattie, Bresnahan [BBB] score23; Basso Mouse Scale for Locomotion [BMS]24; Tarlov scale25; others), (3) the model and level of the injury, (4) study characteristics (time to treatment, time of outcome assessment, animal sex and strain), (5) anesthetic regimen (anesthetic used and method of ventilation), and (6) methodologic quality of the studies according to the 9-point checklist. Cohorts within a stratification accounting for >1% of the total animal number were considered an individual group. Smaller cohorts accounting for <1% were grouped together and referred to as others.

We applied random-effect meta-analysis regression to calculate an overall treatment effect, combining efficacy across studies to assess significant differences between the groups for each stratification criteria. Meta-regression with the metareg function of STATA/SE13 (StataCorp, College Station, TX) was used26 and the significance level was set at \( p = 0.05 \). The meta-regression was performed in a univariate manner because major concerns about the reliability of multivariate analysis have been raised with the need for further statistical investigation.27 Adjusted \( R^2 \) values were calculated to assess how much residual heterogeneity was explained by each independent variable within a stratified dataset.22 The adjusted \( R^2 \) is defined as the ratio of explained variance of the respective stratification to total variance (of all experiments together). Higher \( R^2 \) values are indicative of a higher explanatory value of this stratification. Restricted maximum likelihood was used to estimate the additive (between-study) component of variance tau2 within the meta-regression.26 The dependent variable was, in all cases, the normalized effect size.

Funnel plots28 and trim and fill analyses29,30 were performed with the metatrim function of STATA/SE13. The Duval and Tweedie29 nonparametric trim and fill method estimates the number of missing studies (publication bias) that might have been performed but are not available for a meta-analysis using simple rank-based data augmentation techniques adjusted for missing studies. Normalized mean difference (effect size) was plotted against precision (1/SEM). The metabias function detects the presence of publication bias by Egger regression measured by the intercept from standard normal deviates against precision.31 Figures were drawn with SigmaPlot (Systat Software Inc, San Jose, CA).

**Data availability**

Data not provided in the article will be shared at the request of other investigators for purposes of replicating procedures and results.

**Results**

**Selection of datasets and study details**

The CAMARADES database contained details on 549 experiments from 303 included publications from 1967 to 2014 reporting functional improvement assessed by neurobehavioral scores (table 1).

We included data from 9,535 animals. Overall, the number of control animals used was 20% smaller than the number that received treatment (control animals \( n = 3,765, 39.5\% \); treated animals \( n = 5,770, 60.5\% \)). Six different animal species were used to investigate therapeutic interventions for SCI: 8,108 rats (85%), 1,199 mice (12.6%), 107 dogs (1.1%), 57 monkeys (0.6%), 44 rabbits (0.5%), and 20 sheep (0.2%).

**Study characteristics and their impact on neurobehavioral outcome**

Random-effect meta-regression revealed an overall effect size of 26.3% (95% confidence interval [CI] 24.3–28.3, \( I^2 = 98.6\% \)) when all included neurobehavioral scores were combined. To explore sources of heterogeneity, we applied meta-regression and identified several study design characteristics accounting for significant within- and between-study heterogeneity (figures 1 and 2). Study characteristics are represented in terms of prevalence (width of the bar indicating the number of studies) and variability (CI).

**Animal model**

Experimental SCI in rats constituted the most prevalent injury paradigm and outnumbered mouse SCI, even though the latter offers the advantage of genetic manipulation (figure 1A). Variability as indicated by 95% CIs was larger in mice compared to rats. Effect sizes for larger animal SCI models, including dogs, monkeys, rabbits, and sheep, applied mostly customized scoring systems and were generally larger and linked to greater variability (\( R^2 = 8.0\%, p < 0.0001 \)).
Table 1 Characteristics of different interventions included in the overall analysis

<table>
<thead>
<tr>
<th>Intervention</th>
<th>Decompression</th>
<th>Exercise</th>
<th>Hypothermia</th>
<th>OEG</th>
<th>Rho/ROCK inhibitors</th>
<th>Stem cells</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Publications, n (%)</td>
<td>21 (6.9)</td>
<td>25 (8.3)</td>
<td>16 (5.3)</td>
<td>51 (16.8)</td>
<td>14 (4.6)</td>
<td>176 (58.1)</td>
<td>303</td>
</tr>
<tr>
<td>Experiments, n (%)</td>
<td>79 (14.4)</td>
<td>36 (6.6)</td>
<td>25 (4.6)</td>
<td>62 (11.3)</td>
<td>30 (5.5)</td>
<td>317 (57.7)</td>
<td>549</td>
</tr>
<tr>
<td>Animals, n (%)</td>
<td>873 (9.2)</td>
<td>575 (6.0)</td>
<td>448 (4.7)</td>
<td>1,164 (12.2)</td>
<td>655 (6.7)</td>
<td>5,820 (61.0)</td>
<td>9,535</td>
</tr>
<tr>
<td>BBB scale, n (%)</td>
<td>246 (3.4)</td>
<td>498 (6.8)</td>
<td>218 (3.0)</td>
<td>1,027 (14.0)</td>
<td>524 (7.2)</td>
<td>4,811 (65.7)</td>
<td>7,324</td>
</tr>
<tr>
<td>BMS score, n (%)</td>
<td>0</td>
<td>55 (12.6)</td>
<td>0</td>
<td>0</td>
<td>131 (30.0)</td>
<td>250 (57.3)</td>
<td>436</td>
</tr>
<tr>
<td>Multiple, n (%)</td>
<td>282 (15.1)</td>
<td>398 (21.4)</td>
<td>101 (5.4)</td>
<td>0</td>
<td>0</td>
<td>1,083 (58.1)</td>
<td>1,864</td>
</tr>
<tr>
<td>Inclined plane, n (%)</td>
<td>295 (90.2)</td>
<td>0</td>
<td>32 (9.8)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>327</td>
</tr>
<tr>
<td>Tarlov scale, n (%)</td>
<td>27 (8.6)</td>
<td>0</td>
<td>88 (27.9)</td>
<td>0</td>
<td>0</td>
<td>200 (63.5)</td>
<td>315</td>
</tr>
<tr>
<td>Others, n (%)</td>
<td>151 (23.1)</td>
<td>14 (2.1)</td>
<td>99 (15.1)</td>
<td>137 (20.9)</td>
<td>0</td>
<td>253 (38.7)</td>
<td>654</td>
</tr>
<tr>
<td>Overall effect size, (%)</td>
<td>34.2 (27.4–41.1)</td>
<td>13.1 (4.9–21.4)</td>
<td>29.1 (17.1–41.1)</td>
<td>23.6 (17.3–30.0)</td>
<td>23.2 (14.1–32.4)</td>
<td>27.1 (24.5–29.7)</td>
<td>26.3 (24.3–28.4)</td>
</tr>
</tbody>
</table>

Abbreviations: BBB = Basso, Beattie, and Bresnahan; BMS = Basso Mouse Scale; OEG = olfactory ensheathing glia. Cumulative number of publications, number of experiments, included animals, applied outcome measures, and effect size calculated by meta-regression for all outcome measurements are given. Largest effect sizes relate to spinal decompression surgery and hypothermia. Note that multiple outcome measurements contain experiments assessing various outcome measures for a specific animal group (e.g., BBB score and Tarlov scale).

Neurobehavioral scale

Multiple neurobehavioral outcome scales were used to assess functional recovery after SCI (figure 1B). The BBB or BMS23,24 score for locomotion was most commonly applied, accounting for 66.7% (6,302 of 9,462 animals). The 21-point BBB score evaluates hind-limb movement and fore/hind-limb coordination in an open field test in rats, whereas the BMS score is adapted to recovery patterns in mice.23,24 BBB and BMS scores reported similar improvement of locomotion accounting for 66.7% (6,302 of 9,462 animals). The 21-point BBB score evaluates hind-limb movement and fore/hind-limb coordination in an open field test in rats, whereas the BMS score is adapted to recovery patterns in mice.23,24 BBB and BMS scores reported similar improvement of locomotion accounting for 66.7% (6,302 of 9,462 animals).

Blinded assessment of outcome

Regarding the hypothesis that lack of reporting on study quality criteria is associated with variations in effect sizes, 2 main quality characteristics revealed statistically significant results: blinding of experiments and whether compliance with animal welfare was reported (figure 1, C and D). Fewer than half (48.8%) of the interventional studies reported that they were conducted in a blinded manner. Lack of blinding results in evaluation bias that reduces the internal reliability of the experiment and accounts for significantly inflated effect sizes by 7.2% ($R^2 = 2.6, p = 0.001$) (figure 1C).

Compliance with animal welfare

Reported compliance with animal welfare protocols was lacking in 29.5% (n = 2,815) of the studies. Studies lacking reported compliance were associated with 5.6% larger effect sizes and larger variability (figure 1D). The reporting of statement of compliance ($R^2 = 2.0, p = 0.015$) and blinding during outcome assessment were significantly associated with different effect sizes, both representing items of the 9-point quality checklist.21,22

Impact of the design of experimental SCI studies on neurologic recovery in rodent models

To address the possibility that differences in effect sizes can be attributed to the distinct scoring systems applied to rodent and large animal models, we performed a subanalysis for a BBB/BMS score only population (figure 2). Similar BBB (rat) and BMS (mouse) effect sizes (figure 1B) ensure sufficient comparability as the basis to study the relevance of distinct experimental designs for neurologic recovery. For the BBB/BMS only population, meta-regression calculated an effect size of 24.0% (95% CI 21.9–26.2, $I^2 = 98.9%$; table 2).

Type of treatment

We compared quantitative data about SCI interventions relevant for clinical SCI ranging from acute care to chronic SCI. These include decompression, hypothermia, pharmaceutical (Rho/ROCK pathway) and cellular treatments (stem cells, OEGs), and rehabilitative interventions (voluntary and forced exercise). Acute care interventions are associated throughout with higher effect sizes compared with rehabilitative interventions such as exercise (figure 2A). Improvement in motor
outcome after SCI was most pronounced for decompression surgery (27.1%), followed by stem cells (26.2%), Rho/ROCK inhibition (21.3%), and hypothermia (16.3%). Exercise demonstrated the lowest effect sizes (14.4%). Stem cell transplantation studies accounted for 65.3% of the total number of animals ($R^2 = 3.0\%, p = 0.0022$) included in this subanalysis.

**Level of injury**
The most prevalent lesion level location was Th8, followed by Th9, and Th10 (figure 2B). Level of injury affects the effect size, with closer proximity of the lesion to deafferented afferent target (in the lumbosacral motor pools) resulting in larger effect sizes, mimicking an exponential curve function ($R^2 = 4.5\%, p = 0.0047$). Interpretation of the high cervical SCI data is limited further by small sample size, large variance, and the use of less severe lesion paradigms. Thus, experimental SCI is dominated by low thoracic SCI paradigms.

**Sex**
Female rodents outnumbered males (58.2% females, 27.4% males; figure 2C). Studies including both sexes were characterized by a larger variability resulting in an extended CI ($R^2 = 3.5\%, p = 0.0018$).

**Type of injury**
SCI was induced with 5 different types of injury (figure 2D). Contusion and transection were the most severe injuries and produced the smallest interventional effect. Static compression was associated with slightly better effect sizes, followed by hemisection. Photochemical injuries (focal topical application of high-energy light or chemicals) were rarely applied and induced only minor injury, associated with the largest improvement responses in motor score ($R^2 = 3.8\%, p = 0.005$). Contusion SCI was by far the most prevalent injury paradigm ($n = 3,983$ animals).
Differential aspects of spinal cord injury (SCI) modeling on neurologic recovery were assessed by including only studies that applied Basso, Beattie, and Bresnahan (BBB) or Basso Mouse Scale (BMS) scores. This homogeneous cohort (red square, B) comprises 6,048 of 9,535 animals (63.4%). (A) Six different interventions were included in the meta-regression. (B) Lesion-level matter. A more rostral lesion level (high thoracic) along the neuraxis resulted in smaller outcome gain compared to more distal thoracic lesions. High cervical (C2-C4) lesion levels were considered a separate cohort in this regard in that only unilateral hemisections can be performed at those levels because complete transections or severe contusions are technically challenging or will otherwise cause immediate death of the animal. (C) Female rodents constitute the main cohort tested in experimental SCI. (D) Severe contusion and complete transection injuries resulted in the most impaired neurologic function and lowest effect size, followed by static clip compression. Hemisection and photochemical injuries (focal topical application of high-energy light and/or chemicals) induced only minor injury, associated with the largest improvement. (E) Geographic location of the experimental laboratory influenced the amount of reported neurologic recovery. Individual countries were identified if they contributed >200 animals to the dataset. (F) Effect size of the investigated intervention decreased with extended assessment period after SCI. Effect size depended on the duration of the assessment and started to intermittently drop below the average with observational time frames >8 weeks, whereas the variability remained similar. Gray bars represent the 95% confidence interval (CI) of the global estimate (neurologic recovery). Vertical error bars show the 95% CI of individual means. Width of each bar indicates the percentage of reported animal numbers per specified item. Asterisks indicate stratification items for which the bar width was broadened later for better visibility. All stratifications account for a significant proportion of between-study heterogeneity (p < 0.001). OEG = olfactory ensheathing glia.
Table 2 Analysis of the BBB/BMS only cohort after correction for missing data (delta inflation)

<table>
<thead>
<tr>
<th>Intervention</th>
<th>Reported effect size (random metatrim), % (95% CI)</th>
<th>Adjusted effect size (random), % (95% CI)</th>
<th>Filled experiments (was), n</th>
<th>Absolute overestimation of efficacy, %</th>
<th>Detectable bias with Egger test?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decompression</td>
<td>27.1 (15.1–39.1)</td>
<td>12.8 (1.7 to 24.0)</td>
<td>9 (22)</td>
<td>14.3 (~2.1 to 30.7)</td>
<td>+++ (p &lt; 0.001)</td>
</tr>
<tr>
<td>Exercise</td>
<td>14.4 (6.4–22.3)</td>
<td>1.9 (~6.9 to 10.7)</td>
<td>7 (35)</td>
<td>12.5 (0.7 to 24.3)</td>
<td>(p = 0.72)</td>
</tr>
<tr>
<td>Hypothermia</td>
<td>16.3 (10.3–22.3)</td>
<td>16.3</td>
<td>— (11)</td>
<td>—</td>
<td>(p = 0.799)</td>
</tr>
<tr>
<td>OEG</td>
<td>20.1 (15.2–24.9)</td>
<td>19.2 (14.4 to 24.0)</td>
<td>1 (53)</td>
<td>0.9 (~5.9 to 7.7)</td>
<td>+++ (p &lt; 0.001)</td>
</tr>
<tr>
<td>Rho/ROCK inhibitors</td>
<td>21.3 (16.0–26.6)</td>
<td>15.6 (10.1 to 21.2)</td>
<td>9 (30)</td>
<td>5.7 (~2.0 to 13.4)</td>
<td>+++ (p &lt; 0.001)</td>
</tr>
<tr>
<td>Stem cells</td>
<td>26.2 (23.9–28.5)</td>
<td>26.2</td>
<td>— (278)</td>
<td>—</td>
<td>(p = 0.12)</td>
</tr>
</tbody>
</table>

Abbreviations: BBB = Basso, Beattie, and Bresnahan; BMS = Basso Mouse Scale; CI = confidence interval; OEG = olfactory ensheathing glia.

Location of the laboratory
The regional location of the experimental laboratory affected the reported effect size (figure 2E). Most of the animal experiments were performed in China, the United States, and Japan ($R^2 = 8.3\%$, $p < 0.0001$). Studies from China and Australia were associated with large effect sizes, while studies from Brazil, Sweden, the United States, and Canada were associated with the smallest effect sizes. The required number of total animals for each listed country was 200, excluding those otherwise assigned to others.

Time point of assessment
Studies including long-term outcomes (>2-month assessment time frame) were underrepresented. The effect size of the investigated intervention decreased with extended assessment period after SCI, starting to intermittently drop below the average with observational time frames >8 weeks. Declining effect sizes after 8 weeks point to the risk that interventions provided only transient benefits, diminishing with time and therefore unlikely to translate into long-term results. This is in line with interventions that are able to trigger faster recovery but fail to reach higher functional plateaus at longer follow-up. The pattern of diminished effect size was not dependent on early vs late intervention. This rules out the possibility that decreasing effect sizes after 8 weeks resulted from more studies testing delayed applications. Variability was the highest in cases when the observational time was not indicated ($R^2 = 4.6\%$, $p = 0.0017$; figure 2F).

Detection of publication bias (missing data)
When considering all studies of the homogeneous BBB/BMS only cohort (table 2) comprising data from 6,302 animals stratified for the type of intervention, we observed an overestimation of efficacy due to missing data. Funnel plotting and trim and fill analysis indicated that data were missing in 4 types of SCI interventions. The reported effect size in each study was adjusted for absolute overestimation of efficacy for each study ranging between 0.9% (OEG) and 14.3% (decompression) (figure 3 and table 2). Assessment of the missing data/publication bias with the trim and fill method suggested that the number of missing experiments varies among the tested interventions, ranging from 0% (stem cells, hypothermia) to 1.9% (OEG), 20% (exercise), 30% (Rho/ROCK inhibition), and 40.9% (decompression). Data asymmetry indicative of missing data was confirmed by Egger regression, suggesting significant bias in the BBB/BMS only datasets.

Geographic differences
To explore the differences attributable to the region-specific effect sizes (table 3), we focused on a comparison of the 3 countries with high study numbers and comparably small CIs (Unites States, Japan, and China). We excluded data from other countries with considerably lower study numbers (indicated by smaller bar width, figure 2E) because the analysis would have been underpowered. Between-country differences may be related to varying percentages of interventions associated with larger effect sizes (percentage of stem cell studies: China 31.3%, Japan 25.5%, and United States 17.3%). However, when we focused on stem cell interventions only, a similar geographic pattern was observed, with the highest effect size (31.3%) observed in studies from Chinese laboratories compared to studies from Japan (25.5%) or the United States (17.3%). When we compared the effect sizes of recovery after complete transection as a defined injury type, the response to treatment was highest in studies from China (28.2%), followed by Japan (9.3%) and the United States (8.5%). An association of poor study quality (characterized by the quality score) with larger therapeutic effect sizes has been reported earlier in other CNS lesion paradigms and is...
To allow a more homogeneous group characterized by similar effect sizes (ESs), a rodent analysis was completed identifying the Basso, Beattie, and Bresnahan (BBB)/Basso Mouse Scale (BMS) only group as comparable with regard to the main outcome parameter (improved locomotor function, figure 1B, red square). Funnel plots illustrating the precision (1 divided by the SEM, y-axis) plotted against the standardized ES (x-axis). Each black dot represents 1 experiment. In the absence of publication bias, the points should resemble an inverted funnel. Black dotted line indicates the reported ES before inclusion of missing studies; red line, after trim and fill analysis. (A) Decompression: reported ES 27.1% (95% confidence interval [CI] 15.1–39.1, n = 22 studies), adjusted ES 12.8% (95% CI 1.7–24.0, n = 31). (B) Exercise: reported ES 14.4% (95% CI 6.4–22.3, n = 35), adjusted ES 1.9% (95% CI −6.9 to 10.7, n = 42). (C) Hypothermia: reported ES 16.3% (95% CI 10.3–22.3) without adjustments. (D) Olfactory ensheathing glia (OEG): reported ES 20.1% (95% CI 15.2–24.9, n = 53), adjusted ES 19.2% (95% CI 14.4–24.0, n = 54). (E) Rho/ROCK inhibitors: reported ES 21.3% (95% CI 16.0–26.6, n = 30), adjusted ES 15.6% (95% CI 10.1–21.2, n = 39). (F) Stem cells: reported ES 26.2% (95% CI 23.9–28.5) without detection of missing studies.
Table 3 Study characteristics of the BBB/BMS subanalysis based on geography

<table>
<thead>
<tr>
<th>Type of SCI</th>
<th>China (n = 2,177)</th>
<th>Japan (n = 960)</th>
<th>United States (n = 2,039)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ES 95% CI No. p Value</td>
<td>ES 95% CI No. p Value</td>
<td>ES 95% CI No. p Value</td>
</tr>
<tr>
<td>Quality score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>22.5 4.4 to 40.6 89 0.53</td>
<td>49.1 33.8 to 64.4 132 0.002</td>
<td>38.7 14.3 to 63.1 112 0.06</td>
</tr>
<tr>
<td>2</td>
<td>29.4 18.7 to 40.1 395</td>
<td>32.0 20.1 to 43.9 250</td>
<td>18.3 2.6 to 34. 362</td>
</tr>
<tr>
<td>3</td>
<td>37.2 30.8 to 43.7 613</td>
<td>16.1 8.9 to 23.4 183</td>
<td>17.9 6.4 to 29.3 425</td>
</tr>
<tr>
<td>4</td>
<td>30.4 21 to 39.8 613</td>
<td>23.0 –0.2 to 46.1 78</td>
<td>11.8 –4 to 27.7 369</td>
</tr>
<tr>
<td>5</td>
<td>29.6 19.7 to 39.5 395</td>
<td>18.6 7.3 to 29.9 247</td>
<td>4.0 –12.2 to 20.2 437</td>
</tr>
<tr>
<td>6</td>
<td>27.4 2.6 to 52.2 52</td>
<td>18.6 0.7 to 36.5 70</td>
<td>22.3 5.9 to 38.8 334</td>
</tr>
<tr>
<td>7</td>
<td>22.9 –16.9 to 62.8 20</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Type of SCI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Compression</td>
<td>35.0 22.8 to 47.1 336 0.35</td>
<td>21.5 7.1 to 35.8 117 0.37</td>
<td>10.5 –33.5 to 54.5 45 0.26</td>
</tr>
<tr>
<td>Hemisection</td>
<td>33.9 22.8 to 45. 367</td>
<td>19.6 –1.6 to 40.7 109</td>
<td>30.6 13.6 to 47.6 330</td>
</tr>
<tr>
<td>Transection</td>
<td>28.2 19.8 to 36.5 813</td>
<td>9.3 –12.3 to 30.9 45</td>
<td>8.5 –10.4 to 27.3 172</td>
</tr>
<tr>
<td>Contusion</td>
<td>34.7 28 to 41.3 661</td>
<td>26.0 20.1 to 31.9 689</td>
<td>15.2 9.6 to 20.7 1,492</td>
</tr>
</tbody>
</table>

Abbreviations: BBB = Basso, Beattie, and Bresnahan; BMS = Basso Mouse Scale; CI = confidence interval; ES = effect size; SCI = spinal cord injury.

For a more detailed analysis on varying effect sizes dependent on the location of the publishing laboratory, we focused on countries with a robust number of experiments. Representative countries include Japan (n = 960 rodents), the United States (n = 2,039), and China (n = 2,177). This analysis corresponds to the overall findings displayed in figure 2E, suggesting regional effects with regard to the reporting of study results.

Observables in the SCI studies from Japan (p = 0.002) and the United States (p = 0.06). In contrast, in studies from China, the effect size did not correlate with the quality score.

Discussion

We demonstrate the feasibility and utility of a meta-epidemiologic approach for systematic assessment of outcome variability and detection of the presence of bias in the design and reporting of experimental SCI studies through a meta-analysis of data from 303 publications including 9,535 experimental animals. We provide empirical evidence for bias in experimental SCI research resulting in inflated effect sizes, and we report the effect of study design characteristics on heterogeneity of neurobehavioral recovery.

Rodent females constitute the largest population used in preclinical SCI models, with rats outnumbering mice (figures 1 and 2). The analysis of outcome measures confirms open-field BBB and BMS testing as robust measures of locomotor recovery characterized by comparable effect sizes. The data re-emphasize the evidence that a combination of multiple functional outcome measures examining particular components of recovered motor function results in the reporting of less inflated effect sizes.32 Nonhuman primates were associated with a large effect size and variability. The Tarlov score was developed in 195425 and is outdated.

Independent group assignment (blinded SCI induction, allocation concealment) and blinded assessment (masking) of outcome are crucial to avoid prevalent selection and evaluation bias; more than half of the studies analyzed lacked blinding. Lack of blinding was associated with a significant overstatement of efficacy by 7.2% in the overall cohort when considering all species and interventions. However, relative to the effect size of tested interventions, up to 55% of interventional effect may be inflated and attributable to the lack of blinding (e.g., 7.2% of 13.1% for exercise as an intervention). In addition to blinding, compliance with regulatory requirements is considered a measure of study quality.21 Lack of reporting of this criterion was associated with an overstatement of efficacy by 5.6%.

To provide a more homogeneous subcohort for studying differential effects of intervention type, lesion level, sex, injury modality, geographic location of publishing laboratory, effect durability, and missing data (publication/reporting bias), the subsequent analysis was restricted to a rodent BBB/BMS only cohort (n = 7,613).

Interventions

We identified 6 different therapeutic interventions, including the spectrum of cellular therapies (stem cells, OEG), systemically applied approaches (hypothermia, Rho/ROCK inhibitors), and rehabilitative (exercise) and surgical interventions (decompression). Four of 6 interventions (67%) were characterized by data asymmetry, attributable largely to...
missing data. For 3 of 6 interventions (50%), our analysis identified an overstatement of efficacy (table 2). The effect sizes of classic acute care interventions (decompressive surgery and hypothermia) were larger than the effect sizes of rehabilitation interventions. Effects of exercise interventions display a large CI and point to the existence of variable factors determining rehabilitation efficacy. These results also suggest that rehabilitation cannot fully compensate for what has been lost during acute care and could emphasize the therapeutic value of optimal specialized acute SCI care. Retrospective clinical studies of acute care interventions such as decompressive surgery support this notion.34,35

SCI lesion level
We provide evidence that in addition to varying lesion severity, the distance from the lesion to the lumbosacral motor pools is an important determinant of responsiveness to treatment (figure 2B). The more distant the lesion is, the lower the neurologic recovery is. The increase of effect size per spinal level, descending along the neuraxis, is not linear. The effect size per lesion level distal to Th10 results in a remarkable increased step of effect sizes. This stratification confirmed the biological plausibility of the SCI model, with increasing effect size the closer the lesion is located to lumbosacral motor pools, suggesting the propagation of intrinsic repair/regeneration pathways as the underlying mechanism of recovery. Less severe high cervical models will naturally allow higher compensatory plasticity, contributing to larger effect sizes, which render them mechanistically noncomparable to complete transection or contusion SCI applied to low cervical to low thoracic sites.

Sex
Female rodents displayed smaller effect sizes compared to males (figure 2C). Thoracic SCI is the most prevalent cohort modeling experimental SCI (figure 2B), matching to some degree with lower functional recovery in women after thoracic (T2-8) SCI. However, this is only a transient effect and is no longer detectable 1 year after human SCI.36

Injury type
Contusion injury was the most severe injury, characterized by smallest effect sizes in response to interventional therapy. This is consistent with evidence of injury exacerbation extending both rostrally and caudally from the lesion site after contusion, whereas even complete transection, injury remains localized to the initial lesion site.32,37 Static clip compression injury is a slightly less severe SCI paradigm (figure 2D).

Geographic effects
Stratification for the location of publishing laboratory revealed differences (figure 2E). Study results from Japan and the United States showed an inverse association between effect sizes and quality, confirming findings from a previous meta-analysis in animal research,21 whereas studies from China did not (table 3). Chinese studies demonstrated smaller variability of outcomes compared to studies from Japan and the United States. Larger effect sizes concomitant with smaller CIs are consistent with the suggestion that animals considered to be outliers have been removed from further analysis, contributing to attrition bias.38 The data suggest geographic differences with regard to the reporting of study results and emphasize the importance of reporting attrition and its reasons.

Unstable (transient) treatment effects
We observed an intermittent decline of effect sizes from 8 weeks after SCI on, suggesting the presence of concomitant processes that undermine gained locomotor functions. Several processes may account for this, including delayed neurodegeneration39 and salutatory nerve conduction failure of intact axons at the chronic lesion site.40,41 Our results suggest that an observational time frame of ≥8 weeks is required to report durable effects. The data indicate a risk for transient interventional effects and the need to favor long-term over short-term outcome measures, in line with other models of acute CNS injury.42

Reporting bias (missing data)
We provide empirical evidence of missing data. Trim and fill analysis of the entire BBB/BMS only cohort identified missing data. This was confirmed by Egger regression (table 2). Publication bias is a fundamental problem in preclinical research across many disciplines.30 The presence of such bias leads to inflated effect sizes and skewed biometric planning of clinical trials and jeopardizes the successful translation of SCI therapies from bench to bedside. The fact that for some of the interventions no publication bias could be detected may have several reasons and does not automatically imply that it is not present at all. The analysis of interventions with a small number of experiments is most likely underpowered. In addition, even for interventions with large sample sizes, the precision of experiments may be falsely increased due to unexplored attrition.

Regarding limitations of this study, we were able to include only data from studies in the public domain. Second, for both study quality and study design characteristics, we relied on published information. When relevant data were not available (the sex of a cohort of animals or measures taken to reduce bias), we have either analyzed them as not known or inferred that measures that were not reported did not occur. Third, we present a series of univariate meta-regression analyses. Multivariate meta-regression or stepwise partitioning of heterogeneity might provide more robust insights, but these techniques are not well established today. Fourth, our analysis depended on the validity of the scales used, which has been questioned.43 Fifth, translation into human SCI might be dampened by limitations of the animal SCI model, affecting the external validity, such as the absence of polytrauma (including vertebrae fractures). Finally, we accept the weakness of the presented meta-analysis, which combines various animal models of specific SCI treatments within a very large
dataset and assumes that random-effect meta-regression can account for this. Detecting the likelihood of preventable bias is still a comparatively novel but imperfect science because reporting standards are generally poor. However, while imperfect, the analysis represents best evidence available, representing one of the largest studies on SCI animal models to date.

In this study, we provide empirical evidence for differential outcome characteristics attributable to distinct SCI models and experimental designs. We observed evidence for transient treatment effects, as well as evaluation and publication bias, which diminish the experimental validity and could cause clinical trials to follow false leads. The presence of publication bias in preclinical SCI research confirms the endemic presence of underlying bias in science, rather than being specific to SCI. Bias in animal SCI experiments erodes the predictive value for translation into human SCI trials. Bias is not restricted to experimental testing but also occurs later in the translational spectrum in early clinical testing.

The SCI field is proactively responding to this challenge. Our data inform 4 recommendations to enhance the predictive value of experimental SCI research for a higher likelihood of translational success:

1. Reduce bias at the benchside through allocation concealment and blinded assessment of outcome; by reporting animals excluded from the analysis; and by providing a sample size calculation, inclusion and exclusion criteria, randomization, and transparent reporting of potential conflicts of interest and study funding.

2. Report nonconfirmatory (negative or neutral) results, which are persistently underrepresented in the public domain.

3. Use international experimental and reporting standards in preclinical SCI studies according to the Minimum Information About a Spinal Cord Injury Experiment or Animal Research: Reporting of In Vivo Experiments guidelines to reduce study design variability.

4. Develop data-sharing models and international cooperation in preclinical SCI research to learn more about underlying differences.

**Author contributions**

Conception and design of the study: R. Watzlawick, E. Sena, M. Macleod, D.W. Howells, J.M. Schwab; acquisition and analysis of data: R. Watzlawick, A. Antonic, J. Rind; drafting a significant portion of the manuscript or figures: R. Watzlawick, A. Antonic, M.A. Kopp, U. Dinnagl, D.W. Howells, J.M. Schwab.

**Acknowledgment**

The authors thank Gillian Currie for expert technical assistance. J.M.S. had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. The Department of Clinical and Experimental Spinal Cord Injury Research is an associated member of the European Multicenter Study About Spinal Cord Injury.

**Study funding**

No targeted funding reported.

**Disclosure**

R. Watzlawick, A. Antonic, E. Sena, M. Kopp, J. Rind, U. Dinnagl, M. Macleod, and D. Howells report no disclosures relevant to the manuscript. J. Schwab received funding support from the Else-Kroehner-Fresenius Foundation (No. 2012_A32); the Wings-for-Life Spinal Cord Research Foundation (Nos. DE-047/14, DE-16/16); the Era-Neuron Program of the European Union (SILENCE No. 01EW170A, SCI-Net No. 01EW1710); the National Institute on Disability, Independent Living, and Rehabilitation Research (No. 90SI5020); and the W.E. Hunt & C.M. Curtis Endowment. J.M.S. is a Discovery Theme Initiative Scholar (Chronic Brain Injury) of The Ohio State University. Go to Neurology.org/N for full disclosures.

**Publication history**

Received by Neurology August 27, 2018. Accepted in final form February 11, 2019.

**References**


Outcome heterogeneity and bias in acute experimental spinal cord injury: A meta-analysis
Ralf Watzlawick, Ana Antonic, Emily S. Sena, et al.
Neurology 2019;93:e40-e51 Published Online before print June 7, 2019
DOI 10.1212/WNL.0000000000007718

This information is current as of June 7, 2019

Updated Information & Services
including high resolution figures, can be found at:
http://n.neurology.org/content/93/1/e40.full

References
This article cites 50 articles, 8 of which you can access for free at:
http://n.neurology.org/content/93/1/e40.full#ref-list-1

Subspecialty Collections
This article, along with others on similar topics, appears in the following collection(s):
Plasticity
http://n.neurology.org/cgi/collection/plasticity
Spinal cord trauma
http://n.neurology.org/cgi/collection/spinal_cord_trauma

Permissions & Licensing
Information about reproducing this article in parts (figures, tables) or in its entirety can be found online at:
http://www.neurology.org/about/about_the_journal#permissions

Reprints
Information about ordering reprints can be found online:
http://n.neurology.org/subscribers/advertise

Neurology® is the official journal of the American Academy of Neurology. Published continuously since 1951, it is now a weekly with 48 issues per year. Copyright © 2019 American Academy of Neurology. All rights reserved. Print ISSN: 0028-3878. Online ISSN: 1526-632X.