JOURNAL OF NEUROTRAUMA 28:1525–1543 (August 2011)

DOI: 10.1089/neu.2010.1296

A Grading System To Evaluate Objectively the Strength of Pre-Clinical Data of Acute Neuroprotective Therapies for Clinical Translation in Spinal Cord Injury

Brian K. Kwon,[1] Elena B. Okon,[2] Eve Tsai,[3] Michael S. Beattie,[4] Jacqueline C. Bresnahan,[4] David K. Magnuson,[5] Paul J. Reier,[6] Dana M. McTigue,[7] Phillip G. Popovich,[8] Andrew R. Blight,[9] Martin Oudega,[10] James D. Guest,[11] Lynne C. Weaver,[12] Michael G. Fehlings,[13] and Wolfram Tetzlaff[14]

Abstract

The past three decades have seen an explosion of research interest in spinal cord injury (SCI) and the development of hundreds of potential therapies that have demonstrated some promise in pre-clinical experimental animal models. A growing number of these treatments are seeking to be translated into human clinical trials. Conducting such a clinical trial, however, is extremely costly, not only for the time and money required to execute it, but also for the limited resources that will then no longer be available to evaluate other promising therapies. The decision about what therapies have sufficient pre-clinical evidence of efficacy to justify testing in humans is therefore of utmost importance. Here, we have developed a scoring system for objectively grading the body of pre-clinical literature on neuroprotective treatments for acute SCI. The components of the system include an evaluation of a number of factors that are thought to be important in considering the "robustness" of a therapy's efficacy, including the animal species and injury models that have been used to test it, the time window of efficacy, the types of functional improvements effected by it, and whether efficacy has been independently replicated. The selection of these factors was based on the results of a questionnaire that was performed within the SCI research community. A modified Delphi consensus-building exercise was then conducted with experts in pre-clinical SCI research to refine the criteria and decide upon how to score them. Finally, the grading system was applied to a series of potential neuroprotective treatments for acute SCI. This represents a systematic approach to developing an objective method of evaluating the extent to which the pre-clinical literature supports the translation of a particular experimental treatment into human trials.

Key words: Delphi; grading system; neuroprotection; spinal cord injury

[1]Combined Neurosurgical and Orthopaedic Spine Program, Department of Orthopaedics, University of British Columbia, Vancouver, British Columbia, Canada.
[2]ICORD, University of British Columbia, Vancouver, British Columbia, Canada.
[3]Ottawa Hospital Research Institute, University of Ottawa, Ottawa, Canada.
[4]University of California, San Francisco, Brain and Spinal Injury Center, San Francisco, California.
[5]Kentucky Spinal Cord Injury Research Center, University of Louisville, Louisville, Kentucky.
[6]University of Florida, McKnight Brain Institute, Department Neuroscience, Gainesville, Florida.
[7]Department of Neuroscience, The Ohio State University, Columbus, Ohio.
[8]Department of Molecular Virology, Immunology, and Medical Genetics, The Ohio State University, Columbus, Ohio.
[9]Acorda Therapeutics, Inc., Hawthorne, New York.
[10]Departments of Physical Medicine and Rehabilitation and Neurobiology, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania.
[11]Neurological Surgery and the Miami Project to Cure Paralysis, Miller School of Medicine, Miami, Florida.
[12]Spinal Cord Injury Laboratory, BioTherapeutics Research Group, Robarts Research Institute, London, Ontario, Canada.
[13]Division of Neurosurgery and Spinal Program, University of Toronto, Toronto, Ontario, Canada.
[14]Blusson Spinal Cord Center, University of British Columbia, Vancouver, British Columbia, Canada.

Introduction

The urgent need to establish treatments for spinal cord injury (SCI) has led to the development of numerous therapeutic strategies over the past 30 years. A handful have been evaluated in human trials (Tator, 2006), and many more that are emerging from scientific laboratories are vying for clinical translation (Hawryluk et al., 2008; Rowland et al., 2008). Experience has shown, however, that when promising SCI experimental treatments reach the point of translation into human evaluation, their passage through clinical trials is exceedingly challenging. Several extensive clinical trials have resulted in negative or very modest effects (Tator, 2006). With a relatively low annual incidence of traumatic SCI (compared to other acute neurologic conditions such as stroke or traumatic brain injury), patient recruitment into acute SCI trials can be an extremely slow process (Geisler et al., 2001a, 2001b). Large numbers of patients are typically needed to demonstrate neurologic efficacy in such trials, because spontaneous neurological recovery can occur with substantial variability (Fawcett et al., 2007; Lammertse et al., 2007) and there is far greater heterogeneity in clinical populations compared to experimental animal groups. The tremendous commitment of resources necessary for human clinical trials compels the scientific community to decide carefully which of the many experimental treatments available has been sufficiently studied to justify advancement into human trials (Dietrich, 2003). While the decision to proceed with a clinical trial comes with significant financial costs and time commitments, there is also an opportunity cost associated with the potential inability to study other promising therapies. Aside from the issues of time and expense, the expectation of many SCI patients as clinical trial subjects is that the therapies they are being recruited to test might actually be effective for them – even if the research questions are focused on safety and feasibility. Clearly, the decision to proceed with the clinical translation of experimental treatments is not one to be taken lightly.

Despite the significant resource issues at stake when making the decision to translate experimental therapies into human clinical trials, no objective method exists for characterizing the extent to which a particular therapy for SCI has been scientifically investigated and critically evaluated for its readiness for translation. Members of the stroke field, having suffered substantial frustration in the clinical evaluation of neuroprotective treatments that appeared "promising" in animal studies (O'Collins et al., 2006), have generated useful guidelines to direct the pre-clinical development of such interventions (Fisher, Hanley, et al., 2007; Fisher, Feuerstein, et al., 2009). Developing analogous guidelines in SCI should be valuable, not only to provide direction about the testing of potential SCI therapies, but also to identify important gaps in the scientific validation and pre-clinical preparation of specific treatments.

To generate a measure for evaluating the pre-clinical body of literature for a specific therapy, we took a three-step approach. First, we began with an attempt to garner the perspectives of the scientific and clinical communities regarding what they felt were important "evidentiary landmarks" in experimental SCI studies that should be met prior to proceeding with clinical translation (Kwon et al., 2010a). This was done with a 63-item questionnaire that was circulated to both scientific principal investigators within the SCI field, and clinicians who were involved in the care of spinal injury patients. We then used the results of this questionnaire to generate a preliminary scoring system that incorporated the perspectives of the respondents, particularly those who were scientific principal investigators and clinicians. Second, a focus-group meeting of scientific experts and surgeon-scientists within the SCI community was held, and by conducting a modified Delphi process that involved discussion and anonymous voting, the design and weighting of the scoring system was refined. Third, the resultant scoring system was applied to the body of systematically reviewed literature on specific neuroprotective therapies as an illustration of its utility (see Kwon et al., 2010b).

Methods

Scientific community survey

A 63-item questionnaire was developed in the fall of 2008 by the lead and senior authors (BK and WT) to survey the opinion of clinical and scientific members of the SCI community on: (1) the extent of pre-clinical evidence necessary to justify translating a potential therapy into a clinical trial; (2) the methodology and outcome measures widely utilized in animal-based research to generate such evidence; and (3) the biases that influence the interpretation of that evidence. These questions related to issues that are often discussed amongst SCI researchers. The questionnaire asked respondents to indicate their level of agreement or disagreement with a series of minimally ambiguous statements. For each statement, respondents were asked to indicate if they "strongly disagreed," "mildly disagreed," "neither agreed nor disagreed," "mildly agreed," or "strongly agreed." Additionally, questions were asked about the timing of therapeutic interventions in both pre-clinical and clinical studies.

The three-page questionnaire was distributed at four neuroscience and clinical conferences in November/December, 2008 (the Canadian SCI Solutions Network Annual Meeting and Toronto Rehab Network SCI Precourse Meeting in Toronto, Ontario; the Society for Neuroscience Annual Meeting in Washington, DC; and the Cervical Spine Research Society Meeting in Austin, TX). Additionally, the Microsoft Word version of the questionnaire was distributed via e-mail to spine surgeons, clinicians, scientists, regulatory officials, and industry representatives. Respondents were given the option of including their name and e-mail, or could remain anonymous by either mailing or faxing their questionnaires back. The questionnaire and study design was reviewed and granted approval by the Behavioral Research Ethics Board of the University of British Columbia.

Focus-group meeting and modified Delphi to establish scoring system

A group of 25 scientific experts and spinal surgeon-scientists were invited to participate in the focus-group meeting and modified Delphi consensus-building exercise. For scientific principal investigators, we applied the inclusion criteria that they be actively running a basic science research program in SCI and that they had published as the senior or lead author at least *one* peer-reviewed scientific article in the past year that involved an SCI therapeutic approach. For surgeon-scientists, we applied the inclusion criteria that they be both the

principal investigator in their own basic science laboratory studying SCI *and* were actively involved in the treatment of acute SCI patients.

To generate a preliminary scoring system for pre-clinical literature on a specific SCI therapy, we pooled the questionnaire responses of the scientific principal investigators, clinician scientists, and clinicians (spinal surgeons and non-surgical physicians) on specific issues germane to the question of translating SCI therapies (such as the appropriate animal models to study prior to human trials). This included a total of 235 questionnaire respondents. The preliminary scoring system and the overall rationale and plan for the focus-group meeting was then distributed for review amongst the 25 scientific experts and surgeon-scientists in the SCI field. We then hosted a 3-h focus-group meeting with these individuals during the 2009 Society for Neuroscience meeting. The purpose of the meeting was to garner "expert" opinion regarding the particular weighting of pre-clinical evidence within the scoring system. The questionnaire responses of those who confirmed their attendance were tabulated and presented at the focus-group meeting in addition to the responses from the larger group of 235 scientists and clinicians. Those who had not responded to the original questionnaire completed the questionnaire prior to the meeting so their opinions could be included in the presentation of the group's perspectives. Anonymous voting on subsequent questions related to the questionnaire and preliminary scoring system was then conducted using handheld Audience Response Systems (ARS), and the immediate display of voting responses was used to facilitate further discussion. Using this modified Delphi approach, areas in which consensus existed and where consensus clearly did not exist were identified. We operationally defined 81% to 100% to represent "strong agreement," 61% to 80% to represent "moderate agreement," and 60% or less to represent "poor agreement." The results of the anonymous voting and transcription of the recorded discussions were incorporated into a revised scoring system, which was circulated to focus-group members for revision to establish a final scoring system. There were two married couples in the focus-group meeting; each individual, however, had his/her own ARS keypad for anonymous voting, and correspondence was sent to each individual separately.

Systematic review of pre-clinical literature

We conducted a systematic review of the pre-clinical literature on experimental SCI treatments that were already in human clinical use for some other related or unrelated conditions, or, if not currently in human use, were currently available in a form that could be given to humans (see Kwon et al., 2010b). Common to these agents was their administration by systemic, non-invasive methods (i.e., therapies applied by direct injection or transplantation into the cord were reviewed elsewhere). A PubMed search on the particular "treatment" and "spinal cord injury" was performed, and the resultant articles were included in the systematic review if they met the following criteria: (1) studies in which the testing of the experimental therapy was performed in an *in vivo animal model* of SCI; (2) studies in which the spinal cord was *traumatically* injured; (3) studies in which the application of the therapy was via the *systemic circulation*; and (4) at least *two peer-reviewed publications* available on the therapy. The scoring system was then applied to these therapies to illustrate its use.

A schematic of the methodological steps in devising the scoring system is provided in Figure 1.

Results

A total of 324 responses were received between November 2008 and January 2009, a response rate of 46% of approximately 700 questionnaires that were distributed. A total of 105 respondents classified themselves as "scientific principal investigators" running a laboratory-based research program, 34 as "clinician scientists" (clinicians who additionally operate a laboratory-based research program), 76 as "spinal surgeons," 20 as "clinicians" (non-surgical), 69 as "trainees" (graduate student, post-doctoral students, research associates), and 20 as "others" (industry or research foundation representative, clinical research assistants, regulatory officials). SCI patients were not asked to participate. The analysis of the entire 324 respondents as a whole has previously been reported (Kwon et al., 2010a). For the purposes of this current initiative, we specifically examined the opinions of the scientific principal investigators, clinician scientists, spinal surgeons, and clinicians, as we felt that these were the individuals whose opinions were most relevant to the pre-clinical development of new therapies and the subsequent decision to translate them into human trials. In total, this comprised the opinions of 235 respondents (105 scientific principal investigators and 130 individuals with a clinical background). The data from the questionnaire, which are presented subsequently, represent the responses from these 235 individuals.

For the focus-group meeting and modified Delphi consensus-building exercise, expertise in a variety of aspects of pre-clinical SCI research was sought amongst the attendees. Of the 25 individuals who had confirmed their participation, 21 attended the meeting, which was held over 3 h on the evening of October 17, 2009. Of these, 17 were basic scientists (PhDs) running pre-clinical SCI research programs, and 4 were spine surgeon-scientists (MD/PhDs) who treated SCI patients in addition to running a pre-clinical research program. The attendees of the meeting represented a wide spectrum of expertise in the SCI field, having been involved in SCI research for approximately 22.8 ± 2.0 years and having published approximately 65.3 ± 11.4 peer-reviewed articles on the topic of SCI (mean $\pm$ SEM of self-reported estimates). The attendees divided their research time amongst the following research areas: 18 reported being involved in neuroprotection research, 14 were involved in cell transplantation treatments, 13 studied treatments to promote axonal growth/sprouting (e.g., Chondroitinase ABC), 9 were involved in some form of rehabilitation training research, 5 reported involvement in acute human SCI studies, and 2 reported involvement in chronic human SCI studies. The format was that questions were presented on a screen, and the attendees were asked to respond amongst predetermined choices. The option not to respond if an attendee felt inadequately informed was available. After the initial determination of the range of responses, discussion ensued. For some questions, there was a rewording of the question and a new round of anonymous electronic response to determine if additional consensus could be obtained (Delphi process).

Animal species used in spinal cord injury research

Questionnaire results. The first series of questions related to the importance of different animal species in the

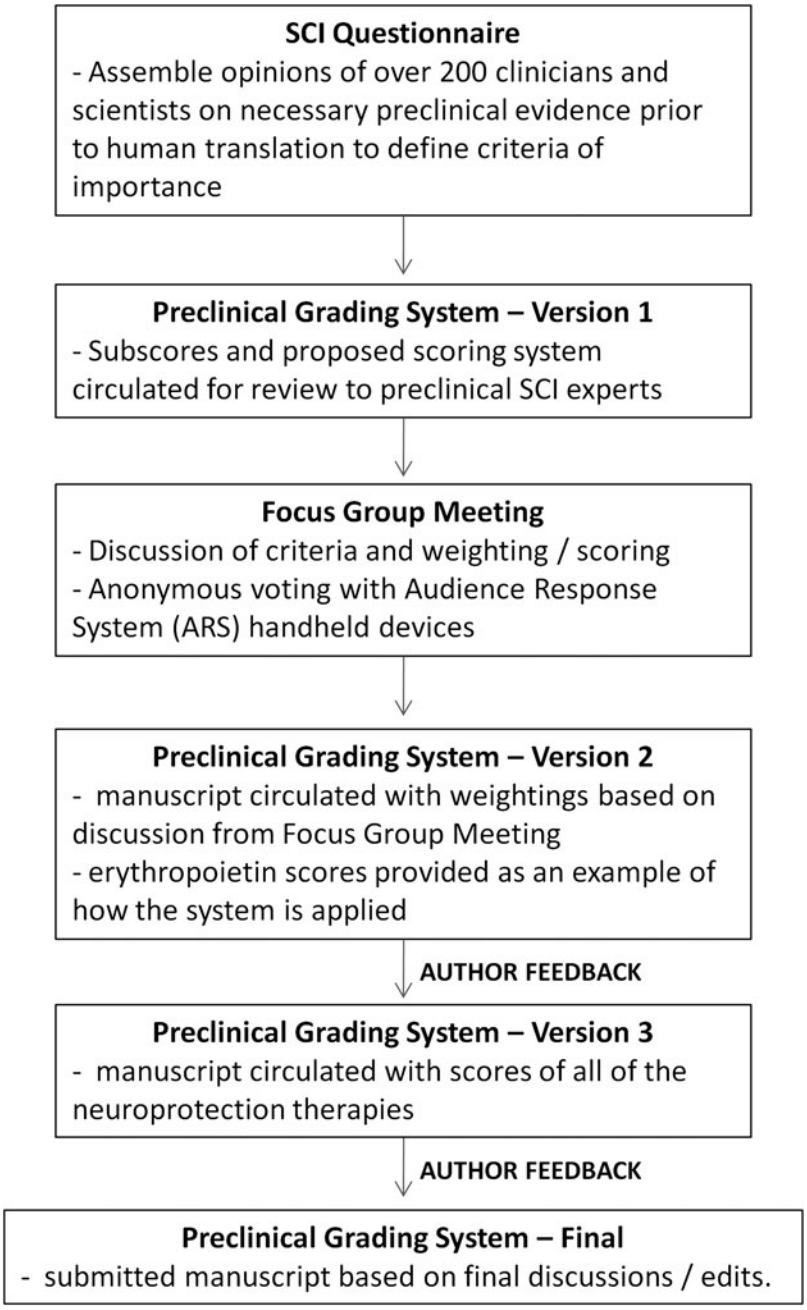

FIG. 1. Schematic of the methodological steps to developing the pre-clinical grading system.

demonstration of efficacy of non-invasive drug therapies in pre-clinical studies (Fig. 2). The majority of respondents (56%) disagreed that the demonstration of therapeutic efficacy in a rodent model of SCI alone was sufficient to proceed with human clinical trials. However, 39% stated that efficacy in rodents was sufficient to proceed with human clinical trials. The study revealed moderate agreement for demonstrating efficacy in a large-animal model of SCI (68% agreed vs. 19% disagreed). The respondents were divided on the need for primate models in the pre-clinical substantiation of such non-invasive drug therapies prior to human translation.

Focus-group meeting discussion. The opinions regarding animal species from the questionnaire were contradicted to some extent by the attendees of the meeting, insofar as 83% either strongly disagreed or mildly disagreed with the need for primate models prior to translating non-invasive drug therapies. Support for large-animal models was also less, with only 50% of the attendees strongly or mildly agreeing with the need for large-animal models in this context. With respect to rodent models, 69% of the attendees voted that studies utilizing rat models were considered to be "more relevant" than those in mouse models, while the remaining 31% felt that they were equal. While the need for large-animal models and primate models for the translation of non-invasive therapies was questioned, 73% of the attendees voted that they would consider a large-animal study to be more clinically relevant than a rodent study (the remainder voted that they would consider

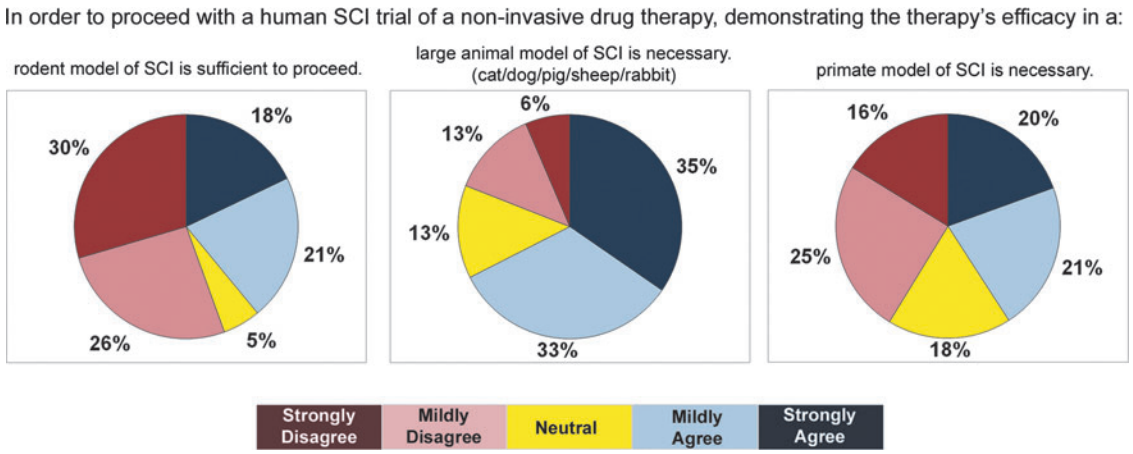

FIG. 2. The opinions of clinicians and scientists about animal species used in pre-clinical SCI research. Color image is available online at www.liebertonline.com/neu.

them equally), and 83% voted that they would consider a primate study to be more clinically relevant than a large-animal study.

Aside from the anonymous voting on specific questions, the dialogue on animal species included a number of comments and perspectives that warrant discussion. Despite the clear preference for studies in rat models over mouse models, it was pointed out that depending on the research question and pathophysiology being studied, murine models might in some cases be more representative of the human condition. Also, while the voting demonstrated a preference toward higher-order animal species, it was pointed out that the sentiment that a therapy would more likely succeed in humans if it were shown to be effective in large-animal or primate models as compared to rodent models alone was a largely intuitive yet untested assumption in SCI research. This highlights the fact that because a convincingly efficacious neuroprotective therapy for human SCI is currently lacking, an example of a "successful" pre-clinical research pathway is lacking. Finally, it was pointed out that while preference for large-animal and primate models exists, such models with well-characterized biomechanical injury parameters and functional outcome measures are not readily available.

Weighting within the animal species subscore. With the results of the questionnaire and the voting of the attendees of the focus group in mind, the following scoring system for the evaluation of animal species utilized in the study of a non-invasive drug therapy was generated. The demonstration of efficacy in a mouse model of traumatic SCI received a score of 2; for rat models, a score of 4; for large-animal models including dog, cat, rabbit, pig, and sheep, a score of 6; and for primate models a score of 8 (see Table 1). Here, we attempted to "balance" the opined hierarchy of mouse, rat, large animal, and primate studies with the fairly strong opinion that primate models and even large-animal models were not necessarily imperative for the development of non-invasive drug therapies. In essence, the question posed here was not whether a primate study was necessary (and thus weighted more heavily), but rather, if assessing a therapy that had been tested in both a rodent model and a primate model, how one would view the primate study in comparison to the rodent study if both studies were equally sound scientifically. The temptation to assign the primate studies more than twice the weight of the rodent studies (they clearly require much more than twice the effort and resources) was mitigated by the opinions on the necessity for such studies.

One of the topics that was extensively discussed at the focus-group meeting was the issue of efficacy. Even though primate and large-animal models of SCI are relatively rare, and behavioral outcome metrics such as the widely used BBB score are not well established, it was still felt that demonstrating "efficacy" in these large-animal models required improvements in both behavioral and non-behavioral outcome measures. Hence, for all aspects of the "animal species" subscore, achieving the criteria of demonstrating efficacy requires both behavioral and non-behavioral improvements. Therefore, if a published study of a therapy utilizes a sheep model and demonstrates histologic improvements but no locomotor improvements, the therapy would not collect a score of 6 in this animal species subscore (unless another publication exists in a large-animal model that demonstrates efficacy in both). For example, in Ozdemir and colleagues (2005), magnesium sulfate was evaluated in a rabbit model of SCI, and while magnesium was reported to promote improvements in biochemical outcomes (reducing lactate and malondialdehyde levels), no behavioral outcomes were reported. Hence, magnesium does not earn a 6 for application in a large-animal model.

Injury models and paradigms utilized in spinal cord injury research

Questionnaire results. A total of 71% of respondents to the questionnaire agreed with the statement that the contusion injury model was the *most* clinically relevant model, while only 20% agreed that the calibrated clip compression model was the *most* clinically relevant model. Importantly, only 9% of the respondents voted that models other than contusion/compression lesions were the most clinically relevant. The group was divided on the question of whether partial transection SCI models were valid for studying acute neuroprotective therapies, although the majority opined that they were not valid (56% vs. 39%). We also asked whether animal models of cervical SCI were necessary before proceeding with human studies in which cervical SCI patients

Table 1. Preclinical Grading Scale

Animal species in which efficacy has been demonstrated*	*Points*
Primate model of traumatic SCI	8
Large animal model of traumatic SCI (dog, cat, rabbit, pig, sheep)	6
Rat model of traumatic SCI	4
Mouse model of traumatic SCI	2
Maximum score:	20
Injury paradigms in which efficacy has been demonstrated*	*Points*
Cervical contusion SCI models	6
Thoracic contusion SCI model	3
Cervical clip compression SCI model	6
Thoracic clip compression SCI models	3
Cervical partial transection sharp SCI model	1
Thoracic partial transection sharp SCI model	1
Maximum score:	20
*Time window of efficacy**	*Points*
Efficacy demonstrated with a treatment delay of 12 h or more	8
Efficacy demonstrated with a treatment delay of 4 h or more, but less than 12 h	6
Efficacy demonstrated with a treatment delay of 1 h or more, but less than 4 h	3
Efficacy demonstrated when treatment given immediately at the time of injury or within less than 1 h	2
Efficacy demonstrated when treatment is given prior to injury	1
Maximum score:	20
Demonstration of "clinically meaningful" efficacy	*Points*
THORACIC SCI MODEL: Achievement of plantar weight support (i.e., BBB of 9) versus controls that do not, or the achievement of consistent forelimb–hindlimb coordination (i.e., BBB of 14) versus controls that do not, in a study with associated improvements in non-behavioral outcomes (e.g., tissue sparing).	4
THORACIC SCI MODEL: Significant improvement in other locomotor or motor behavioral tests (e.g., quantitative gait analysis, inclined plane, swimming) or other non-motor behavioral tests (e.g., pain, autonomic dysreflexia) in a study that also demonstrates associated improvements in non-behavioral outcomes (e.g., tissue sparing).	4
CERVICAL SCI: Significant improvement in some motor function test (e.g., food-pellet reaching, grasping, quantitative "gait" assessment) in a study that also demonstrates associated improvements in non-behavioral outcomes.	4
CERVICAL SCI: Significant improvements in other non-motor behavioral tests (e.g., pain, autonomic dysreflexia) in a study that also demonstrates associated improvements in non-behavioral outcomes.	4
DOSE RESPONSE: Demonstrated in a single study using either the thoracic or the cervical SCI model. The dose "response" is defined by improvements in *either* behavioral or non-behavioral outcomes with changing doses of the therapy.	4
Maximum score:	20
Independent reproducibility/replication	*Points*
More than 10 independent laboratories report on the beneficial effects of the therapy	20
5–10 independent laboratories report on the beneficial effects of the therapy	12
3–4 independent laboratories report on the beneficial effects of the therapy	7
2 Independent laboratories report on the beneficial effects of the therapy	3
1 independent laboratory reporting on the beneficial effects of the therapy	0
1 independent laboratory reports on the negative results on the therapy	−3
2–3 independent laboratories report on the negative results of the therapy	−7
4–9 independent laboratories report on the negative results of the therapy	−12
More than 9 independent laboratories report on the negative results of the therapy	−20
Maximum score:	20

would be enrolled. This statement received agreement from the majority (55%) of respondents (Fig. 3).

During the focus-group meeting, we addressed three issues: the preference for contusion over compression injuries, the validity of incomplete transection injuries for neuroprotective studies, and the importance of cervical injury models. Analysis of the attendees' responses to the questionnaire revealed similar opinions to the larger body of questionnaire respondents with respect to the injury models. A total of 80% strongly or mildly agreed with the statement that the contusion model was the most clinically relevant model, while 28% had mildly agreed with the statement that the clip compression model was the most relevant. When asked to vote specifically about the contusion versus compression, 55% voted that the contusion model was the most relevant, while 38% viewed contusion and compression models to be equal in relevance, suggesting that a strong consensus on this issue did not exist. When asked how they would "weight" a study utilizing the contusion injury versus the compression injury, again, there was a lack of clear consensus, with 41% assigning

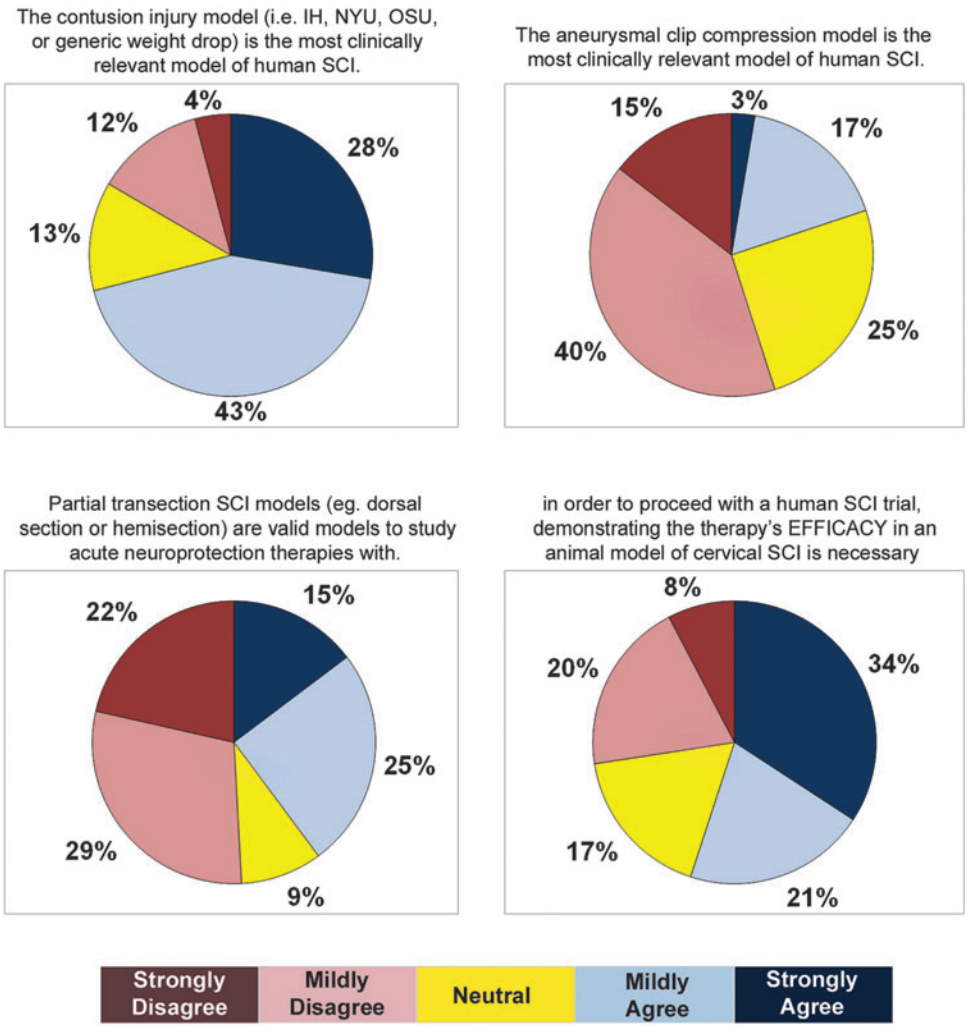

FIG. 3. The opinions of clinicians and scientists about injury models used in pre-clinical SCI research. Color image is available online at www.liebertonline.com/neu.

more weight to contusion injuries, 18% assigning more weight to compression injuries, and 41% assigning equal weight to them.

With respect to the issue of incomplete transection models in the questionnaire, 64% of the attendees had strongly or mildly disagreed with the statement that these were valid models to study acute neuroprotective therapies, while 36% had strongly or mildly agreed that they were valid. When asked during the focus meeting specifically about the relative relevance of blunt (contusion or compression) and incomplete transection models, 89% voted that they viewed the blunt injury model to be of greater relevance, while 11% voted that blunt and partial section models were of equal relevance. When asked to assign a weighting to blunt models with respect to partial section models, 50% and 25% indicated that they would assign, respectively, triple or double the weight to studies utilizing blunt injury models.

Finally, with respect to the issue of cervical versus thoracic injury models, the attendees were first asked if they considered one particular model to be more relevant than the other for the development of acute neuroprotective therapies. Worded in this fashion, 61% voted that they were of equal relevance, while the remaining 39% voted that the cervical models were more relevant. After some discussion, the question was posed again within the context of a therapy being administered to a cervical SCI patient. Under these conditions, 94% voted that the cervical models were more relevant than thoracic models, and only 6% voted that the two were equally relevant.

The discussion of injury models was vigorous, particularly around the issue of the contusion versus compression injury models. It was commented that the apparent preference for the contusion injury model – both in the large body of questionnaire respondents and amongst the "SCI experts" attending the meeting – might in part reflect the relative familiarity with and the relative use of these various models. Merits of both contusion and compression models were discussed at length, and emerging from this came the general agreement that demonstrating efficacy in *both* injury models would be an important evidentiary landmark, recognizing that, across the spectrum of human cord injury mechanisms, there are differing rates of sudden and violent contusive force and varying degrees and durations of compression. There was a lack of a consensus regarding contusion versus clip compression, although it was evident that some felt strongly about the superiority of the contusion injury model, while

others felt strongly about the superiority of the compression model. The attendees clearly were in consensus regarding the relative importance of blunt injury models over partial section models, and cervical injury models for the development of treatments administered to cervical SCI patients.

Weighting within the injury model subscore. Based on the data from the questionnaire and the discussion during the focus-group meeting, the following weighting scheme was proposed (see Table 1). Due to the lack of clear consensus regarding the relative merit of contusion versus clip compression, we assigned these equal weighting. While this could be debated given the opined preference of the contusion injury model in the questionnaire, a clear consensus was not achieved about the relative weighting of the contusion and clip compression injuries, and hence we weighted them equally. Given that any clinical trial of a non-invasive and potentially neuroprotective intervention will likely be administered to cervical SCI patients (and possibly *exclusively* to cervical SCI patients), the strong consensus within the group prompted us to weight cervical injury models substantially higher than thoracic injury models. And considering the strong preference for blunt injury models, we weighted partial section models lightly, irrespective of whether they were conducted in the cervical or thoracic region. Hence, we assigned cervical contusion or clip compression injuries a score of 6, thoracic contusion or clip compression injuries a score of 3, and partial section injuries a score of 1. To recognize the perceived value in having efficacy demonstrated in different injury models, each of these injury models was weighted individually, so that a therapy could accrue more points for being shown in both a contusion and compression injury model, for example. Hence, in theory, a treatment with efficacy demonstrated in cervical and thoracic contusion, clip compression, and partial injury models would score a total of 20 points. As in the "animal species" subscore, efficacy here represents the demonstration of improvements in both behavioral and non-behavioral outcome measures. Therefore, if a published study of a therapy utilizes a cervical contusion and demonstrates histologic improvements but no locomotor improvements, the therapy would not earn a 6 in this injury paradigm subscore for this study.

Time window of efficacy

Questionnaire results. In this question, we posed the hypothetical scenario of an acute neuroprotective therapy that was administered within 12 h of injury in humans and asked respondents what time window of efficacy would be needed in pre-clinical studies to justify this inclusion criteria in human trials. Responses were grouped in the following manner: 0–1 h, 2–4 h, 6–8 h, or 12–24 h. Interestingly, the most prevalent responses were 12–24 h followed by 6–8 h, 2–4 h, and fewer than 10% of individuals believed that a 0–1 h time window was sufficient (Fig. 4).

Focus-group meeting discussion. The attendees' responses to the questionnaire on this topic had approximately 33% choose 2–4 h, 18% for 6–8 h, and 42% for 12–24 h. The attendees were asked specifically whether they felt that the time window of pre-clinical studies should match that of the human inclusion time window, or whether a shorter time window could be appropriately used. A total of 44% voted for the equivalent time window, while 56% voted that a shorter time window could be used. The attendees then voted on what they felt a "minimum" time window would be for an acute neuroprotective therapy being administered to human patients within 12 h of injury. A total of 59% and 24% of the attendees voted for 4 h or 8 h respectively. We then posed specific questions to address the other end of the time-window spectrum: those studies in which the therapy was applied either before the time of injury or immediately after the injury was induced. For studies in which the therapy was administered prior to the injury, 39% voted that they would assign no weight to such studies, while the remaining 61% voted that while they would not ignore these "pre-treatment" studies, they would assign them the lowest weight possible. For studies in which the therapy was administered right at the time of injury, 88% felt that these should be weighted either equal or just marginally heavier than "pre-treatment" studies.

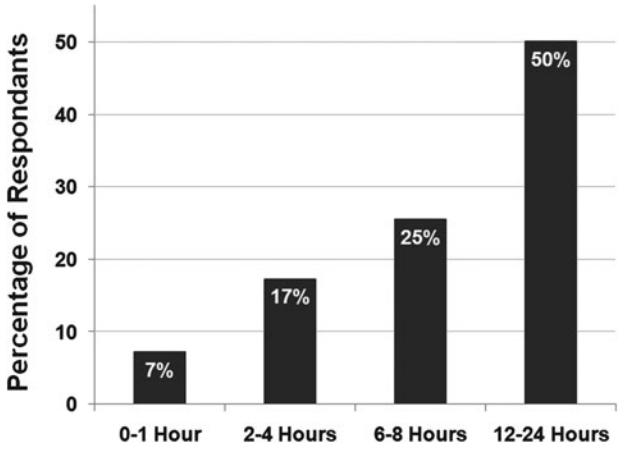

FIG. 4. The opinion of clinicians and scientists about the time window of efficacy requirements for acute neuroprotective therapies.

The discussion around this topic revealed that there is great uncertainty about the temporal progression of pathophysiologic processes in acute human SCI and how this correlates with animal models (rodents in particular). Given this, the discussion regarding appropriate time windows for pre-clinical studies is hampered by speculation and the recognized limitations of our current understanding. There was reluctance therefore in establishing a firm "threshold" for what the minimum time window of efficacy should be as a "standard" for moving a therapy forward into clinical trials. Nevertheless, there was little argument that for a therapy whose efficacy had been demonstrated when administered prior to or at the time of injury, it would be highly desirable to see further evidence of efficacy at later time points.

Weighting within the time-window subscore. With this, we established the following scoring system for time window of efficacy (see Table 1). Studies in which efficacy was demonstrated with pre-treatment were assigned a score of 1, the lowest score possible. Studies with immediate treatment or treatment within 1 h later were given a score of 2. Treatment delays of 1 h but less than 4 h were given a score of 3. Given that the majority felt that 4 h was a "minimum," we assigned studies with a treatment delay of at least 4 h but less than 12 h

a score of 6. Efficacy demonstrated with a delay of 12 h or more was given a score of 8. Here, we would define "efficacy" as the demonstration of improved behavioral *and* non-behavioral outcomes. So, for example, in an evaluation of magnesium, Kwon and colleagues (2009) tested a treatment delay of 2, 4, and 8 h post injury. At 8 h post injury, magnesium promoted statistically significant reduction in lesion size, but the behavioral recovery (BBB) was not statistically significant. At 4 h post injury, magnesium promoted a statistically significant improvement in both lesion size and BBB scores. Hence, for this study, the maximal time window of efficacy is 4 h, not 8 h.

Demonstration of "clinically meaningful efficacy"

Questionnaire results. A myriad of behavioral and non-behavioral outcome measures are used to report the results of experimental treatments in pre-clinical animal SCI studies. These include such measures as locomotor recovery (e.g., BBB scores, catwalk, ladder footfall testing), non-motor recovery (e.g., mechanical allodynia), and histological/anatomical changes (e.g., lesion size, white matter sparing). Here, we sought perspectives on what results and findings were viewed to represent "clinically meaningful efficacy." There was moderate agreement (79%) for the statement that a dose response relationship should be demonstrated before a therapy is tested in clinical trials. With respect to locomotor recovery, there was moderate agreement for the statement that the achievement of plantar weight support on the BBB scale (versus controls that did not achieve weight support) represented "clinically meaningful efficacy." Similarly, the achievement of forelimb/hindlimb coordination on the BBB scale was also felt to represent clinically meaningful efficacy. The majority (61%) regarded non-locomotor behavioral recovery in such aspects as autonomic function and neuropathic pain to be clinically meaningful. In the absence of associated behavioral improvements, the majority (61%) did not view improvements in histological/biochemical/physiological outcomes to be important (Fig. 5).

Focus-group meeting discussion. This topic expectedly generated the most intense discussion (and anticipating this to be the case, we addressed this topic first during the meeting). To frame the discussion for the attendees, we clarified at the beginning of the meeting that the purpose of this exercise was to consider the pre-clinical evidence supporting the argument for human translation of a particular therapy; within this context, it was proposed that studies focused on biologic mechanisms might have less importance than studies with behavioral outcomes. There was vigorous discussion of the merits of this proposition (as one might expect given the scientific background of the attendees). When asked to distinguish "behavioral outcomes" such as locomotor function or "non-behavioral outcomes" such as histological evidence of tissue sparing, 70% voted that the behavioral outcomes were more important than the non-behavioral outcomes in determining a therapy's potential for translation, and 30% voted that they were equal. However, in the ensuing discussion, it was clear that many attendees found this distinction to be

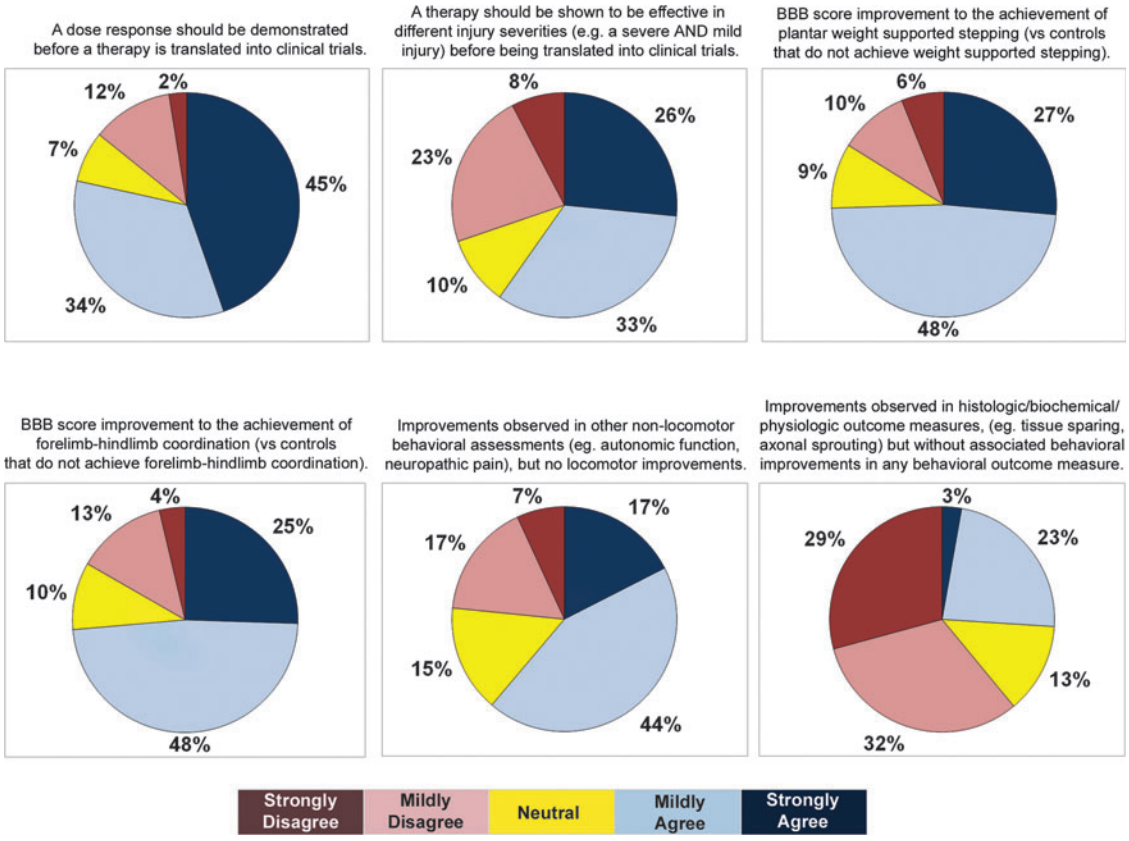

FIG. 5. The opinion of clinicians and scientists about what constitutes "clinically meaningful efficacy." Color image is available online at www.liebertonline.com/neu.

difficult, and they assigned great importance to the demonstration of *both* behavioral and non-behavioral improvements. While there was general agreement that one would never translate a treatment into humans solely on the basis of histological findings in the absence of some demonstrable functional benefit (and hence the importance of behavioral outcomes), there was also a strong sentiment that a treatment with only a behavioral effect (e.g., improved hindlimb function) but no associated improvement in any non-behavioral outcome measure (e.g., tissue sparing, axonal sprouting) would be viewed as being premature for translation.

With regards to behavioral outcomes measured in preclinical research, there are many different metrics used, and we attempted to delineate how these were viewed with respect to one another. The attendees were asked about the relative importance of those outcomes assessing some form of motor recovery and those "non-motor" outcomes addressing such issues as autonomic dysreflexia or pain. For acute neuroprotection studies, 58% felt that all of these types of outcomes were equal, while 37% felt that the motor outcomes were more important. Amongst the motor outcomes available for thoracic SCI models, 65% felt that improvements on the BBB scale were just as important as improvements on other locomotor tests (e.g., catwalk/quantitative gait assessment, ladder stepping), while 30% felt that the BBB score improvements were less important. With respect to the BBB scale, 88% voted that the achievement of weight-supported stepping in treated animals compared to control animals that did not achieve weight-supported stepping was considered to be "clinically meaningful efficacy." A total of 80% voted that the achievement of forelimb–hindlimb coordination in treated animals compared to control animals also represented "clinically meaningful efficacy."

Recognizing that most of the outcome measures related to thoracic SCI models, we also attempted to garner the perspectives of the attendees on behavioral outcomes in cervical SCI models. Given the choice between the food pellet reaching test, grasping/grip strength, forelimb use in locomotion (catwalk/quantitative "gait" assessment), cylinder test (rearing/vertical exploration within a cylinder), horizontal ladder (forelimb errors while crossing), and sticker removal test, 31% voted that they considered all to be equal, 31% voted that the food pellet reaching test was most relevant, and 19% voted that the grasp/grip strength test was most relevant. It was implied during the discussion that, for cervical models, behavioral metrics should include an assessment of forelimb function, given that upper extremity function is the chief priority of quadriplegic patients.

Finally, on the issue of "dose response," we asked whether the "response" would have to be in both behavioral and non-behavioral outcome measures in order to fulfill this criterion for "clinically meaningful efficacy." Here, 100% of the attendees voted in favor of considering either a behavioral or non-behavioral effect at different doses as evidence of a "dose response."

Weighting within the clinically meaningful efficacy subscale. We attempted to include the following general perspectives voiced at the focus-group meeting in developing the subscore for the demonstration of clinically meaningful efficacy. It was evident that while behavioral outcomes were considered by some to be more important than non-behavioral outcomes for the decision to translate a specific SCI therapy, the attendees assigned greater strength to the demonstration of both. Hence, in this subscore, points for the demonstration of clinically meaningful efficacy were assigned only if a study reported *both* a behavioral improvement and some associated non-behavioral improvement. For thoracic SCI models, milestones for what was considered to be a "significant improvement" on the BBB scores were applied based on the discussion, but improvements in other tests of motor function or tests of non-motor functions (e.g., pain) were also considered with equal importance.

For cervical SCI models, because of the lack of consensus regarding what the "most relevant" motor outcome measure was, we graded them all equally. It is recognized that as cervical models become more commonly employed, further behavioral outcome measures for forelimb function will be developed. We also recognized that hindlimb locomotor testing may also be done in cervical models, and these results are difficult to discount, even if upper extremity recovery is a priority for cervical quadriplegics. Hence, for any motor improvement in the cervical model, we assigned a score of 4. We additionally assigned equal weight to non-motor behavioral outcome measures such as pain or autonomic dysreflexia in these cervical models. The subscore therefore included "equal opportunity" for accruing points in thoracic and cervical SCI models. Hindlimb tests such as the BBB score in cervical models would be evaluated using the same criteria as in thoracic models, but would be considered under the heading of "any motor function test" for cervical SCI models. This eliminates the possibility that a cervical SCI study that noted significant improvements in forelimb motor recovery, pain, BBB scores, and catwalk would achieve a maximal score of 8 and not 16 in the "clinically meaningful efficacy" section.

Dose response (as defined as either behavioral or non-behavioral effects) could be demonstrated in either cervical or thoracic injury models. Given that the subscore title of "clinically meaningful efficacy" implies some element of relevance to the human setting, we propose that results from studies in which a drug was applied prior to the injury not be included in the calculation of this subscore. Such studies which reveal efficacy (both behavioral and non-behavioral) are not ignored, insofar as they do achieve a score of 1 in the "time window of efficacy," but it did not seem reasonable to consider their demonstration of efficacy in the same manner as that of studies that instituted a lengthy delay between injury and intervention.

For the BBB scores of achieving plant weight support (BBB of 9) or consistent forelimb–hindlimb coordination (BBB of 14), we would use the "rounded up" BBB score reported in the article, given that the BBB scale only has integers. So, for example, if the average BBB was 8.8, we would round up to 9.

Independent reproducibility/replication

Questionnaire results. The question that evoked the most unanimous response related to the issue of independent replication. Over 95% of respondents agreed with the statement "If a promising therapy has been developed primarily by a single laboratory, its efficacy should be replicated by an independent laboratory before it is translated into clinical trials." The response was strongly agree in 75% of the respondents, and mildly agree in 21%. This strong sentiment

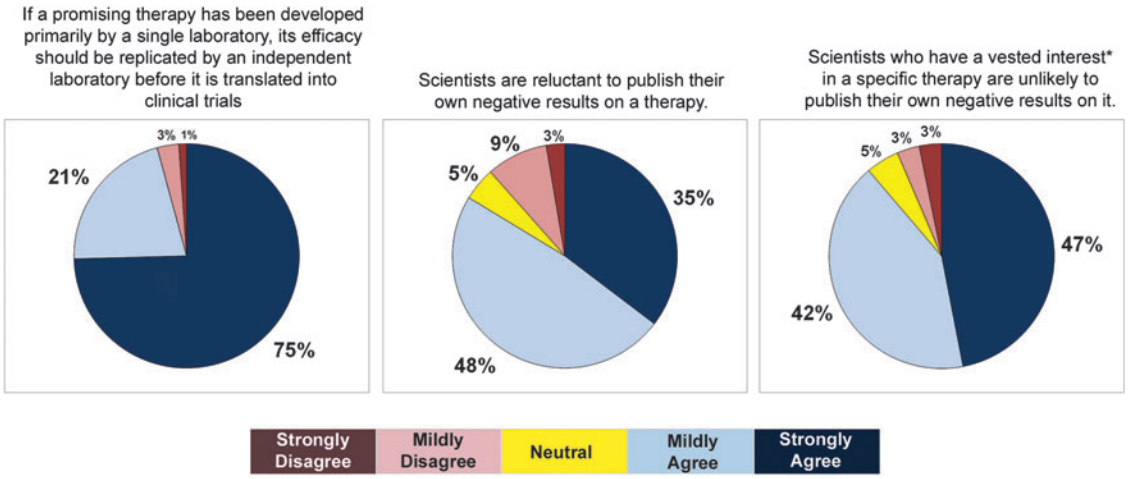

FIG. 6. The opinion of clinicians and scientists about perceived bias in pre-clinical SCI research. Color image is available online at www.liebertonline.com/neu.

for independent replication may be attributable in part to a widely perceived bias against the publication of negative data. A total of 83% of respondents agreed with the statement "Scientists are reluctant to publish their own negative results on a therapy." For scientists who had a vested interest in a specific therapy (meaning that they were either well known for it or had commercialized it), the expectation that negative results would not be published was even higher (89%; Fig. 6).

Focus-group meeting discussion. The questionnaire responses from the attendees were quite similar to that of the larger group of clinicians and scientists, with 17% mildly agreeing and 83% strongly agreeing that independent replication was necessary. The issues we attempted to address during the meeting were how to view studies reporting negative results, particularly when these were "formal replication" studies. All attendees voted that they would consider negative studies in the overall assessment of a therapy's "promise," and provided that a negative study was as well executed and carefully described (i.e., considered to be of "equal quality"), 81% voted that they would assign equal weight to such a study as they would a positive study. The equal weighting of negative studies was interesting on two fronts. First, by assigning equal weight to the publication of a well-performed negative study, there was no "adjustment" in the weightings to take into account the very strong perception that many negative studies are not being submitted for publication (a phenomenon that was personally confirmed by individuals in the room). Second, and perhaps most importantly, the equal weighting was viewed to be a strong message to the scientific community and editors of peer-reviewed scientific journals that a carefully performed experimental study with negative results should be considered with equal merit and importance to a positive study, despite the commonly perceived "lack of novelty" of the former.

An interesting discussion then ensued around the issue of formal replication studies and how to view negative results from these studies in comparison to negative results from labs not explicitly attempting to reproduce the experimental conditions described by another laboratory. There were a number of important points raised during this discussion. First, there was general agreement that there was indeed some merit to attempting to reproduce the efficacy of a therapy using exactly the same experimental conditions, in an effort to confirm the robustness of the therapeutic approach. There was, however, an opinion that the robustness of the therapy might be better revealed by demonstrating efficacy in *different* experimental conditions – an opinion that we attempt to capture, for example, in the subscores that assign weight to the demonstration of efficacy in different animal models and injury models. Second, a number of individuals pointed out that even for the most well-intended and carefully performed formal replication study, it was difficult – if not impossible – to reproduce every experimental condition from the "index" study. An anecdotal example was provided about marked differences in secondary damage observed after the same thoracic contusion injury in the same rat species of the same age, gender, and weight obtained from the same company but from different distributors.

Counterbalancing these concerns about the limitations of formal replication studies was the described analogy between such experimental studies and Phase 2 or Phase 3 clinical trials. It was put forth that the vast majority of scientific experimental studies (and their subsequent publications) are akin to Phase 2 clinical trials where the treatment is administered, a host of behavioral and non-behavioral outcomes are measured, and the outcomes that are positive are reported as evidence that the treatment was effective. In contrast, formal replications are more like Phase 3 clinical trials where the exact primary outcome that will be used to determine the effectiveness of the treatment is determined *a priori*, and must be achieved as prescribed in order for the agent to be considered "efficacious." This is a considerably higher bar to set in both clinical and experimental settings, and thus argues in favor for the importance of replication studies (despite their inherent limitations). Along the lines of this discussion, it was commented that a positive replication study might then be considered a very important evidentiary finding. When then asked to vote about this issue, 56% voted that they would

weight a negative formal replication study with equal weight to a positive study – in essence, viewing the negative formal replication study to be no different from any other negative study done without a formal attempt to replicate the results of another lab. The remaining 44% assigned the formal replication study twice the weight, indicating a perceived distinction for these studies. It was recognized that the achievement of independent "replication" would be an important demonstration of an experimental therapy's "robustness," particularly given the difficulties to date in replicating promising data, and the recognition that study-to-study variation inevitably exists, even in the most carefully conducted "replication" study.

Weighting within the independent replication subscale. The weightings within this subscale were mathematically arranged to reflect the following two sentiments: (1) the strong agreement (81%) that negative studies should be considered equally (but not necessarily more) to positive studies; and (2) the lack of consensus on the relative weighting of formal replication studies (56% equal weight versus 44% greater weight; Table 1). Given that everyone in the focus-group meeting indicated that negative studies should be considered in some manner (i.e., not ignored), it is proposed in the subscale to assign points for the accumulation of studies describing positive results on a therapy, and then to assign equivalent "negative" points for the accumulation of studies describing negative results on the same therapy. Given the arguments around replication studies, these were not assigned any distinct weighting either for negative or positive studies. Also, given that this category reflects "independent" replication and reproducibility, the weightings were assigned not according to the number of published studies, but rather on the number of independent laboratories that had reported on the therapy.

There were some additional considerations in the assigning of scores. Importantly, if an article reported a beneficial/positive effect of the therapy, it was counted as a positive study, even if the reported effects did not achieve points on the "clinically meaningful efficacy" subscore. For example, if a therapy had significantly improved BBB scores from 10 to 12, while this would not be sufficient to get the four points in the clinically meaningful efficacy subscore, we would count it as a "positive" study for the purpose of assessing independent replication. Also, if an article reported positive behavioral outcomes (e.g., improved BBB scores) but negative non-behavioral outcomes (e.g., no change in lesion sparing), we viewed this to be a "neutral" study, and it was neither "positive" nor "negative." The same applied for a study with improved histology but unimproved behavioral outcomes: it seemed appropriate to describe this as neither a positive nor a negative study. In the scoring system, it is assumed that there is always an "index" positive study from one lab, and so the therapy gets a score of 0 if there are no other studies from any other lab. If there is another study reporting beneficial effects from an independent lab, then the therapy has been reported by two independent laboratories and it receives a score of 3. Hence, the equivalent score of −3 points would be given with one additional negative study.

A consideration warranting comment is that this subscore seems to punish severely those therapies that come exclusively from a single laboratory. A therapy that has been extensively but exclusively studied by a single laboratory would actually score 0 in this subscale, even if its neurological efficacy might be demonstrable in numerous studies from the same laboratory. Nonetheless, it is hard to justify a different score for the criteria of "independent" replication. For a treatment that has been extensively published on – albeit always from the same laboratory – the "harshness" in this replication subscale will be mitigated by the other four subscores, where it is likely that such a treatment will pick up more points for fulfilling other criteria (particularly, with respect to the demonstration of clinically meaningful efficacy, and the establishment of a time window of therapeutic efficacy). Conversely, a therapy that has not been extensively studied and has only been reported by a single lab would do poorly throughout, as would be expected.

Example of the scoring of a neuroprotective treatment

As a demonstration of how this scoring system is applied, we take the example of erythropoietin (Table 2). The systematically collected literature on this drug included 19 studies (see Kwon et al., 2010b). All of these studies were done using rat species, with the exception of one study that used a mouse model. Therefore, for the "animal species" subscore, erythropoietin receives a score of 6: 4 for the publication of studies demonstrating behavioral and non-behavioral efficacy in rats, and 2 for the same in mice.

The majority of erythropoietin studies employed a thoracic contusion SCI models, although there were six aneurysmal clip or rod compression models and one model that employed a unilateral thoracic hemisection. No cervical injury models were employed in the study of erythropoietin. Therefore, for the "injury paradigm" subscore, erythropoietin receives a 6: 3 for efficacy in a thoracic contusion model, and 3 for a thoracic clip compression model, While a 1 would also be scored for the study by King and colleagues (2007) using a partial thoracic transection model, the positive results in this study were only in histological measures and no behavioral outcomes were reported, and so it does not satisfy the criteria of demonstrating both behavioral and non-behavioral efficacy.

In terms of the time window of efficacy, in approximately half of the studies, erythropoietin was administered at the time of injury. The delay was 30 to 60 min in their remainder of studies. The study by Gorio and colleagues (2005) evaluated a 24- and 48-h time window, but in these, only a small improvement in BBB was noted for up to 3 weeks post injury, but at the final 4-week post-injury assessment, animals treated with these extended time windows of intervention were not improved over saline controls. The drug did not appear to have efficacy with this extended time window. Therefore, for the "time window of efficacy" subscore, erythropoietin receives a 5: 3 for efficacy demonstrated with a delay of 1 h or more but less than 4 h, and a score of 2 for efficacy with a delay less than 1 h.

For the demonstration of clinically meaningful efficacy, erythropoietin scores very well, in part due to the large number of studies. A dose effect was demonstrated by Kaptanoglu (2004), Gorio (2005), and Kontogeorgakos (2009). The achievement of plantar weight supported stepping and forelimb/hindlimb coordination on the BBB scale was reported. Improvements in other motor behavioral outcomes such as

Table 2. Application of Pre-Clinical Grading System to Erythropoietin

	Author	Animal species	Injury model	Maximum time window of efficacy	Clinically meaningful efficacy: Thoracic SCI model: BBB Scores: Plantar wt support or coordination	Clinically meaningful efficacy: Thoracic SCI model: Other motor and non-motor improvement	Clinically meaningful efficacy: Cervical SCI model: Motor improvement	Clinically meaningful efficacy: Cervical SCI model: Non-motor improvement	Dose response	Reproducibility/ replication
1	Guízar-Sahagún Spinal Cord, 2009	Rats[4]	T9 contusion	0 h						Negative 1
2	Huang J. Int. Med. Res., 2009	Rats	T10 contusion[3]	0 h						Positive 1
3	Kontogeorgakos Arch. Orthop. Trauma Surg., 2009	Rats	T10 Clip[3] compression	0 h					1000 IU/kg[4]2 vs. 13 doses	Positive 2
4	Yazihan Injury, 2009	Rats	T9–11 Clip compression	1 h						Positive 3
5	Fumagalli Eur. J. Pharmacol., 2008	Rats	T9 contusion	30 min[2]	Coordination[4]					Positive 4
6	Mann Exp. Neurol., 2008	Rats	T9/10 Contusion	1 h tested						Negative 2
7	Pinzon Exp. Neurol., 2008	Rats	T3 Clip compression and T9 contusion	0 h, 24 h, or 48 h tested						Negative 3
8	Vitellaro-Zuccarello Neuroscience, 2008	Rats	T9 Contusion	30 min	Plantar WS and coordination					Positive (same as 5)
9	King Eur. J. Neurosci., 2007	Rats	T10/11 Hemisection	30 min						Positive 5
10	Okutan J. Clin. Neurosci., 2007	Rats	T9 Contusion	0 h						Positive 6
11	Vitellaro-Zuccarello Neuroscience, 2007	Rats	T9 Contusion	30 min	Plantar WS and coordination					Positive (same as 5)
12	Arishima Spine, 2006	Rats	T9 Weight compression	15 min						Positive 7
13	Cetin Eur. Spine J., 2006	Rats	T3 Clip compression	40 min		Swimming[4]			1000 IU/kg 1 vs. 3 doses	Positive 8
14	Grasso J. Neurosurg. Spine, 2006	Rats	T3 Clip Compression	0 h	Coordination					Positive (same as 17)
15	Boran Restor. Neurol. Neurosci., 2005	Rats	T6/7 Contusion	1 h		Swimming				Positive 9
16	Gorio PNAS, 2005	Rats	T9 Contusion	30 min					100, 500, 5000 IU/kg	Positive (same as 5)
17	Brines PNAS, 2004	Mice[2]	T3 Compression	0 h		BMS				Positive (same as 5)
18	Kaptanoglu Neurosurg. Rev., 2004	Rats	T8 Contusion	1 h						Positive (same as 10)
19	Gorio PNAS, 2002	Rats	T3 Clip Compresion and T9 contusion	1 h[3]	Coordination				500 vs. 5000 IU/kg	Positive (same as 5)
Scores		4 + 2 = 6	3 + 3 = 6	2 + 3 = 5	4	4			4	12–7 = 5
TOTAL						34				

In the Vitellaro-Zuccarello studies, the EPO-treated animals achieved BBB scores ~15, while the controls were ~8, indicating that the treated animals achieved both weight support and consistent forelimb–hindlimb coordination while the controls did not achieve weight support in stance. Histologic improvements were also reported in both studies.

In the Fumagalli study, the BBB scores were 13.9 versus 9.5 for EPO versus control. Given that 14 is "consistent" forelimb–hindlimb coordination, we felt that the treated animals earned the score of for achieving coordination. In the Gorio 2005 paper, however, the BBB scores were 13 versus 9, and we therefore would not assign this as being "clinically meaningful" as per the BBB criteria because both groups achieved weight support, but the EPO animals did not achieve consistent forelimb–hindlimb coordination.

For reproducibility, the studies in which Gorio and Brines were involved were viewed as one independent lab, hence the positive studies of Fumagalli, Vitellaro-Zuccarello, Gorio, and Brines counted as one. With this, there were nine independent reports demonstrating positive effects of EPO (earning it a score of 12), and three independent reports of negative effects (earning it a score of (7). Hence, the final score is 5.

the swimming test were reported (although these were assessed no later than 2 weeks post injury). Improvements in non-motor outcomes such as mechanical allodynia or autonomic functions have not been demonstrated. This gives erythropoietin a total of 12 out of 20 possible points on the "clinically meaningful efficacy" subscale.

In the final consideration of independent reproducibility and replication, nine independent laboratories reported on the beneficial effects, while three reported negative results (including one attempted formal replication). The nine positive articles gives erythropoietin a score of 12, but the three negative articles subtracts seven points. Therefore, for "independent reproducibility/replication," erythropoeitin receives a score of 5.

The final tally of scores related to each of the above measures gives erythropoietin a total score of 34 $(6+6+5+12+5)$.

To put such a score into perspective, one could consider translating a treatment that was found in a single laboratory to be effective in a rodent model of SCI. Such studies are quite common: a drug administered at the time of a thoracic contusion injury promotes tissue sparing and a significant improvement in BBB scores, with controls getting an 8 and treated animals achieving a 12. If this one study constituted the "body of literature" on the particular treatment, it would get a 4 in "animal species" for demonstrating efficacy in a rodent model, a 3 for "injury paradigms" for demonstrating efficacy in a thoracic contusion model, a 2 for "time window of efficacy" for demonstrating efficacy when given at the time of injury, a 4 for "clinically meaningful efficacy" for the BBB score improvement, and a 0 for "independent reproducibility/replication," given that it was the only independent laboratory reporting the beneficial effects on the therapy. That would give it a final score of 13 $(4+3+2+4+0)$.

Applying the score to systematically reviewed acute neuroprotective therapies

The other method for putting erythropoietin's score of 34 into perspective is to measure it against other systemically administered neuroprotective therapies whose pre-clinical literature was systematically reviewed in the same fashion. Table 3 includes all of the treatments in our systematic review. The therapies are listed in the order of number of published studies that met the criteria for the systematic review, with the exception of "NSAIDS," which we divided into the two NSAIDS most commonly studied: ibuprofen and indomethacin. The individual tables, which include the scoring of each therapy, are included in the Supplementary Data (Supplementary Data are available online at www.liebertonline.com/neu).

At the outset, it is noted that the criteria upon which the scores were based were derived in a systematic, stepwise fashion. First, by using the results of the questionnaire circulated in 2008, we came up with the first version of a scoring system that included the five general subscores: animal species, injury model, time window of efficacy, clinically meaningful efficacy, and independent reproducibility/replication. Then, after utilizing the discussion and voting results of the focus-group meeting held in October 2009, we derived a second version of the scoring system with revised criteria for how each subscore would be specifically graded. This was then sent around to all the authors and, based on comments and suggestions, a final version of the scoring system was generated. These criteria were then applied consistently across all of the neuroprotection therapies. To minimize bias, it was felt that the first step was to decide upon the subscores and the criteria by which they would be scored, based on what the researchers felt was appropriate. We then applied this scoring scheme to all of the treatments uniformly. Looking at the final scores might naturally stimulate further discussions to revise elements of the scale. While this is totally appropriate, we felt that the most "unbiased" way to develop an objective scoring system would be to focus on the criteria, and then once agreement upon the criteria were reached, to apply this to the treatments. Changing the criteria according to the final scores of the therapies introduces the possibility of bias, as researchers who "perceive" that one therapy is better than another can alter the scores by weighting specific categories more heavily than others. Ultimately, we were aiming to develop a scoring system that could be "objectively" applied to the body of literature that exists on a particular therapy.

Table 3. Scores for Neuroprotective Therapies

Therapy	*Animal species*	*Injury model*	*Maximum time window of efficacy*	*Clinically meaningful efficacy*	*Reproducibility/ replication*	*Total*
Erythropoietin	6	6	5	12	5	34
Systemic Hypothermia	4	18	3	12	5	42
Ibuprofen	4	4	10	8	7	33
Indomethacin	10	6	2	8	5	31
Anti-CD11d antibodies	4	3	9	8	0	24
Minocycline	6	8	13	16	5	48
Progesterone	4	3	2	4	0	13
Estrogen	6	6	3	0	9	24
Magnesium	4	6	11	12	12	45
Riluzole	4	9	5	8	7	33
Polyethylene Glycol	10	12	11	4	4	41
Atorvastatin	4	3	10	4	3	24
Inosine	6	1	8	4	7	24
Pioglitazone	4	3	5	12	3	27

Discussion

The scoring system that we devised is an academic perspective on the issue of translation, and we recognize that other considerations, particularly that of the commercial potential of a therapy, have a significant influence on the process of translating an experimental therapy from bench to bedside. The use of such a grading system is meant to provide an objective measure of the "translational potential" of a specific therapy based on a systematically collected set of literature supporting its application in acute SCI. The general elements of the grading system were based on perspectives and opinions provided by over 200 clinicians and scientists in the SCI field in the questionnaire that was distributed in the fall of 2008 (see Kwon et al., 2010a). The results of this questionnaire highlighted strong opinions about the importance of animal species, injury models, time windows of efficacy, demonstrations of efficacy, and independent replication. To some extent, the survey also provided important guidance about what criteria to exclude, such as pre-clinical efficacy in other acute or chronic neurologic conditions (e.g., stroke, TBI, Parkinson's, MS, ALS). We also considered it reasonable to exclude any consideration of safety, as this is ultimately a regulatory issue for all neuroprotective agents. The issue of pain is included as a "non-behavioral" outcome metric as a number of investigators are evaluating therapeutics that might reduce neuropathic pain.

To refine the grading system, we conducted a modified Delphi consensus-building approach at a focus-group meeting in October 2009, which included scientists and surgeon-scientists in the SCI research field. Although both clinicians and scientists are important participants in the dialogue on translating therapies from bench to bedside, many clinicians lack detailed and current knowledge of the animal and injury models utilized in pre-clinical experimental research. The opinions of clinicians should not be discounted, given that they ultimately will have a major role in deciding for which experimental therapies they will choose to recruit patients. However, for the purposes of deciding how to prioritize the significance of specific pre-clinical data elements, the scientist principal investigators who have a more intimate knowledge of the many nuances of such experiments are arguably better equipped. While the overall structure of the scoring system and subscales were based on the opinions voiced by both clinicians and scientists, how the pre-clinical data would be scored and weighted within each subsection was felt to be best determined by those who are very experienced with pre-clinical experimental methodology (particularly as it pertains to animal and injury models). It is for this reason that only scientists or spine surgeon-scientists (all of whom are principle investigators of a scientific research laboratory) were invited to participate in the focus-group meeting, during which the weighting within each subscale was discussed and debated. The weightings that emerged from the meeting were distributed amongst the focus-group attendees, and an example of how the scores were applied to a therapy (erythropoietin) was provided. The final scoring system was then established based on comments and feedback, and this was applied to all of the therapies in the systematic review on neuroprotective treatments (see Kwon et al., 2010b).

A few observations can be made based on scoring of the therapies by the pre-clinical grading scale. First, the more extensively studied treatments with greater numbers of publications do better, provided that they are studied by different labs. This is reflected in the scores for systemic hypothermia and minocycline, for example. The multiple different labs increase the chance that the treatment will gain more points for being studied in different animal species, using different injury models, and with longer delays in intervention being attempted. Additionally, there is a greater chance of gaining points for satisfying different criteria for "clinically meaningful efficacy," and for achieving independent replication. Conversely, a treatment where all of the publications come from a single laboratory scores relatively poorly, not only because of the reproducibility issue, but also because of the homogeneity of the animal species and injury model, which would be expected within a single lab. Studies where the therapy was administered prior to injury fared poorly, because even though they received a token score of 1 on the "time window of efficacy," we did not consider any of their behavioral improvements to represent "clinically meaningful efficacy" because they were attained in a therapeutic paradigm that is clearly inapplicable clinically. And finally, therapies scored poorly if most of their studies reported only non-behavioral improvements without any behavioral outcomes, given the requirement for behavioral efficacy in the animal species, injury model, time window of efficacy, and 16 of 20 points in the clinically meaningful efficacy subscore.

This scoring scheme is a first attempt to generate a quantitative system to evaluate, objectively, pre-clinical evidence for potentially promising experimental treatments for acute SCI. As an academic exercise contributed to by both clinicians and scientists in the SCI field, it is by no means intended to represent a regulatory guidance document describing what must absolutely be done in pre-clinical studies prior to human translation. The scoring system, in essence, attempts to reflect how extensively a particular therapy has been studied, and any given treatment logically accrues points and a higher score as the body of peer-reviewed literature on it incrementally grows. For the subscores of "time window of efficacy" and "independent replication," there is an ordinal progression of the weightings reflecting the simple concept that longer time windows and more independent studies would be more predictive of clinical efficacy. For the other subscores of "animal species," "injury models," and "demonstration of clinically meaningful efficacy," the weightings are assigned categorically, reflecting the value in a therapy being studied in a multitude of settings and therefore accruing credit for the breadth of investigation. This scoring system also provides an objective perspective regarding the "translational readiness" of an SCI therapy that has been demonstrated in one single study from one lab. While such a "perspective" might seem obvious, human translation has proceeded in the past on the basis of this very limited peer-reviewed pre-clinical literature.

While this initiative was inspired to some extent by the leadership of the stroke field in establishing the STAIR guidelines, the end result differs importantly in that we have put forth a quantifiable measure based upon similar perspectives on analogous guidelines, such as the use of different animal models and the need for therapeutic window analysis. We acknowledge that the merits of proposing a quantifiable measure versus simply laying out guidelines could be debated (and indeed were within the members of the focus

group). Given that STAIR-like guidelines do not currently exist for SCI, it would not have been a trivial contribution to simply outline a series of similar guidelines for consideration in pre-clinical SCI research (such as, for example, the desirability for studies in multiple animal models and the need for time-window studies). However, in this initiative, we attempted to achieve a deeper perspective on these important issues by applying some measure of quantification. So, for example, if multiple animal models are considered important, which animal models should they be? Are they all considered equally, or are some more "relevant" than others? Is one injury mechanism (e.g., contusion) as important as another (e.g., laceration)? If everyone believes that time-window studies are important for neuroprotective agents, what time windows of efficacy should be demonstrated? Is efficacy in one time window (e.g., 4 h post injury) viewed the same as another time window (e.g., 1 h)? The grading system attempts not only to identify which types of studies are important in the pre-clinical development of a therapy, but also to provide perspective on how many have actually been done.

Having said this, this method of scoring has a number of inherent limitations, which are important to point out. First, in the rather simplistic way of adding up scores as they are achieved by the publication of further results on a particular therapy, a proposed treatment can potentially achieve a relatively high score even in the absence of some important evidentiary elements. For example, a therapy that is studied by many independent labs but with a time window of efficacy of no longer than 10 min post injury might potentially accrue a lot of points for animal species, injury models, clinically meaningful efficacy, and independent replication, but would be seemingly quite far from being "ready" for human translation without showing efficacy with a more clinically relevant delay in intervention. Conversely, a therapy studied only in one laboratory might, over time, accumulate quite a high score if that lab continued to study it in different animal models, injury models, and time windows of efficacy; but yet, despite a high score, the important issue of independent replication would remain unsatisfied. There are no "must-haves" identified in the scoring system, although one could reasonably argue that there should be some in order to truly justify translation. For example, it could be argued that no matter what the total score is, the failure of a treatment to have its efficacy independently replicated in some manner makes it not ready for human translation. It would be reasonable to suggest that readiness for translation might be assessed more completely by a combination of an overall score and a checklist of required elements, including those identified by the scoring subsections.

On a broader sense, it would be reasonable to question how the opinions of the SCI community (as expressed in the survey of 324 individuals) and the opinions of the SCI experts (as discussed within the focus-group meeting and in subsequent communication) were balanced in the generation of the scoring system. We have attempted to provide the reader with some insights into this process by simply including the content of the discussion at the focus-group level in the text of the manuscript, so as to reveal the dialogue and considerations that surrounded the generation of the scores. The five broad considerations of "animal models," "injury models," "time window of efficacy," "clinically meaningful efficacy," and "independent replication" formed the basis of the scoring system because we actually had hard data on the opinions of clinicians and scientists to serve as the basis for discussion. These data were presented to the group as the "baseline information" and context for subsequent discussion. We then attempted to use the focus group of SCI experts to provide guidance about the specific weightings of evidentiary elements within these five broad considerations. It is acknowledged that in some cases (a good example being the issue of how to weight contusion versus compression injuries) what emerged from the focus-group discussion seems a bit at odds with what the survey results indicated. The nature of this discussion is included in the text of the manuscript, and we accept that others might choose to revisit this issue down the road in subsequent refinements of a pre-clinical grading system. In general, however, while the broad considerations for the scoring system were identified in the questionnaire, the specific weightings were established primarily by the focus group.

Two additional methodological issues warrant discussion. We acknowledge that the process by which this grading system was developed did not strictly adhere to the rigors of a true Delphi consensus-building exercise. We did, however, attempt to maintain important aspects of the Delphi that we felt were particularly relevant to this process, which included providing participants with a common set of data or "knowledge," providing an opportunity for discussion around topics, providing anonymity for voting on specific questions, providing further opportunity for discussion, and then providing opportunity for further anonymous voting. Additionally, we did not insist on establishing a consensus for everything, but rather we felt that it was valuable to also identify areas in which strong consensus was lacking. Such was the case, for example, of the question around contusion versus compression mechanisms of inducing SCI. While the majority of experts favored the contusion injury, this preference did not – strictly speaking – reach the level of "strong consensus" even after much discussion and repeated voting. Hence, the scoring scheme reflected this lack of a strong consensus. Here, as in other aspects of the scoring system, it was not so essential to arrive at the "single best injury model," but rather, to come to some agreement as to whether one could be considered to have more merit than another (and then to have that merit reflected in the weightings). In the case of the injury model (contusion vs. compression), the focus group did not arrive at a consensus that one could be weighted more heavily than another. It could be argued that an entire formal Delphi process could be conducted around this single question, and we acknowledge that if we delved further into each specific question that came up in this initiative, it may have been possible to arrive at some additional areas of consensus. Finally, we also acknowledge that while scientists from Europe and Australia were also invited to the focus-group meeting (and the meeting did include members from Europe and Australia), the scoring system is heavy in its Canadian and American input. While we do not think that this invalidates the end product, opinions and perspectives from a more international group would have certainly been desirable, and we hope that future work to improve upon this first iteration of a pre-clinical grading scale would have more global representation.

While scoring systems and checklists may have some "common sense" academic appeal, the SCI field has previ-

ously moved forward with clinical trials of a number of potential SCI treatments that have not met many of the criteria that are included in this grading scale (e.g., independent replication, demonstrated efficacy in more than one animal or injury model, time-window studies). Clearly, there are also other forces at play that influence the translation of a particular treatment into human patients. The unfortunate reality imposed by the cost of conducting human SCI trials is that an "off-patent" drug with significant pre-clinical substantiation may have less chance of being translated into clinical trials than a patented technology with less animal data but a motivated industry sponsor. The issues of intellectual property and commercialization are beyond the scope of this scoring scheme, which focuses strictly on published scientific data. Nevertheless, the evaluation of the body of literature around a given approach using a rational scoring scheme should be useful to industry, as well as academia. In the commercial development of therapies, considerable non-published pre-clinical data may exist that were generated in the refinement and optimization of the treatment. However, without such data in the peer-reviewed scientific literature, it is impossible to factor them into this scoring scheme.

Perhaps the most important limitation of the scoring system is that by being a measure of the body of literature on a specific treatment, it does not factor in the "quality" of the actual studies that are included in the systematic review. This is an important distinction to make: the scoring system proposed here applies to a given treatment, not to a given manuscript. A scoring system for individual scientific pre-clinical SCI studies that is analogous to the Downs and Black (1998) criteria for assessing the quality of individual clinical studies does not currently exist. Important clinical criteria from the Downs and Black scoring such as the presence of randomization, blinding, and use of controls are typically required methodological elements in pre-clinical studies, and so such criteria would be ineffective at distinguishing "low quality" from "high quality" studies. By focusing on the body of literature and not the quality of the literature, the scoring system has the potential to inflate the "readiness" of a particular therapy, for which studies of relatively modest "quality" might earn the treatment as many points (if not more) as an extremely well-conducted study. As one focus-group member put it, with the failures that we have had thus far, it is hard to envision that simply "more of the same" is going to bring us success in the foreseeable future. Tackling the issue of how to distinguish objectively and quantitatively high- from low-quality studies is a challenge that was not addressed here, but we recognize that the scoring system is ultimately dependent upon the quality of the studies that get included for review.

It should also be noted that the pre-clinical grading system was developed specifically for systemically administered neuroprotective therapies, as it was our *a priori* intention to subject the therapies that were reviewed in a systematic fashion in our previous work (see Kwon et al., 2010b) to whatever form this grading system took after the Delphi process. The grading system would need some modification to be relevant for other therapies such as cell transplantation treatments and directly applied biological therapies. While the overall structure could remain the same (given that animal species, injury paradigms, time windows, independent replication, and clinically meaningful efficacy are still important considerations), the scoring within each subsection would need to be changed in order to be relevant to cell transplant treatments or directly applied biologic therapies. Different time windows for demonstrating efficacy, for example, would need to be defined. The weightings of partial transections versus blunt injuries might be reconsidered (particularly for treatments that intend to promote sprouting/plasticity). We are undertaking such initiatives to define scales that would be relevant to these different types of therapies.

The history of human translation of experimental therapies in SCI has not been very long, nor has it witnessed the successes that can guide the translation of other therapies. In the future, the demonstration of an efficacious treatment for human SCI will be extremely helpful, as it may provide some guidance regarding which pre-clinical steps are most important. Such a demonstration will likely also change how we view certain aspects of pre-clinical evidence, and in turn require recalibration of the scoring system. Until that happens, we are dependent upon the perspectives and opinions of the members of our scientific community. The subscores that were selected (animal species, injury paradigms, time window, demonstration of clinically meaningful efficacy, and reproducibility) were based upon the questionnaire results. How they are scored and weighted could certainly be the subject of further debate, but we felt that the experience of the attendees of the focus-group meeting were extremely helpful in providing an expert opinion that could guide this process. This represents a first attempt to provide a structure for how to assess objectively the body of pre-clinical evidence for a potentially "promising" SCI therapy. We acknowledge that it is an imperfect, unproven system, and the methodology of its establishment – while systematic – is somewhat removed from a formal Delphi process. Nonetheless, we feel that it does capture the perspectives of a large body of SCI researchers, in addition to the opinion of a panel of recognized SCI experts. We will welcome further dialogue on this important issue and encourage further improvements and refinements of this grading system, given the compelling need to establish truly effective therapies for individuals who suffer this injury.

Acknowledgments

Support for the focus-group meeting was generously provided by the Rick Hansen Institute (formerly the Canadian SCI-Solutions Network). BKK holds a CIHR New Investigator Award. WT is the Rick Hansen Man In Motion Chair in Spinal Cord Injury Research).

Author Disclosure Statement

No competing financial interests exist.

References

Arishima, Y., Setoguchi, T., Yamaura, I., Yone, K., and Komiya, S. (2006). Preventive effect of erythropoietin on spinal cord cell apoptosis following acute traumatic injury in rats. Spine 31, 2432–2438.

Boran, B.O., Colak, A., and Kutlay, M. (2005) Erythropoietin enhances neurological recovery after experimental spinal cord injury. Restor. Neurol. Neuroscience 23, 341–345.

Brines, M., Grasso, G., Fiordaliso, F., Sfacteria, A., Ghezzi, P., Fratelli, M., Latini, R., Xie, Q.W., Smart, J., Su-Rick, C.J., Pobre, E., Diaz, D., Gomez, D., Hand, C., Coleman, T., and Cerami,

A. (2004). Erythropoietin mediates tissue protection through an erythropoietin and common beta-subunit heteroreceptor. Proc. Natl. Acad. Sci. USA 101, 14907–14912.

Cetin, A., Nas, K., Buyukbayram, H., Ceviz, A., and Olmez, G. (2006). The effects of systemically administered methylprednisolone and recombinant human erythropoietin after acute spinal cord compressive injury in rats. Eur. Spine J. 15, 1539–1544.

Dietrich, W.D. (2003). Confirming an experimental therapy prior to transfer to humans: What is the ideal? J. Rehabil. Res. Dev. 40, 63–69.

Downs, S.H., and Black, N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. J. Epidemiol. Community Health 52, 377–384.

Fawcett, J.W., Curt, A., Steeves, J.D., Coleman, W.P., Tuszynski, M.H., Lammertse, D., Bartlett, P.F., Blight, A.R., Dietz, V., Ditunno, J., Dobkin, B.H., Havton, L.A., Ellaway, P.H., Fehlings, M.G., Privat, A., Grossman, R., Guest, J.D., Kleitman, N., Nakamura, M., Gaviria, M., and Short, D. (2007). Guidelines for the conduct of clinical trials for spinal cord injury as developed by the ICCP panel: Spontaneous recovery after spinal cord injury and statistical power needed for therapeutic clinical trials. Spinal Cord 45, 190–205.

Fisher, M., Feuerstein, G., Howells, D.W., Hurn, P.D., Kent, T.A., Savitz, S.I., and Lo, E.H. (2009). Update of the stroke therapy academic industry roundtable preclinical recommendations. Stroke 40, 2244–2250.

Fisher, M., Hanley, D.F., Howard, G., Jauch, E.C., and Warach, S. (2007). Recommendations from the STAIR V meeting on acute stroke trials, technology and outcomes. Stroke 38, 245–248.

Fumagalli F., Madaschi L., Brenna P., Caffino L., Marfia G., Di Giulio A.M., Racagni G., and Gorio A. (2008). Single exposure to erythropoietin modulates Nerve Growth Factor expression in the spinal cord following traumatic injury: comparison with methylprednisolone. Eur. J. Pharmacol. 578, 19–27.

Geisler, F.H., Coleman, W.P., Grieco, G., and Poonian, D. (2001a). Recruitment and early treatment in a multicenter study of acute spinal cord injury. Spine 26, S58–S67.

Geisler, F.H., Coleman, W.P., Grieco, G., and Poonian, D. (2001b). The Sygen multicenter acute spinal cord injury study. Spine 26, S87–S98.

Gorio, A., Gokmen, N., Erbayraktar, S., Yilmaz, O., Madaschi, L., Cichetti, C., Di Giulio, A.M., Vardar, E., Cerami, A., and Brines, M. (2002). Recombinant human erythropoietin counteracts secondary injury and markedly enhances neurological recovery from experimental spinal cord trauma. Proc. Natl. Acad. Sci. USA 99, 9450–9455.

Gorio, A., Madaschi, L., Di, S.B., Carelli, S., Di Giulio, A.M., De, B.S., Coleman, T., Cerami, A., and Brines, M. (2005). Methylprednisolone neutralizes the beneficial effects of erythropoietin in experimental spinal cord injury. Proc. Natl. Acad. Sci. USA 102, 16379–16384.

Grasso, G., Sfacteria, A., Erbayraktar, S., Passalacqua, M., Meli, F., Gokmen, N., Yilmaz, O., La, T.D., Buemi, M., Iacopino, D.G., Coleman, T., Cerami, A., Brines, M., and Tomasello, F. (2006). Amelioration of spinal cord compressive injury by pharmacological preconditioning with erythropoietin and a nonerythropoietic erythropoietin derivative. J Neurosurg. Spine 4, 310–318.

Guìzar-Sahaún G., Rodriguez-Balderas C.A., Franco-Bourland R.E., Martinez-Cruz A., Grijalva I., Ibarra A., and Madrezo I. (2009). Lack of neuroprotection with pharmacological pretreatment in a paradigm for anticipated spinal cord lesions. Spinal Cord. 47, 156–160.

Hawryluk, G.W., Rowland, J., Kwon, B.K., and Fehlings, M.G. (2008). Protection and repair of the injured spinal cord: A review of completed, ongoing, and planned clinical trials for acute spinal cord injury. Neurosurg. Focus 25, E2–E13.

Huang H., Fan S., Ji X., Zhang Y., Bao F., and Zhang G. (2009). Recombinant human erythropoietin protects against experimental spinal cord trauma injury by regulating expression of the proteins MKP-1 and p-ERK. J. Int. Med. Res. 37, 511–519.

Kang C.E., Poon P.C., Tator C.H., and Shoichet M.S. (2009). A new paradigm for local and sustained release of therapeutic molecules to the injured spinal cord for neuroprotection and tissue repair, Tissue Eng. Part A 15, 595–604.

Kaptanoglu, E., Solaroglu, I., Okutan, O., Surucu, H.S., Akbiyik, F., and Beskonakli, E. (2004). Erythropoietin exerts neuroprotection after acute spinal cord injury in rats: effect on lipid peroxidation and early ultrastructural findings. Neurosurg. Rev. 27, 113–120.

King, V.R., Averill, S.A., Hewazy, D., Priestley, J.V., Torup, L., and Michael-Titus, A.T. (2007). Erythropoietin and carbamylated erythropoietin are neuroprotective following spinal cord hemisection in the rat. Eur. J. Neurosci. 26, 90–100.

Kontogeorgakos V.A., Voulgaris S., Korompilias A.V., Vekris M., Polyzoidis K.S., Bourantas K., and Beris A.E. (2009). The efficacy of erythropoietin on acute spinal cord injury. An experimental study on a rat model. Arch. Orthop. Trauma Surg. 129, 189–194.

Kwon, B.K., Roy, J., Lee, J.H., Okon, E.B., Zhang, H., Marx, J.C., and Kindy, M.S. (2009). Magnesium Chloride in a polyetheline glycol formulation as a neuroprotective therapy for acute spinal cord injury: Preclinical refinement and optimization. J. Neurotrauma 26, 1379–1393.

Kwon, B.K., Hillyer, J., and Tetzlaff, W. (2010a). Translational research in spinal cord injury: A survey of opinion from the SCI community. J. Neurotrauma 27, 21–33.

Kwon, B.K., Okon, E.B., Hillyer, J., Mann C.M. Baptiste D., Weaver L.C., Fehlings M.G., and Tetzlaff, W. (2010b). A systematic review of non-invasive pharmacologic neuroprotective treatments for acute spinal cord injury. J. Neurotrauma, Epub ahead of print.

Lammertse, D., Tuszynski, M.H., Steeves, J.D., Curt, A., Fawcett, J.W., Rask, C., Ditunno, J.F., Fehlings, M.G., Guest, J.D., Ellaway, P.H., Kleitman, N., Blight, A.R., Dobkin, B.H., Grossman, R., Katoh, H., Privat, A., and Kalichman, M. (2007). Guidelines for the conduct of clinical trials for spinal cord injury as developed by the ICCP panel: Clinical trial design. Spinal Cord 45, 232–242.

Mann, C., Lee, J.H., Liu, J., Stammers, A.M., Sohn, H.M., Tetzlaff, W., and Kwon, B.K. (2008). Delayed treatment of spinal cord injury with erythropoietin or darbepoetin–a lack of neuroprotective efficacy in a contusion model of cord injury. Exp. Neurol. 211, 34–40.

O'Collins, V.E., Macleod, M.R., Donnan, G.A., Horky, L.L., van der Worp, B.H., and Howells, D.W. (2006). 1,026 experimental treatments in acute stroke. Ann. Neurol. 59, 467–477.

Okutan, O., Solaroglu, I., Beskonakli, E., and Taskin, Y. (2007). Recombinant human erythropoietin decreases myeloperoxidase and caspase-3 activity and improves early functional results after spinal cord injury in rats. J. Clin. Neurosci. 14, 364–368.

Özdemir, M., Cengiz, S.L., Gurbilek, M., Ogun, T.C., and Ustun, M.E. (2005). Effects of Magnesium sulfate on spinal cord tissue lactate and malondialdehyde levels after spinal cord trauma. Magnes. Res. 18, 170–174.

Pinzon, A., Marcillo, A., Pabon, D., Bramlett, H.M., Bunge, M.B., and Dietrich, W.D. (2008). A re-assessment of erythropoietin

as a neuroprotective agent following rat spinal cord compression or contusion injury. Exp. Neurol. 213, 129–136.

Rowland, J.W., Hawryluk, G.W., Kwon, B., and Fehlings, M.G. (2008). Current status of acute spinal cord injury pathophysiology and emerging therapies: Promise on the horizon. Neurosurg. Focus 25, E2–E13.

Tator, C.H. (2006). Review of treatment trials in human spinal cord injury: Issues, difficulties, and recommendations. Neurosurgery 59, 957–982.

Vitellaro-Zuccarello, L., Mazzetti, S., Madaschi, L., Bosisio, P., Fontana, E., Gorio, A., and De, B.S. (2008). Chronic erythropoietin-mediated effects on the expression of astrocyte markers in a rat model of contusive spinal cord injury. Neuroscience 151, 452–466.

Vitellaro-Zuccarello, L., Mazzetti, S., Madaschi, L., Bosisio, P., Gorio, A., and De, B.S. (2007). Erythropoietin-mediated preservation of the white matter in rat spinal cord injury. Neuroscience 144, 865–877.

Yazihan N., Uzuner K., Salman B., Vural M., Koken T., and Arslantas A. (2008). Erythropoietin improves oxidative stress following spinal cord trauma in rats. Injury 39, 1408–1413.

Address correspondence to:
Brian K. Kwon, M.D., Ph.D., FRCSC
Combined Neurosurgical and Orthopaedic Spine Program
Department of Orthopaedics, University of British Columbia
Room 6196, Blusson Spinal Cord Center, VGH
818 West 10th Avenue
Vancouver, British Columbia V5Z 1M9
Canada

E-mail: brian.kwon@vch.ca