Solutions to methodological problems in rehabilitation research?

Bouter, L.M.; van der Windt-Mens, D.A.W.M.

published in
Biomechanical aspects of manual wheelchair propulsion: the state of the art II
1999

document version
Publisher's PDF, also known as Version of record

Biomedical Aspects of Manual Wheelchair Propulsion
L.H.V. van der Woude et al. (Eds.)
IOS Press, 1999

Solutions to methodological problems in rehabilitation research?

L.M. Bouter and D.A.W.M. van der Windt
Institute for Research in Extramural Medicine, Faculty of Medicine, Vrije Universiteit, Amsterdam, The Netherlands

1. Introduction

This paper will focus on methodological problems in the field of rehabilitation medicine from the perspective of evidence-based medicine (EBM) [1]. Firstly, we will discuss what type of evidence we mean when we talk about EBM and how to use it. Secondly, some examples will be given of variations among systematic reviews and clinical guidelines that are supposed to form the fundaments of EBM. Thirdly, examples of common methodological flaws in randomised clinical trials (RCTs) will be discussed, showing that things can be seriously wrong with regard to the most solid elements of evidence. The fourth methodological problem will deal with the selection of the patients for whom the evidence holds, while the fifth focuses on the ability of outcome measures to detect clinically relevant changes over time. The methodological problems will be illustrated by examples from our own research group dealing with low back pain and shoulder pain.

2. What evidence and how to use it

What evidence are we talking about? The shortest answer to this question is that it should be medicine-based evidence, which means that we are talking about data collected in clinical care settings, evaluating patients as opposed to cell cultures and other type of laboratory data. Applied clinical research provides the type of evidence we need in EBM. The principles that are used in EBM have been developed in the field of clinical epidemiology. The central issues are typically about effectiveness/efficacy of interventions on the one hand and efficiency/cost effectiveness on the other hand, focussing on value for money in a comparative perspective. The outcome measures are usually patient-oriented, dealing with complaints and symptoms that are important to the patient. Furthermore, these outcomes typically focus on disabilities, handicaps, and quality of life. Again, the research has often little to do with laboratory work, but first and foremost with measurements in everyday clinical care. This does not mean that pathophysiological insights are not important. Most treatments that make sense will be based on these insights. However, pathophysiological knowledge is insufficient to provide the evidence needed to enable EBM. There are many examples in the literature of treatments that should work from a pathophysiological point of view, but were not effective when applied in health care situations. In a way, we could say that less emphasis is put on the authority of experts and more on the results of clinical research. And therefore, efforts are made to replace authority-based medicine with EBM.

Once the evidence on a specific health care problem is available, there are several ways to obtain the evidence and use it in clinical practice. The simplest way is to try to find an original primary publication dealing with patients similar to the one requiring your attention. The problem will often be that you lack the time to look for it, and once you have identified a publication, it is often not clear on the issues you consider being important. Moreover, there might be severe methodological flaws that you may not be able to detect. Besides, you may have no idea how this publication fits in the whole range of available evidence. A good alternative is to look for a

review that has systematically evaluated all relevant publications on the topic [2,3]. Such systematic reviews sometimes try to include not only published, but also unpublished material. But again, the problem is that many reviews are not answering practical questions and may come up with statements like "we are not sure" or "more research is needed", which is not very useful to the clinician. That is the reason why clinical guidelines are becoming very popular, in general practice as well as in other fields. Clinical guidelines for diagnosis or treatment are usually based on the available evidence, but try to supplement the evidence with sensible expert opinions on the topic [4,5]. Nonetheless, uncertainty concerning diagnosis or optimal treatment always remains, due to incomplete evidence and clinical variation among patients. The last remark is a tricky one, as this may be the most popular excuse for not adhering to clinical guidelines. Olli Miettinen, who is a kind of godfather in epidemiology, always explains that the problem is not that the patients are unique, but that the doctors are so unique. It might be a good idea to try to adhere more to clinical guidelines.

3. Variation among reviews and clinical guidelines

When looking at the available evidence you might get worried by the fact that there is such a broad variation among reviews and clinical guidelines. One example comes from the field of low back pain (LBP) research [6]. In 1995 36 publications were available reporting the results of randomised clinical trials (RCTs) on the effectiveness of spinal manipulation for LBP. However, without too much effort, no less than 51 reviews on this topic were identified that included at least several of these trials. Perhaps surprisingly, there was substantial variation in the conclusions of the reviews. Some were positive, some negative, and some were doubtful about the effectiveness of spinal manipulation for LBP. Obviously, they cannot all be right. This may cause a serious problem when you would like to use the evidence in your treatment decisions. We developed an instrument to assess the quality of the reviews. This 'review of reviews' showed a wide variation in quality among the 51 reviews. The conclusions were more likely to be positive in high quality reviews, in reviews that had identified a larger proportion of the available RCTs, and when one of the authors of the review was a manipulator. These findings suggest that the optimal review would show that manipulation works. However, we were not really convinced ourselves. The main problem is that the majority of the RCTs that were included in the reviews were of poor methodological quality, limiting the strength of evidence in favour of spinal manipulation [6,7,8].

As yet, many clinical guidelines for LBP have been proposed, including those published by the Quebec Task Force, Canada, 1987; the Agency for Health Care Policy and Research, USA, 1994; the Clinical Standards Advisory Group, UK, 1996; the Royal College of General Practitioners, UK, 1996; and the Dutch College of General Practitioners, NL, 1996. The conclusions from these guidelines differ slightly on some issues, but substantially on others, for example on the use of spinal manipulation. And again, as not all these guidelines can be right, this may cause problems when you would like to use clinical guidelines in caring for your patients. One of the reasons for the conflicting recommendations is, of course, that the reviews on these topics varied in their conclusions. Another reason may be that those guidelines vary across countries that have different health care systems. For example, guidelines may differ in countries with relatively low or high numbers of chiropractors and manipulative therapists. A third reason may be that the members of the expert panels responsible for constructing the guidelines differ with respect to profession and expertise. This example shows that there may be more variation in guidelines than you would suspect from a rational point of view.

4. Methodological flaws in RCTs

The results of studies on shoulder pain will be used to illustrate the third problem; methodological flaws in RCTs. Intrinsic shoulder complaints (pain or stiffness) are very common in the general population and in primary care. We studied 392 consecutive patients who consulted their general practitioner (GP) for shoulder complaints over a period of one year. Eleven practices (18 GPs) participated in this observational study [9]. The cumulative incidence was 11.2 per 1000 patients per year. We studied the diagnostic classification of shoulder complaints according to the practice guidelines issued by the Dutch College of General Practitioners in 1990 (Table 1) [10].

Table 1 Summary of the clinical guidelines for the diagnosis of shoulder pain (Dutch College of General Practitioners [10]).

Syndrome	*Diagnostic criteria*
Capsular syndrome (capsulitis, arthrosis, frozen shoulder, etc.)	Restriction of lateral rotation, abduction, and medial rotation. Pain in C5 dermatome.
Acute bursitis	Restriction of abduction. Severe pain in C5 dermatome. Acute onset, no evident preceding trauma.
Acromioclavicular syndrome	Restriction of horizontal adduction. Pain in the area of the acromioclavicular joint and/orC4 dermatome.
Subacromial syndrome Rotator cuff tendinitis Chronic bursitis Rotator cuff tears	Painful arc during abduction. Pain in the C5 dermatome. No restriction in passive range of motion. At least one positive resistance test: Bursitis: variable/little pain, normal power Tendinitis: pain, normal power Cuff tears: little pain, loss of power
Others Unclear clinical pictures Extrinsic causes	

These guidelines are largely based on the methods of diagnosing and treating shoulder disorders proposed by an English orthopaedist, James Cyriax [11]. Table 2 presents distribution of occurring syndromes in our study. The content of treatment at the first consultation is presented in Table 3. The outcomes are roughly in agreement with the guidelines that suggest physiotherapy for tendinitis and local injections for bursitis.

Table 2 Syndrome distribution over categories [9].

Syndrome	*Percentage (%)*
capsular syndrome	22
acute bursitis	17
acromioclavicular syndrome	4
subacromial syndrome	48
rotator cuff tendinitis	(30)
chronic bursitis	(13)
cuff tears	(5)
unclear	9

Based on these results it could be concluded that the GP acts according to the guidelines in many cases. However, several questions remained unanswered, three of which will be discussed. Firstly, how reproducible is this diagnostic classification? Would a doctor, when examining the

same patients a second time, make the same diagnostic classification? What about another doctor and what about physiotherapists? This second question also refers to the relevance of a diagnostic classification for the choice of treatment: does it really matter what diagnosis is made at the first consultation? Thirdly, are these treatments helping the patients at all?
We will first address this last question. Three systematic reviews were conducted to evaluate the effectiveness of Non-Steroidal Anti-Inflammatory Drugs (NSAIDs), corticosteroid injections, and physiotherapy [12,13,14]. Evidence from many RCTs was available, but on average their methodological quality was rather poor. Therefore, it was impossible to formulate strong conclusions regarding the effectiveness of the interventions. Table 4 summarises the main results of the reviews.

Table 3 Treatment distribution [9]

Treatment at first consultation	*Percentage (%)*
wait and see	28
medication (NSAIDs)	38
physiotherapy	30 (tendinitis overrepresented)
injection (steroid)	22 (bursitis overrepresented)

The results of trials with acceptable methods demonstrated superior short-term effects of NSAIDs and corticosteroid injections, whereas the effects of physiotherapeutic ultrasound were not significantly better than placebo. Methodological flaws and inconsistencies among RCTs of acceptable methods were still present. And again, inconsistencies are worrying when you would like to use the results of trials to select the optimal treatment for a patient. Another problem was the fact that only few RCTs directly compared different treatment alternatives, such as NSAIDs versus physiotherapy, or corticosteroid injections versus physiotherapy, while these are the important questions from a practical point of view.

Table 4 Main results of systematic reviews on the effectiveness of interventions for shoulder pain [12,13,14]

	N	*Acceptable methods*	*Conclusions*
NSAIDs	19	5	superior short-term effects compared to placebo
steroid injections	16	3	superior short-term effects compared to placebo
physiotherapy	20	6	ultrasound is ineffective

It should be noted that many trial reports were insufficiently clear regarding important aspects of trial methodology, including the method of randomisation, concealment of allocation, prognostic similarity at baseline, and operationalisation of treatments. This hampers the assessment of methodological quality, and leaves you uncertain about the best way to treat your patients with shoulder pain. The recent publication of guidelines for reporting clinical trials will hopefully enhance the informativeness of trial reports [15]. An additional problem is that the trials used many different inclusion criteria and outcome measures, indicating considerable clinical heterogeneity. Sample sizes were often too small to detect relevant effects. This problem may be solved by statistically pooling the results of these small trials, but pooling was not considered to be sensible given the clinical heterogeneity of the trials. Other prevalent methodological flaws concerned a lack of blinding of patients and outcome assessment, no control of co-interventions, high proportions of withdrawals and missing values, and limited follow-up periods. However, the review also made clear that high quality trials might be feasible in this field [e.g. 16,17]. Table 5 shows the checklist we use to assess the methodological quality of a trial [18].

Table 5 Criteria List for the Methodologic Quality Assessment [18].

Patient selection	
a. Were the eligibility criteria specified?	Yes/No/Don't know
b. Treatment allocation	
1) Was the method of randomisation performed?	Yes/No/Don't know
2) Was the treatment allocation concealed?	Yes/No/Don't know
c. Were the groups similar at baseline regarding the most important prognostic indicators?	Yes/No/Don't know
Interventions	
d. Were the index and control interventions explicitly described?	Yes/No/Don't know
e. Was the care provider blinded to the intervention?	Yes/No/Don't know
f. Were co-interventions avoided or comparable?	Yes/No/Don't know
g. Was the compliance acceptable in all groups?	Yes/No/Don't know
h. Was the patient blinded to the intervention?	Yes/No/Don't know
Outcome measurement	
i. Was the outcome assessor blinded to the intervention?	Yes/No/Don't know
j. Were the outcome measures relevant?	Yes/No/Don't know
k. Were adverse effects described?	Yes/No/Don't know
l. Was the withdrawal/drop-out rate described and acceptable?	Yes/No/Don't know
m. Timing follow-up measurements	
1)Was a short-term follow-up measurement performed?	Yes/No/Don't know
2) Was a long-term follow-up measurement performed?	Yes/No/Don't know
n. Was the timing of the outcome assessment in both groups comparable?	Yes/No/Don't know
Statistics	
o. Was the sample size for each group described?	Yes/No/Don't know
p. Did the analysis include an intention-to-treat analysis?	Yes/No/Don't know
q. Were point estimates and measures of variability presented for the primary outcome measures?	Yes/No/Don't know

Internal validity criteria: b, e, f, g, h, i, j, l, n, p.
Descriptive criteria: a, c, d, k, m.
Statistical criteria: o, q.

5. Picking the right patients: diagnostic issues

A diagnostic classification is useful if it is reproducible and has consequences for management. The patients in our observational study on shoulder complaints that were referred to a physiotherapist were also examined and classified by the participated physiotherapists. The kappa statistic, which quantifies the magnitude of agreement adjusted for agreement by chance on a scale from –1 to 1 (0 indicating that agreement is no better than chance), was used to study the agreement regarding the diagnosis between GPs and physiotherapists [19]. The results show that reproducibility was rather poor (Table 6). This outcome is alarming when you realise that a decision concerning treatment may often be based on the diagnosis. Table 7 represents the treatment given by the physiotherapist.

Table 6 Interobserver agreement between GP and physiotherapist [19]

Syndrome	*Kappa*
capsular syndrome	0.48
acute bursitis	-0.03
acromioclavicular syndrome	0.36
subacromial syndrome	0.33
OVERALL	0.31

Table 7 Distribution of treatment among the patients [19].

Treatment	*Percentage (%)*
exercise therapy	56 (caps. syndrome overrepresented)
passive mobilisations	44 (caps. syndrome overrepresented)
deep friction massage	68 (tendinitis overrepresented)
physical applications	48

Another study on the diagnosis of shoulder complaints, conducted in a different population, evaluated agreement between two experienced and well-trained physiotherapists [20]. The outcome was better, but still only moderate reproducibility could be demonstrated (Table 8). Diagnostic disagreement was influenced by several factors, including bilateral involvement that was associated with a twice as large likelihood of disagreement. In patients with complaints that existed longer than 6 months, the likelihood of disagreement was also twice as large. This suggests that diagnosing shoulder pain is particularly difficult in patients with long-lasting, chronic symptoms. A three-fold likelihood of disagreement was found in patients with severe pain, which may be explained by difficulties in performing the physical examination in these cases (Table 9) [20].

A lack of agreement causes unintended variation in treatment decisions. This problem can be solved in three ways; by simplifying the taxonomy, by further operationalisation of the diagnostic categories, and by training the observers. However, the relevance of a (pathophysiological) diagnostic classification of shoulder pain in primary care is still unclear. Currently, a revision of the Dutch practice guidelines for shoulder complaints is underway, in which a simplification of the diagnostic classification is proposed.

Table 8 Interobserver agreement between two physiotherapists [20].

	kappa
capsular syndrome	0.63
acute bursitis	0.50
acromioclavicular syndrome	0.24
subacromial syndrome	0.56
OVERALL	0.45

Table 9 Determinants of disagreement between two physiotherapists [20].

	odds ratio
bilateral involvement	1.9
> 6 months complaints	2.0
severe pain	2.7

6. Selecting the right outcome: Responsiveness

It should be clear that using the wrong outcome measures in RCTs can be disastrous. Using the wrong outcome measure may lead to irrelevant positive or incorrect negative conclusions of RCTs and reviews. The ability to detect a clinically relevant change over time (responsiveness) is, therefore, an important characteristic of an outcome measure [21,22]. Responsiveness is closely connected to sample size and power of a trial. When the responsiveness is low, very large sample sizes are needed to detect a clinically relevant change after treatment. In the study of shoulder complaints, improvement of functional disability is important, next to the reduction of pain. One third of 55 RCTs included in our reviews only used a single question to measure functional disability. Such a rough assessment may be inadequate and not responsive to change. Therefore, the Shoulder Disability Questionnaire (SDQ) was designed (Table 10) [23]. This instrument consists of 16 items representing activities or situations that may cause symptoms in patients with shoulder pain. The response options are 'yes', 'no' or 'not applicable'. The final score ranges between 0 (no disability) to 100 (all applicable items positive).

Table 10 Shoulder Disability Questionnaire [23]

INSTRUCTION FOR SDQ COMPLETION

When your shoulder hurts, you may find it difficult to do certain things you normally do. This list contains 16 sentences that people have used to describe themselves when they have shoulder pain. When you read them, you may find that some stand out because they describe you today (last 24 hours). As you read the list, think of yourself today (last 24 hours). Ask yourself if you performed the activity today.

SDQ ITEMS

01	I wake up at night because of shoulder pain.
02	My shoulder hurts when I lie on it.
03	Because of pain in my shoulder it is difficult to put on a coat or a sweater.
04	My shoulder hurts during my usual daily activities.
05	My shoulder hurts when I lean on my elbow or hand.
06	My shoulder hurts when I move my arm.
07	My shoulder hurts when I write or type.
08	My shoulder is painful when I hold the driving wheel of my car or handlebar of my bike.
09	When I lift and carry something my shoulder hurts.
10	During reaching and grasping above shoulder level my shoulder hurts.
11	My shoulder is painful when I open or close a door.
12	My shoulder is painful when I bring my hand to the back of my head.
13	My shoulder is painful when I bring my hand to my buttock.
14	My shoulder is painful when I bring my hand to my low back.
15	I rub over my painful shoulder more than once during the day.
16	Because of my shoulder pain I am more irritable and bad tempered with people than usual.

Table 11 Responsiveness of three outcome measures for shoulder disability [23].

	SDQ	PSS	FSQ
RR_{1m}	2.2	2.5	-
RR_{6m}	1.9	2.2	-
AUC_{1m}	0.84	0.84	0.72
AUC_{6m}	0.88	0.86	0.79

The responsiveness of the SDQ was studied in the same series of patients mentioned before; 349 consecutive incident cases in general practice [24]. Follow-up questionnaires were sent to all patients after 1 and 6 months. Improvement reported by the patients themselves was used to discriminate between clinical stability, deterioration, and improvement at follow-up. The responsiveness of the SDQ (0-100) was compared to that of a Pain Severity Score (PSS) (0-100), and a single question on functional disability (FSQ) (1,2,3). Responsiveness was evaluated using two methods. First, the Responsiveness Ratio (a 'signal to noise ratio') was computed as the ratio of the mean change in clinically improved patients to the standard deviation of the mean change in clinically stable patients [21]. When this ratio (RR) equals 1 there is no responsiveness. When the RR > 1, the signal is larger than the noise, and responsiveness increases. Second, Receiver Operating Characteristic (ROC) curves were constructed. These curves provide the sensitivity and specificity for distinguishing between improved and stable patients at different cut-off points of an outcome measure. The area under the ROC curve (AUC) gives the probability of correctly classifying random pairs of improved and stable patients [25]. An AUC of 0.5 indicates no responsiveness. The results of RR and AUC for the measurements of the SDQ, PSS, and FSQ after 1 and 6 months are presented in Table 11. It can be concluded that the responsiveness of the SDQ is adequate (RR > 1 and AUC > 0.5), similar to that of the PSS, and somewhat better than a single question on disability (FSQ). A cut-off point of 18.75 points (an improvement of at least 3 items) was the optimum trade-off with a sensitivity of 74% and a specificity of 77%. The mean change (ΔSDQ) in clinically improved patients was 40 points (±6 items) with a sensitivity of 46% and specificity of 98% if used as a cut-off point distinguishing between improved or stable patients. A drawback to this method of

evaluating responsiveness is the fact that the criterion for clinical relevant improvement was based on subjective judgement (improvement according to the patients themselves). Consequently, results on responsiveness cannot easily be generalised to other populations. At present, several instruments that measure shoulder disability are available. In stead of designing yet another shoulder disability questionnaire, effort should be put in directly comparing their responsiveness in a single population.

7. Final remarks

There appears to be substantial variation among clinical guidelines and systematic reviews, which may be partly caused by the fact that many RCTs are poorly executed and reported, and that RCTs of acceptable quality often show inconsistent results. In addition, there seems to be a failure to focus on clinically homogeneous groups of patients. Trials that make direct comparisons between alternative treatment options are needed. They will provide the evidence that will facilitate clinical decisions. With respect to diagnosing soft-tissue disorders, poor reproducibility is common, and the relevance of diagnosis for treatment decisions is often unclear. The responsiveness of primary outcome measures in RCTs needs further evaluation.
Finally, Evidence-Based Medicine is, as yet, more an ideal than reality, but substantial progress is being made in making the evidence available to those responsible for patient care.

References

[1] L.M. Bouter, Prevalence of methodological errors in rehabilitation research, *Rheumatology in Europe* **24 (Supplement 2)** (1995) 85-87.
[2] C.D. Mulrow *et al.*, Systematic reviews – critical links in the great chain of evidence, *Annals of Internal Medicine* **126** (1997) 389-391.
[3] C. Mulrow *et al.*, Integrating heterogeneous pieces of evidence in systematic reviews, *Annals of Internal Medicine* **127** (1997) 989-995.
[4] D.J. Cook *et al.*, The relation between systematic reviews and practice guidelines, *Annals of Internal Medicine* **127** (1997) 210-216.
[5] D.J. Cook *et al.*, Systematic reviews – synthesis of best evidence for clinical decisions, *Annals of Internal Medicine* **126** (1997) 376-380.
[6] W.J.J. Assendelft *et al.*, The relation between methodological quality and conclusions in reviews of spinal manipulation, *JAMA* **274** (1995) 1942-1948.
[7] B.W. Koes *et al.*, Methodological quality of randomized clinical trials on treatment efficacy in low back pain, *Spine* **20** (1995) 228-235.
[8] L.M. Bouter *et al.*, Methodologic issues in low back pain research in primary care, *Spine* **23** (1998) 2014-2020.
[9] D.A.W.M. van der Windt *et al.*, Shoulder disorders in general practice: incidence, patient characteristics and management, *Annals of the Rheumatic Diseases* **54** (1995) 959-964.
[10] J.F. Bakker *et al.*, Standaard Schouderklachten, *Huisarts en Wetenschap* **33** (1990) 196-202.
[11] J. Cyriax, Textbook of Orthopaedic Medicine, chapters 9 and 10, London, Baillière Tindal (1981) 190-254.
[12] D.A.W.M. van der Windt *et al.*, The efficacy of non-steroidal anti-inflammatory drugs (NSAIDS) for shoulder complaints: a systematic review, *Journal of Clinical Epidemiology* **48** (1995) 691-704.
[13] G.J.M.G. van der Heijden *et al.*, Steroid injections for shoulder disorders: a systematic review of randomised clinical trials, *British Journal of General Practice* **46** (1996) 309-316.
[14] G.J.M.G. van der Heijden *et al.*, Physiotherapy for patients with soft tissue shoulder disorders: a systematic review of randomised clinical trials, *British Medical Journal* **315** (1997) 25-30.
[15] D.G. Altman, Better reporting of randomised controlled trials: the CONSORT statement, *British Medical Journal* **313** (1996) 570-571.
[16] D.A.W.M. van der Windt *et al.*, Effectiveness of corticosteroid injections versus physiotherapy for treatment of painful stiff shoulder in primary care: randomised trial, *British Medical Journal* **317** (1998) 1292-1296.
[17] G.J.M.G. van der Heijden et al. The efficacy of ultrasound therapy and electrotherapy for shoulder disorders. Results of a randomized placebo-controlled clinical trial. In press.
[18] M.W. van Tulder *et al.* (Back Group Editorial Board), Methodologic guidelines for systematic reviews in the Cochrane Collaboration back review group for spinal disorders, *Spine* **22** (1997) 2323-2330.

[19] C. Liesdek *et al.*, Soft-tissue disorders of the shoulder: a study of inter-observer agreement between general practitioners and physiotherapists and an overview of physiotherapeutic treatment, *Physiotherapy* **83** (1997) 12-17.
[20] A.F. de Winter *et al.*, Diagnostic classification of shoulder disorders: inter-observer agreement and determinants of disagreement, Submitted for publication.
[21] G.H. Guyatt *et al.*, Measuring change over time: assessing the usefulness of evaluative instruments. *Journal of Chronic Diseases* **40** (1987) 171-8.
[22] G.H. Guyatt *et al.*, Measuring health status: what are the necessary measurememt properties?, *Journal of Clinical Epidemiology* **45** (1992) 1341-1345.
[23] G.J.M.G. van der Heijden *et al.*, Shoulder disability questionnaire: design and responsiveness of a functional status measure, In press.
[24] D.A.W.M. van der Windt *et al.*, The responsiveness of the Shoulder Disability Questionnaire, *Annals of the Rheumatic Diseases* **57** (1998) 82-87.
[25] RA Deyo *et al.*, Assessing the responsiveness to change of functional scales to clinical change: an analogy to diagnostic test performance, *Journal of Chronic Diseases* **39** (1986) 897-906.