Int. Disabil. Studies, **12**, 61 – 65

Reliability of the assessment of impairments and disabilities in survey research in the field of physical therapy

E. F. van Triet, J. Dekker, J. J. Kerssens and E. Chr. Curfs

Netherlands Institute of Primary Health Care, Utrecht, The Netherlands

Accepted for publication: May 1990
Correspondence to: J. Dekker, Netherlands Institute of Primary Health Care, PO Box 1568, 3500 BN Utrecht, The Netherlands

Key words Diagnostic assessment – ICIDH – Physical therapy – Survey research

Summary *Objective:* to evaluate the reliability of diagnostic assessment based on clinical observations by physical therapists. *Design:* interobserver studies between two pairs of physical therapists. *Setting:* two primary-care physical therapy practices. *Patients:* all applying to the practices for treatment. *Exclusions:* physical therapy in previous 3 months, or need for acute treatment. *Assessments:* schedule derived from the ICIDH. *Results:* agreement on disabilities better than on impairments, the latter revealing problems particularly with pain and restricted range of motion. *Conclusions:* reliability of assessments of most of the categories considered was reliable; in two categories the reliability was poor, and the categories were modified.

Introduction

Epidemiological surveys ideally rely on a battery of well standardized diagnostic tests. However, such an approach can create logistic problems because many tests require technical or personnel assistance that may not be regularly available. An alternative approach is to rely on clinical observation. This approach has been used, for example, in morbidity studies in general practice.[1] Clinical observation may be less well standardized, but it is more practical. Especially in a relatively new field of investigation, where little prior information is available, the benefits of standardizing testing do not easily outweigh the costs.

Physical therapy is such a field. There is almost no survey research available in this field. Well standardized tests, if available, often require specialized equipment and extra personnel. Therefore clinical observations seem the obvious means for collecting diagnostic data in survey research in physical therapy. If this approach is chosen, however, it is important to determine whether it leads to reliable (and reproducible) results. The present study set out to test the interobserver reliability of diagnostic assessment based on clinical observations by physical therapists.

Diagnostic assessment by physical therapists is concerned with functional disorders that arise mainly as a result of disease. There is no accepted system to classify assessment findings in physical therapy. Disease classifications, such as the *International Classification of Diseases* (ICD)[2,3] or the *International Classification of Primary Care* (ICPC),[4] are not adequate because they do not contain a systematic description of functional disorders. The *International Classification of Impairments, Disabilities, and Handicaps* (ICIDH)[5] is regarded as an excellent point of departure for the classification of assessment findings by physical therapists.[6–8] However, quite a number of categories that are important to a physical therapist do not occur in the ICIDH.

This is particularly true at the level of impairments, the consequences of disease at the organ level. At this level physical therapists formulate their findings in terms such as restricted range of motion and diminished muscle strength.[9,10] It is not possible to code these categories adequately in the ICIDH. For example, diminished strength of the muscles of the upper leg may be coded as 'other paralysis of lower limb – other weakness' (code I 73.46). The use of such 'other' categories is not satisfactory, because it is not very informative. It was therefore thought necessary to develop a revised schedule. This schedule draws on the conceptual framework of the ICIDH but, predominantly at the impairment level, diagnostic categories are different.

Methods

SCHEDULE

Drafting the revised schedule was guided by two considerations. First, diagnostic assessment in physical therapy is primarily at the level of impairment and disability, i.e. disorders at the level of the organ and the person. Assessment of handicap, disorders at a societal level, appeared to be a less appropriate task for a physical therapist,[11] and so handicaps were not included in the recording schedule. Secondly, in the Netherlands most physical therapists (about 65%) work in primary health care.[12] The majority of the patients in primary care (84%) are referred for complaints relating to the musculoskeletal system or injuries resulting from accidents.[13] In the schedule emphasis was therefore placed on impairments and disabilities arising in these disorders that are actually treated in the primary-care setting.

The principal elements of the schedule are shown in Table 1. For each impairment one or more localizations were indicated, the latter not being specified (i.e. an open format) but being left to the therapist to identify and record. Disabilities were subdivided into a number of categories (Table 1). The severity of an impairment, with its specified localization/s, and of a disability were indicated on scales (also noted in Table 1).

ASSESSMENT

Assessment data according to the schedule were based on clinical observations by physical therapists. In order to evaluate interobserver reliability, a pair of therapists assessed a series of patients independently. Two physical therapy practices in a primary-care setting were included in the study, so that two pairs of therapists were involved. The latter were all male, aged between 30 and 36 years, and details of their training and experience are recorded in Table 2.

The therapists received a brief period of training in use of the schedule, together with written information about its use. Practical experience was gained with five patients who were not included in the definitive study; mutual comparisons of records were made, on which the researchers commented. Impairments were evaluated by clinical tests and observations, but disabilities were assessed with the help

0379-0797/90 $3.00

of information in the records of patients' histories. After the training period there was no further consultation between the therapists on the subject of records.

All patients applying for treatment were eligible for inclusion in the study, and clinical details are shown in Table 3. Patients were excluded if they had received physical therapy in the previous 3 months, or if they required acute treatment or refused to participate. Medical diagnoses were established by referring physicians, and were classified according to ICPC.[4] The average age of patients in Practice 1 was 49 years (SD 20 years), and in Practice 2 was 36 years (SD 14 years); the average age of excluded patients was 40 years (SD 16 years). Each patient was assessed twice; the first assessment took place when the patient applied for treatment and the second, by a second therapist, after an average of 2 days (median 2 days, with a maximum of 7 days). Treatment was initiated after the second assessment.

ANALYSIS

Data on the localizations of impairments were subdivided into a number of areas (noted in the Appendix); when a particular impairment was recorded at more than one site, that with the highest severity rating was selected for analysis. Severity scores were dichotomized because the full scales, which reflected a clinical approach, offered too much detailed information for epidemiological survey research. 'Normality' (i.e. no impairment present) was taken as a score of 0 for muscle tone or pain, and all other scores were equated with 'impairment present'. A similar approach was followed with other impairments, 'normality' being regarded as a score of 5 for muscle strength, up to and including 5° reduction in range of motion, up to and including 1 cm of swelling, or absence of postural impairment; the other impairments were not analysed.

The degree of agreement between the pairs of observers was determined in two ways. The simplest was to compute the percentage agreement (i.e. number of patients on which the observers agreed as a proportion of all patients studied), but this does not allow for agreement arising from chance. The other measure was to use Cohen's kappa, which does correct for chance agreement.[18] The maximum value of kappa is 1, but this is seldom achieved.[19-21] Interpretation is assisted by the terminology of Fleiss;[20] a kappa value of more than 0.75 indicates an excellent degree of agreement, one of between 0.40 and 0.75 is 'fair to good' (which we designated satisfactory), and a value of less than 0.40 indicates a low level of agreement.

The frequency distribution of many impairments and disabilities appeared to be skewed, in that they were identified among only a few patients. This caused small differences in assessments to have a major influence on kappa. For example, if a therapist reported that an impairment was present in three patients and absent among the 47 others, whilst the other therapist recorded one and 49, respectively, then kappa was 0.49, but if they both reported impairment in three and absence in 47 then kappa was 1. To our knowledge there are no criteria for relating kappa to the skewness of observations (see also [21]). In the absence of such criteria we adopted the following rules: (i) kappa was determined only if each therapist within the pair recorded an impairment (with a certain localization) or a disability in at least 10% of patients — with a sample of 50 patients this meant an impairment or disability had to be observed in at least five patients, and the expected frequency in the cell in which both therapists reported an impairment present should therefore be at least 0.5; (ii) kappa was calculated but had to be interpreted carefully if the therapist observed an impairment or disability in 10 – 20% of the patients (expected frequency between 0.5 and 2); (iii) if an impairment or disability was reported present by both therapists in 20% or more of patients, kappa was calculated and interpreted in a straightforward manner (expected frequency greater than or equal to 2).

Table 1 Impairments and disabilities specified in the assessment schedule

Impairment or disability category	*Components*	*Severity*
Impairments		
Range of motion	Active or passive restriction	Degrees of deviation from normal (from starting positions of AAOS)[14]
Muscle tone	Increase or decrease	Scored on 6-point scale (0 = no impairment; 5 = serious)
Muscle strength	Diminished	Scored on 6-point scale (after Kendall *et al.*)[15]
Swelling	cf. non-affected joint	Scored in centimetres
Posture	Kyphosis, lordosis, or scoliosis	Scored 'absent' or impairment of first, second, or third degree[16]
Respiratory	Sputum retention or shortness of breath	Frequency of shortness of breath
Pain	On muscle contraction or stretch, joint compression, pain on pressure of other tissues, or radiating pain	Scored on 6-point scale (0 = no pain; 5 = severe or intense)[17]
Pain complaints (by patient)	At rest, on movement, or transition rest/movement	Severity scored by frequency (0 = never; 5 = continuously)
Miscellaneous	Cardiac; circulatory; sensitivity, proprioception, and psychomotor functions; or involuntary movements	Scored on a 6-point scale (0 = none; 5 = severe or intense)
Disabilities†		
Self care	Washing, dressing, using lavatory, eating	
Physical control	Grasping and gripping, writing, kneeling, bending, keeping balance	
Mobility	Getting in and out of bed, walking, climbing stairs, cycling, driving a car	
Occupational	Doing shopping, preparing meals, changing beds, doing housework, caring for other household members, using telephone, sitting for long periods, standing for long periods, lifting, maintaining a normal tempo, stress resistance	
Recreational	Running, jumping, throwing or catching	

†Severity assessed on the 7-point ICIDH scale.[5]

Table 2 Training and experience of physical therapists

Characteristic	*Physical therapist*			
	Practice 1		*Practice 2*	
	I	*II*	*III*	*IV*
Training institution†	A	B	B	C
Year of graduation	1980	1983	1980	1984
Experience (years)	8	5	8	4
Additional training	Chiropractic, kinematics	Manual therapy	Acupuncture	Orthopaedic medicine, podology, osteopathy

†Letters indicate institutions of training.

Table 3 Characteristics of patients considered for entry into the study

Characteristic	*ICPC codes*[4]	*Patients included*		*Patients excluded*
		Practice 1	*Practice 2*	
Sex				
male		18	29	16
female		33	21	12
Medical diagnosis†				
cervical spine disorders	L01, L83, L84	5	1	3
lumbar spine disorders	L86	23	12	10
musculoskeletal injury	L72, L73, L76, L77, L79, L81	4	7	4
degenerative disorders:				
general	L91	3	2	0
spine	L84, L85	0	4	0
hip	L89	1	1	0
knee and ankle	L15, L90, L91, L97	3	3	3
back disorders, general	L02	0	0	2
shoulder syndrome	L08, L92	2	5	2
tennis elbow	L93	1	2	0
meniscus damage	L96	1	1	0
headache	N01	6	0	1
other diagnosis	L09, L54, L66 L98, L99, P78	3	2	3
no diagnosis stated		6	13	3
Total patients		51	50	28

†Diagnostic entries not exclusive (i.e. if more than one diagnosis made) and so item totals exceed column total.

Table 4 Estimates of reliability of assessments of impairments

Impairment	*Practice 1*		*Practice 2*	
	Percentage agreement	*kappa*	*Percentage agreement*	*kappa*
Restricted range of motion:				
active	78	0·43	76	0·31
passive	+	+	76	0·31
Increased muscle tone	78	0·57	84	0·62
Diminished muscle strength	+	+	76	0·45
Swelling	+	+	78	0·29†
Postural impairment:				
kyphosis	80	0·33†	+	+
lordosis	76	0·26†	+	+
scoliosis	90	0·49†	+	+
Pain:				
on stretching	75	0·47	64	0·29
on pressure	78	0·34	88	0·19
on contraction	71	−0·06†	84	0·66
joint compression	28	0·22	+	+
Circumstances of pain:				
at rest	90	0·77	80	0·59
on movement	88	0·65	92	0·31
rest/movement	90	0·76	82	0·38

+impairment confirmed in less than 10% of patients.
†impairment confirmed in 10–20% of patients.

Results

Assessments of agreements were calculated for each pair of therapists, and the results for impairments are shown in Table 4; respiratory and miscellaneous impairments were not analysed because they were observed too infrequently (in less than 10% of patients). In most cases the percentage agreement was within acceptable limits (76–92%), though for contraction and compression pain in Practice 1 and for stretch pain in Practice 2 the results were less satisfactory (i.e. less than 75%).

The kappa values revealed slightly less encouraging results once chance agreement had been taken into account, and the impact of the findings was diminished by the frequency with which only small numbers of patients with certain specific impairments were encountered (Table 4). Only two excellent results were observed for impairments, both relating to the circumstances of pain recorded in one of the practices. Of the remaining results in Table 4, half of those in Practice 1 showed satisfactory agreement whereas somewhat less than half in Practice 2 were satisfactory. The reliability of agreement varied between the two pairs of therapists in regard to individual impairments, but again small numbers limited the conclusions that could be drawn from such comparisons.

The results for disabilities showed greater consistency (Table 5), and in all instances where there were adequate numbers of patients the percentage agreement was within acceptable limits (i.e. greater than 76%). Kappa values reinforced this encouraging picture. Numbers were inadequate in both practices in regard to two disabilities, but of the remaining disabilities a quarter were excellent in one practice and a third in the other; in both practices a further half of the results were satisfactory.

Table 5 Estimates of reliability of assessments of disabilities

Disability	*Practice 1*		*Practice 2*	
	Percentage agreement	*kappa*	*Percentage agreement*	*kappa*
Self care:				
washing	94	0·77†	96	0·83†
dressing	96	0·87†	84	0·57
using lavatory	96	0·83†	94	0·76†
eating	+	+	+	+
Physical control:				
grasping/gripping	+	+	96	0·91
writing	+	+	98	0·90†
kneeling	88	0·74	76	0·47
bending	90	0·80	80	0·56
keeping balance	78	0·26†	+	+
Mobility:				
in and out of bed	86	0·65	88	0·73
walking	88	0·73	96	0·91
climbing stairs	90	0·76	88	0·74
cycling	86	0·66	88	0·63
driving a car	82	0·63	94	0·84
Occupational:				
doing shopping	77	0·47	86	0·71
preparing meals	78	0·26†	92	0·56†
changing beds	77	0·41	100	1·00†
doing housework	77	0·43	88	0·69
caring for others	80	0·34†	+	+
using telephone	+	+	+	+
sitting long periods	80	0·60	84	0·63
standing long periods	90	0·80	82	0·62
lifting	94	0·88	86	0·71
normal tempo	86	0·68	78	0·40
stress resistance	88	0·56†	+	+
Recreational:				
running	80	0·61	90	0·79
jumping	82	0·65	88	0·72
throwing/catching	84	0·59	+	+

+ disability confirmed in less than 10% of patients.
† disability confirmed in 10–20% of patients.

Discussion

The results of this study of the reliability of assessments using a schedule derived from the ICIDH varied from poor to excellent. This indicates that, while some parts of the assessment were adequate, others may need modification. There were greater degrees of agreement about disabilities, as compared with impairments. Of the three disabilities where agreement in one practice was poor, one concerned 'caring for household members'. This has more to do with social role than with activity performance, i.e. it is a handicap rather than a disability, and the activity can therefore be disregarded in the present context. There was no ready explanation for the other unreliable items, 'preparing a meal' and 'keeping balance', and these perhaps merit further study to elucidate reasons for the low reliability.

In relation to impairments the conclusions are more varied. The assessment of pain does not appear to be reliable. Physical therapists are accustomed to assessing pain during evaluation of other impairments, such as pain during testing of range motion. The assessment schedule called for pain to be recorded independent of other impairments, which may have given rise to ambiguities. Reliability might be improved if pain occurring during evaluation of other impairments was recorded, and the schedule has been modified accordingly. Therapists are now required to record whether pain occurs during assessment of range of motion, muscle strength, or muscle tone, and the reliability of this approach will be examined in future work. Another problematic area was the circumstances under which pain was experienced. Although assessment of pain at rest was satisfactory, the reliability of pain on motion and in transition from rest to motion left something to be desired in Practice 2. It is possible that training in regard to these categories had not been adequate in the second practice.

As can be seen in Table 4, assessments varied in reliability for a number of other impairments. This was particularly true for muscle tone. Muscle strength also caused some difficulty. The therapists reported that they had used manual muscle tests to established the degree of loss of muscle strength, and the ability of these tests to determine small differences is limited.[22] However, the results do indicate that the results are reliable if they are used just for broader judgements, such as the presence or absence of diminished muscle strength.

Assessment of ranges of motion also showed some variability, and the reliability of the range in both active and passive motion was poor in Practice 2. This is probably a reflection of the enormous variety of 'normal values' reported in the literature.[23] One remedy could be to provide a list of normal values for therapists participating in future research, although this is scarcely a realistic option for large-scale work using the schedule. Another solution might be to use broader categories, such as recording impaired range of motion as none, some, or severe. Such an assessment would be less precise, but for survey research it might be adequate; the reliability of this approach will be evaluated in future work.

The conclusion is that diminished muscle strength, increased muscle tone, and the circumstances under which pain is experienced can be assessed with sufficient reliability, but that for pain and restricted range of motion some modification of the schedule is advisable. The frequency of decreased muscle tone and respiratory impairments was sufficiently low that no useful analysis could be made. With swelling and impairments of posture it is doubtful whether the frequency is sufficient for meaningful analysis.

As regards the design and overall results of the study some further observations are apposite. There was an average of 2 days between the two diagnostic assessments made on

patients, with a maximum of 7 days. Some real changes in impairment and disability could have taken place during this interval. In such instances the therapists could have recorded discrepant findings without this necessarily indicating unreliability in assessment. The implication is that the present study yields a conservative estimate of the reliability of assessments. Secondly, in research on interobserver reliability there are generally one series of patients and two or more observers, all of whom are employed in the same institution. The present study involved two pairs of therapists working in two different practices. The degree of agreement between the conclusion reached in each practice encourages some confidence in generalization of the results to other therapists.

The design of the present study should be distinguished from research on the reliability of a classification system. The latter is intended to determine whether observations can be unambiguously assigned to a category, given particular diagnostic findings,[24] and the reliability of the findings themselves is not then the subject of research. Our work tested the reliability of diagnostic assessments based on clinical observations, the design being the only one that makes it possible to determine whether survey research may rely on such observations. Our results indicate that for most of the categories considered the reliability was sufficiently high. However, in two cases the reliability was insufficient, and for these two the assessment schedule has been modified.

Acknowledgements

The schedule used in this study was derived from a schedule developed by the National Institute for Research and Training in Physical Therapy (Stichting Wetenschap en Scholing Fysiotherapie). The authors are grateful for co-operation from Dr R. A. B. Oostendorp and Dr C. D. Mischner-van Ravensberg. The authors are also grateful for co-operation by W. Kroon, T. van Loosbroek, Dr H. J. M. Obbens, and N. Stam. The manuscript was prepared by Mrs R. de Groot.

Appendix

For certain impairments (restricted range of motion, swelling, postural impairment, and pain) localizations were aggregated as follows, with recorded localizations shown in parentheses:

1 Head.
2 Spine (cervical, thoracic, lumbar, sacroiliac, coccyx, total spine, and spine NOS [not otherwise stated]).
3 Neck/back (neck, neck – shoulder, back – thoracic, back – lumbar, back – sacral, back – total, back NOS, pelvis, and buttock).
4 Front (cervix, armpit, thorax, breast, abdomen, groin, flank).
5 Lung.
6 Shoulder (shoulder – soft tissue, shoulder joint, clavicle, scapula).
7 Arm (upper arm, elbow, elbow joint, forearm, wrist, wrist joint, hand, hand joint, finger, finger joints, arm, and arm NOS).
8 Hip (hip, hip joint).
9 Leg (upper leg, knee, knee joint, shin/calf, ankle, ankle joint, foot, foot joint, toe, toe joint, leg, leg NOS).

For increased muscle tone and diminished muscle strength localizations were aggregated as follows:

1 Head (muscles of head, including face).
2 Neck/shoulder (muscles of neck, muscles of cervix, trapezius, levator scapulae, and deltoid).
3 Back (muscles of back, erector trunci, rhomboid, latissimus dorsi).
4 Arm (dorsal muscles of upper arm, ventral muscles of upper arm or forearm, muscles of hand and fingers).
5 Hip (iliopsoas, pelvic muscles, gluteus).
6 Abdomen and breast (pectoralis, muscles of abdomen).
7 Upper leg (dorsal or ventral muscles of upper leg, adductor of upper leg, tensor fasciae latae).
8 Shin/calf (dorsal or ventral muscles of shin/calf, muscles of foot and toes).

References

1 Royal College of General Practitioners and Office of Population Censuses and Surveys. *Morbidity Statistics from General Practice, Third National Study 1981 – 1982.* London: Her Majesty's Stationery Office, 1986.
2 World Health Organization. *Manual of the International Statistical Classification of Diseases, Injuries and Causes of Death,* 9th revision. Geneva: WHO, 1977.
3 Commission on Professional and Hospital Activities. *The International Classification of Diseases, 9th Revision, Clinical Modification (ICD – 9 – CM).* Ann Arbor MI: CPHA, 1978.
4 Lamberts H and Wood M (Eds). *International Classification of Primary Care.* Oxford: Oxford University Press, 1987.
5 World Health Organization. *International Classification of Impairments, Disabilities, and Handicaps.* Geneva: WHO, 1980.
6 Wagstaff S. The use of the International Classification of Impairments, Disabilities, and Handicaps in rehabilitation. *Physiotherapy* 1982; **68:** 233 – 234.
7 Nationale Raad voor de Volksgezondheid. *Vooronderzoek fysiotherapie en classificatie.* Zoetermeer: NRV, 1987.
8 Mischner-van Ravensberg CD, Paauw NJM, van Gestel JLM. De fysiotherapeutische werkdiagnose in relatie tot de medische diagnose. *Ned T. Fysiotherapie* 1988; **98:** 104 – 107.
9 Leffelaar EG. *Nederlands Leerboek der Fysiotherapie.* Assen: van Gorcum, 1966.
10 Cyriax J. *Textbook of Orthopaedic Medicine.* London: Baillière Tindall, 1978.
11 Granger CV, Seltzer GB, Fishbein CF. *Primary Care of the Functionally Disabled.* Philadelphia PA: Lippincott, 1987.
12 Ministerie van Welzijn, Volksgezondheid en Cultuur. *Notitie Taakstellingen Beroepskrachtenplanning Gezondheidszorg.* Rijswijk: MWVC, 1988.
13 Kerssens JJ, Curfs E Chr, Groenewegen PP. *Fysiotherapie in de Nederlandse Gezondheidszorg: klachten van patienten, indicatiestelling van (huis)artsen en fysiotherapeutische behandelingen.* Utrecht: Nivel, 1988.
14 Heck CV, Hendryson IE, Rowe CR. *Joint Motion: Methods of Measuring and Recording.* Chicago: American Academy of Orthopedic Surgeons, 1965.
15 Kendall H, Kendall F, Wadsworth G. *Muscle Testing and Function,* 2nd edn. Baltimore: Williams and Wilkins, 1971.
16 Stijns HJ. *Het klinisch onderzoek van het bewegingsstelsel.* Leuven: Acco, 1976.
17 Melzack R. The short-form McGill Pain Questionnaire. *Pain* 1987; **30:** 191 – 197.
18 Cohen J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 1960; **20:** 37-46.
19 Soeken KL, Prescott PA. Issues in the use of kappa to estimate reliability. *Medical Care* 1986; **24:** 733 – 741.
20 Fleis JL. *Statistical Methods for Rates and Proportions,* 2nd edn. New York: John Wiley, 1981.
21 Schouten HJA. *Statistical Measurement of Interobserver Agreement* (dissertation). Utrecht: Elinkwijk, 1985.
22 Lamb RL. Manual muscle testing. In Rothstein JM (Ed.), *Measurement in Physical Therapy,* New York: Churchill Livingstone, 1985.
23 Miller PJ. Assessment of joint motion. In Rothstein JM (Ed.), *Measurement in Physical Therapy.* New York: Churchill Livingston, 1985.
24 van der Horst F, Seelen A, Vissers F, Plagge H, von Höfen R, de Geus C, Kleynen J. Registratie in de huisartsenpraktijk. *Huisarts en Wetenschap* 1985; **28:** 229 – 234.