Systematic Review of Randomized Controlled Trials of Clinical Prediction Rules for Physical Therapy in Low Back Pain

Shilpa Patel, C. Psychol, Tim Friede, PhD, Robert Froud, PhD, David W. Evans, PhD, and Martin Underwood, MD

**Study Design.** Systematic review.

**Objective.** To evaluate randomized controlled trials validating the effects of a clinical prediction rule for patients with non-specific low back pain (LBP). The outcomes of interest were any back pain or pain-related measures.

**Summary of Background Data.** LBP is a common and costly condition. Interventions for back pain seem to have, at best, small to moderate mean beneficial effects. Identifying subgroups of patients who may respond better to certain treatments may help to improve clinical outcomes in back pain. The development of clinical prediction rules is an attempt to determine who will respond best to certain treatments.


**Results.** We identified 1821 potential citations; 3 articles were included. The results from the available data do not support the use of clinical prediction rules in the management of non-specific LBP.

**Conclusion.** There is a lack of good quality randomized controlled trials validating the effects of a clinical prediction rule for LBP. Furthermore, there is no agreement on appropriate methodology for the validation and impact analysis. The evidence for, and development of, the existing prediction rules is generally weak.

**Key words:** clinical prediction rules, back pain, subgroups, predictors of response to treatment, systematic review. Spine 2013;38:762–769

**Level of Evidence:** 1

Low back pain (LBP) is a common and costly condition. Effective interventions for LBP have, at best, small to moderate effects, when averaged over populations. The true potential value of these interventions might have been underestimated because most trials encompass all people with non-specific LBP as a single group, assuming homogeneity.

The identification of subgroups is an important research priority. There is growing interest in the development and use of clinical prediction rules (CPRs) in the physical therapy literature, where focus has been on using such rules to determine who will best respond to a given intervention. It is likely that outcomes will be improved if subgroups of patients with LBP could be identified and better matched to treatment.

Clinical prediction rules are defined as:

... the process by which combinations of clinical findings that have been statistically demonstrated to be meaningful predictors of a condition or outcome of interest are used to categorize a heterogeneous group of patients into subgroups based on a shared likelihood of the presence of that condition or outcome.

CPRs can be useful in determining prognosis, assessing the likelihood of the presence or absence of a condition, and to help classify patients into groups more likely to benefit from treatment. Strictly speaking the latter is not covered by the definition given above, but for the purpose of this review we take a wider view. They can help with screening patients to decide when further investigations are likely or unlikely to yield meaningful findings.
Developing a CPR should be a 3-step process of derivation, validation, and impact analysis. During the derivation stage, it is important to identify known individual predictors of response to treatment. For validation, external validity should be examined by applying the rule to different settings, patients, and clinicians before generalizing. Unfortunately, validation studies are rarely reported in the literature. If these tools are well-designed and validated in appropriate populations, they will have the potential to identify those patients most likely to benefit from a particular treatment, which in turn will help improve clinical decision making and practice. There is currently considerable interest in reviewing the literature on CPRs for back pain. This largely focuses on synthesizing the outcomes of the original studies. If, however, the underpinning research methods used are not robust, little weight can be given to the conclusions. Here, we report a systematic review evaluating validation studies of CPRs for patients with LBP using randomized controlled trials (RCTs). Our outcomes of interest were any back pain or pain-related measures. We provide a methodological and statistical critique of the included articles, different to articles already published in this area.

MATERIALS AND METHODS

Inclusion Criteria
We included RCTs that validated the effects of a CPR. We defined a CPR as any clinical tool with various components drawn from the history, examination, and laboratory tests used to inform treatment choices. We included studies of non-specific LBP, that is, we excluded studies concerned with diagnosis and management of malignancy, infection, fracture, or inflammatory disorders (e.g., ankylosing spondylitis). Trials were of interventions to treat LBP of any duration with outcomes of pain, disability, and psychological distress. The age of participants in trials was restricted to adults (18 yr or older). Only articles published in English were included. Studies described as RCTs that did not allow the performance of the CPR to be assessed using random allocation were excluded. We excluded studies of surgical techniques and derivation studies (studies prospectively examining the predictive ability of selected variables for a CPR).

Search Strategy
We searched the following electronic databases from 1980 to 2009: MEDLINE, EMBASE, PsycINFO, AMED, PubMed, ISI Web of Knowledge, and the Cochrane Library. We used MeSH (Medical Subject Heading) terms to identify articles containing material relevant to “back pain” and “low back pain.” We combined these using Boolean operators with keyword terms “subgroup” or “subgroups” or “classification” or “criteria” or “diagnostic criteria” or “rule” or “rules” or “decision rules” or “prediction rule” or “clinical predictors.” Databases were limited to RCTs. We also searched reference lists of relevant articles for further citations. The validation studies included in a recent review of CPRs for the management of LBP were assessed for inclusion in this review.

Inclusion Process
Two reviewers independently reviewed titles and abstracts of citations identified from the electronic searches (Figure 1). We found that the inter-observer reliability for screening titles and abstracts was poor; we therefore used a third reviewer to screen all titles and abstracts. The third reviewer screened the titles and abstracts and developed a list of the potentially included articles. These were compared with those selected by reviewers 1 and 2. Those that matched were included and those that did not were discussed by all 3 reviewers to reach a consensus. The agreed upon full articles were obtained and reviewed by 2 reviewers, SP and RF independently extracted data and assessed the quality of each article. We extracted data on the derivation and composition of the CPR tested, the treatments being compared, and the study design (Table 1). All disagreements over data extraction and quality assessment were resolved by means of discussion.

Quality Assessment
We used the Pincus quality assessment tool for studies of treatment moderators. Each study was assessed for quality independently by 2 reviewers. Studies that met all 5 criteria were classified as providing confirmatory evidence, those complying with criteria 3, 4, and 5 were classified as providing...
exploratory evidence. All other studies were classified as providing insufficient evidence (Table 2).

RESULTS
We identified 1821 citations for possible inclusion from bibliographic searches. We obtained 33 full texts for detailed inspection. Following examination and discussion, 32 articles were excluded, 27 articles were not validating the effects of a prediction rule, 3 presented results from mixed samples, 1 study was not randomized, and 1 study recruited some participants under the age of 18 years. Subsequently, 3 studies were included in this systematic review (Table 1). Statistical pooling of the data was not performed because of the small number of studies found as well as heterogeneity in the interventions applied, outcome measures collected, and populations from which samples were drawn.

Brennan et al reported a 3-armed trial of manipulation, stabilization, and exercise. The CPR used was on the basis of work by Fritz et al. They examined the inter-rater reliability of individual examination items for a classification decision-making algorithm. At baseline history and physical examination, data were collected before randomizing patients to one of the 3 treatments. After completion of the study, the baseline signs and symptoms were used by 2 physical therapists to decide which subgroup the patient fitted and a third therapist was consulted if agreement could not be reached. Patients were then classified as “matched” (if they received the treatment that matched their subgroup classification) or “unmatched” (if they received a treatment different to their subgroup classification). The results were analyzed using a 3-way interaction between randomized treatment, classification, and time that was found to be significant, whereas both the 2-way interaction between randomized treatment and time and that between classification subgroup and time were not statistically significant. The authors reported greater change in the Oswestry Disability Questionnaire for matched subjects than unmatched both at 4-week and 1-year follow-up.

Childs et al used a CPR in a 2-arm trial of spinal manipulation and exercise. The rule had previously been developed in a prospective cohort study with a small sample of participants from an army medical center. The prediction rule identified patients with LBP who had a good prognosis when treated with spinal manipulation, based on 5 clinical factors. Data on these 5 factors were collected at baseline by a physical therapist blind to the participant’s treatment group. Those participants who met 4 or more of the 5 criteria were classified as positive and therefore likely to respond to manipulation. Those classified with 3 or fewer positive criteria were classified as negative. The authors tested for a 3-way interaction between patients’ status on the rule, treatment group, and time. The authors concluded that those positive on the prediction rule that received a form of lumbar sacral spinal manipulation had the greatest treatment benefit at 4- and 6-month follow-up.

Hancock et al evaluated the spinal manipulation rule used by Childs et al to assess generalizability of setting and sample of patients receiving spinal manipulation. They carried out a preplanned analysis of data from an attention-controlled trial of spinal manipulation therapy. After collection of baseline data, participants were randomized to one of 4 groups. Participants were assessed on the 5 criteria to determine status on the prediction rule. A researcher who was blinded to allocation collected data on 2 of the 5 criteria, the treating physiotherapist collected data on the remaining 3 criteria. The authors conducted a 3-way interaction between a patient’s status on the rule, treatment group, and time, which was not statistically significant at the 5% level. The authors concluded no clinically worthwhile interaction effects between treatment group and status on the rule for either pain or disability at any time point.

Critique of Included Articles
Brennan et al adopted a pragmatic approach to treatment progression, patients moved to the second subacute stage if they achieved a predefined reduction in their Oswestry Disability Index score. For these patients the therapist could select only those treatments permitted based on the patient’s original treatment group. Therapists were permitted to use their own clinical judgment to determine exercise dosage for individual patients. Although this is more representative of normal clinical practice, it is difficult to determine confounding factors that may have an effect on response to treatment. Statistically, comparing those that are classified as matched with those that are unmatched is likely to produce larger effects than if comparing those randomized with the different treatments because some participants will receive the correct treatment by chance. A closer inspection reveals that the statistical analysis used by Brennan seems not to have included a test for an interaction between all subgroups and treatment, making it difficult to conclude which subgroup of patients would respond best to which treatment. In addition, quality assessment of this article provided inconclusive evidence.

In the analysis conducted by Childs et al, the authors do test for a 3-way interaction between CPR, treatment group, and time and concluded a positive effective for the CPR. However, the CPR has been criticized as comprising of items that would ordinarily be associated with a favorable prognosis, which challenges the usefulness of such a rule in clinical practice. Despite their positive result we think that the methods used to develop this rule are not robust and the items included in the rule select those that would get better ordinarily. The quality assessment of this article provided inconclusive evidence.

Hancock et al, when independently assessing the findings reported by Childs et al, found no significant interactions between the treatment group and status on the rule for either pain or disability. They found that a positive status on the rule predicted better prognosis regardless of treatment received for pain and disability at 2 and 12 weeks. Importantly, because Hancock et al did not find a difference between intervention and control treatments, it would be unlikely that there would be an important interaction between CPR and treatment group unless it was hypothesized the intervention was harmful for at least some individuals. These results do not
<table>
<thead>
<tr>
<th>Author</th>
<th>Setting and Inclusion Criteria</th>
<th>Sample Size</th>
<th>Baseline Data</th>
<th>Treatments Being Compared</th>
<th>Details of Clinical Prediction Rule</th>
<th>Outcome Measure and follow-up</th>
<th>Authors Results and Conclusions</th>
<th>Quality Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brennan et al</td>
<td>Clinics in Utah. Patients aged 18 to 65 yr, LBP of &lt;90 d, with or without referral into lower extremity, and an OSW score of ≥25%</td>
<td>123</td>
<td>History and standardized physical examination. Demographic data, current pain intensity, FABQ, modified OSW</td>
<td>Manipulation vs. specific exercise vs. stabilization</td>
<td>Retrospectively baseline data was used to classify them into one of 3 groups: manipulation, specific exercise, or stabilization (classification was based on previous work by Fritz et al [17]) by 2 therapists blind to the treatment group assignment. If agreement was not reached a third therapist was consulted</td>
<td>OSW at 4 wk after baseline assessment. Long-term follow-up at 1 yr</td>
<td>3-way interaction between randomized treatment, classification and time significant. Greater change in OSW in matched than unmatched for both short- and long-term follow-up. 2-way interactions not significant. 4 wk treatment effect 0.4</td>
<td>Inconclusive evidence</td>
</tr>
<tr>
<td>Childs et al</td>
<td>Clinics mainly with facilities within the US Air Force. Patients aged 18 to 60 yr, LBP with or without referral into lower extremity, and an OSW score of at least 30%</td>
<td>131</td>
<td>History and standardized physical examination. Demographic data, body diagram to assess symptoms, current pain intensity, and best and worst pain in last 24 hr, FABQ, modified OSW</td>
<td>Spinal manipulation plus an exercise program vs. exercise program alone</td>
<td>Patients were examined on the basis of the 5 criteria for spinal manipulation (Flynn et al [17]) after randomization, therapists were not instructed in the rules criteria and were unaware of the patients’ status on the rule. After completion of the trial an examiner blinded to the patients treatment assignment determined patients status on the rule using baseline data (positive = 4/5 criteria met, negative = 3 or fewer)</td>
<td>OSW at 1 wk. A therapist blinded to the patients’ status on the rule repeatedly the history and physical examination 1 and 4 wk after randomization. 6-mo follow-up questionnaire</td>
<td>3-way interaction between status on rule, treatment group, and time significant. Patients positive on the rule that received spinal manipulation had great improvement in pain and disability at 1 and 4 wk and 6-mo follow-up than positive on the rule and got exercise or negative on the rule and received spinal manipulation</td>
<td>Inconclusive evidence</td>
</tr>
<tr>
<td>Hancock et al</td>
<td>Primary care. Patients with LBP of &lt;6 wk, pain between the 12th rib and buttock crease causing moderate pain and moderate disability</td>
<td>239</td>
<td>Demographics, history, RMDQ, FABQ, catastrophizing, self-statement, coping self-statement, numerical pain rating scale, patient specific functional scale.</td>
<td>PL SMT &amp; PL diclofenac vs PL SMT &amp; AT diclofenac vs AT SMT &amp; PL diclofenac vs AT SMT &amp; AT diclofenac 4-arm trial</td>
<td>Classification based on Childs et al [14] spinal manipulation clinical prediction rule initially developed by Flynn et al [17]</td>
<td>Pain (11 point scale) and disability (RMDQ) at 1, 2, 4, and 12 wk</td>
<td>3-way interaction between treatment group, clinical prediction rule status and time not significant for pain or disability. No significant interactions between treatment group and status on rule for either pain or disability. Positive status on rule predicted better prognosis regardless of treatment received for pain and disability at 2 and 12 wk</td>
<td>Providing exploratory evidence</td>
</tr>
</tbody>
</table>
support the wider use of the Childs rule and we agree that this conclusion is supported by the analysis of results from the Hancock trial.

The differences in findings between Hancock et al. and Childs et al. can be attributed to a variation on the type and application of spinal manipulation, the setting, and the patient population. More specifically, the spinal manipulation delivered in both studies was not the same. Hancock et al. reported baseline pain and disability scores between the spinal manipulation groups and placebo group, and these were small and not statistically significant. Therefore, testing for an interaction effect may be misleading. This was the only trial providing exploratory evidence.

DISCUSSION

In this review, we focused on evaluating validation studies of CPRs for LBP using RCTs. We have therefore not included studies looking specifically at the derivation of rules. Two systematic reviews, published after we had completed our review, support our conclusions. The first, examined the validity of CPRs and concluded that derivation studies were mainly of high quality, whereas the cross-sectional validation studies were weak, limiting application of the rules in clinical practice. The second, a more recent review of CPRs for the management of LBP conclude the current evidence does not support the clinical application of these rules. Our findings add to this work by demonstrating that the evidence from randomized trials validating CPRs for non-specific LBP is weak. Only 3 articles met our inclusion criteria, 2 of which were small studies with a total sample of less than 150 patients. In all cases, the prediction rule had been developed in small selected populations compromising external validity. Even the largest of these studies (Hancock et al., n = 239) is likely to be too small to validate a CPR adequately, and thus their negative findings are unlikely to be robust due to the possibility of type II error; indeed one of the challenges to performing interaction tests is having adequate statistical power.

After completing the work for this review, the results of the STarT back trial, a well conducted RCT of subgrouping and targeted treatment for patients with LBP, which achieved a positive result, were published. This trial would not have met our inclusion criteria as the actual subgrouping tool was not being tested but rather the targeted treatments were being tested. To test the targeting tool the trial design would need to randomize patients to receive the “tool” versus “no tool,” with those receiving the tool being allocated to treatment and those without being randomly assigned to treatment. This trial design would allow for the statistical comparison of tool versus no tool, giving a better indication of the tools ability to subgroup patients to targeted treatments.

A systematic review by Haskins et al. excluded the article by Brennan et al., in which a classification approach was used. We have included this article in our review as we have used a slightly different definition of CPRs that allows trials using classification and categorization to be included. We have excluded the article by Cleland et al. which was included in this recent review because the authors only included patients

TABLE 1. (Continued)

<table>
<thead>
<tr>
<th>Author</th>
<th>Setting and Inclusion Criteria</th>
<th>Sample Size</th>
<th>Baseline Data</th>
<th>Treatments Being Compared</th>
<th>Details of Clinical Prediction Rule</th>
<th>Outcome Measure and Follow-up</th>
<th>Authors Results and Conclusions</th>
<th>Quality Assessment (+ or -) on the rule. Researcher blinded to treatment allocation collected data on 2 items (FABQ work subscale and duration of current episode). Treating therapist collected data on the remaining 3 criteria in the initial assessment.</th>
<th>LBP indicates low back pain; FABQ, Fear Avoidance Belief Questionnaire; OSW, Oswestry; RMDQ, Roland and Morris Disability Questionnaire; PL, placebo; SMT, spinal manipulative therapy; AT, Active.</th>
</tr>
</thead>
</table>

LBP indicates low back pain; FABQ, Fear Avoidance Belief Questionnaire; OSW, Oswestry; RMDQ, Roland and Morris Disability Questionnaire; PL, placebo; SMT, spinal manipulative therapy; AT, Active.
positive on the CPR, therefore not allowing for the rule to be validated appropriately.

Kent et al. conclude that treatments targeted to subgroups may be effective; however, the results of the studies need to be interpreted with caution. In their review they include a trial by Long et al., which we have excluded from our review because the authors include patients only with a directional preference and exclude those without, therefore not allowing the clinical prediction rule to be tested appropriately. The discussions within the review by Kent et al. focus on the effect sizes of the included trials and significance of this. In our article we focus on the authors ability to test a clinical prediction rule effectively in a RCT. We present a methodological and statistical critique, different than that presented by Kent et al.

The development of the CPRs tested in these studies lacked methodological rigor. It remains unclear as to which candidate domains should be included in a prediction rule for LBP. It seems that identification of such domains, or indeed concluding that such domains cannot be identified will require a substantial study with an a priori design. To develop a rule that enables clinicians to choose between interventions, the developmental work of derivation and validation needs to be thoroughly designed and systematically validated.

The methodology for quality assessing studies of CPRs is poorly developed. We used a tool developed for a different study of moderators. Based on these criteria only one of the included studies provides exploratory evidence, none fall within the remit of confirmatory evidence (Table 2). The focus of this review was to critique the statistical methods used for the testing of a CPR. Therefore, we did not provide any extensive data on the quality of the underpinning trial. If we had found an apparently robust evaluation of a CPR on which changes in clinical practice might be based, then it would be very important to know the quality of the trials by conventional measures.

At present, we do not know if the disappointing performance of CPRs in RCTs is because inappropriate rules have been tested, the trials have been poorly designed, underpowered, or indeed whether it is impossible to develop CPRs that are fit for this purpose. Although most of the derivation studies have been of high quality, our review, as well as previous reviews, have found the quality of validation studies to have been poor.

In a perspective article, Hancock et al. make a number of useful suggestions for future work on predictors of response to treatment which could also be applied to validation of a CPR. Alternatively, a CPR could be validated as the intervention in a RCT. Any such trials are likely to need very large numbers of subjects. The effect size for main treatment effects in current positive LBP pain trials are typically small to moderate. Any interaction effect is unlikely to be greater than the main treatment effect; indeed if it was, it would suggest that for a substantial group of people the intervention was likely to have no positive effect or for an identifiable group to make their condition worse. Thus, as a rule of thumb, trials at least 4 times the size of current large trials of LBP treatments; perhaps 2000 to 3000 participants will be needed to validate CPRs satisfactorily.

Ideally, it would be better to have CPRs that can be applied to a varied range of clinical decisions, but this is unlikely. CPRs would be useful in LBP as the treatment effects currently seen in populations remain small; this is probably, at least in part, due to the heterogeneity of the un-subgrouped back pain population. In the physical therapy literature there are currently 10 prediction rule derivation studies that have not been validated. Notwithstanding this, authors continue to cite the effective management of LBP using CPR.

There is a need for future well-designed validation studies of these rules to enable better matching of patients to treatment that in turn may lead to better patient outcomes and less health care usage and thus cost to the health care system. The task of developing, validating, and testing such CPRs should not be underestimated. It is difficult to justify the very large cost entailed in taking one CPR through full testing to inform just one treatment choice, therefore our research efforts and our funders’ resources may be better directed in alternative directions. However, we warn against the application of CPRs without sufficient evidence because patients in need of treatment may be denied treatments that they might have benefited from.

Table 2. Methodological Quality of Included Articles

<table>
<thead>
<tr>
<th>Question</th>
<th>Brennan</th>
<th>Childs</th>
<th>Hancock</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Was the subgroup analysis specified a priori?</td>
<td>Yes</td>
<td>Yes</td>
<td>Preplanned secondary analysis</td>
</tr>
<tr>
<td>2. Was the selection of subgroup factors for analysis theory/evidence driven?</td>
<td>Yes—based on work by Fritz et al.</td>
<td>Yes—based on work by Flynn et al.</td>
<td>Yes—Independent evaluation of work by Childs et al. whose rule was based on Flynn et al.</td>
</tr>
<tr>
<td>3. Were subgroup factors measured prior to randomization?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>4. Were subgroup factors measured by adequate (reliable and valid) measurements, appropriate for the target population?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>5. Does the analysis contain an explicit test of the interaction between moderator and treatment?</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Key Points

- Identifying subgroups of patients with LBP who can be better matched to treatments might help to improve clinical outcomes.
- The existing evidence for CPRs is weak. Despite this, prediction rules are still being promoted in the physical therapy literature.
- There is a need for well-designed validation studies of CPRs to enable progression to clinical implementation.

Acknowledgment

The authors thank Professor Sallie Lamb for her contribution in the early stage of this review.

References


