TABLE 1 Criteria List for Assessing Methodological Quality of Trials

Heading	Subheading	Descriptor
Title		Identify the study as a randomized trial.
Abstract		Use a structured format.
Introduction		State prospectively defined hypothesis, clinical objectives, and planned subgroup or covariate analyses.
Methods	Protocol	Describe
		Planned study population with inclusion or exclusion criteria.
		Planned interventions: their nature, content, and timing.
		Primary and secondary outcome measure(s) and the minimum important difference(s), and indicate how the target sample size was estimated.
		Reasons for statistical analyses chosen, and whether these were completed on an intention-to-treat basis.
		Mechanisms for maintaining intervention quality, adherence to protocol, and assessment of fidelity.
		Prospectively defined stopping rules (if warranted).
	Assignment	Describe
		Randomization (e.g., individual, cluster, geographic).
		Allocation schedule method.
		Method of allocation concealment.
	Masking (blinding)	Describe
		Mechanism for maintaining blind and allocation schedule control.
		Evidence for successful blinding.
Results	Participant flow and follow-up	Provide a trial profile summarizing participant flow, numbers and timing of randomization assignment, interventions, and measurements for each randomized group.
	Analysis	State estimated effect of intervention on primary and secondary outcome measures, including a point estimate and measure of precision (confidence interval).
		State results in absolute numbers when feasible (e.g., 10/20, not 50%).
		Present summary data and appropriate descriptive and interferential statistics in sufficient detail to permit alternative analyses and replication.
		Describe prognostic variables by treatment group and any attempt to adjust.
		Describe protocol deviations.
	Discussion	State specific interpretations of study findings, including sources of bias and imprecision (internal validity) and discussion of external validity, including appropriate quantitative measures when possible.
		State general interpretation of the data in light of the available evidence.

Source: Modified from Machin et al. [21], with permission.

In spite of application of well-defined inclusion/exclusion criteria, it is commonly thought that about half of all clinical trials do not achieve their planned sample size [30]. As a recruitment strategy, necessary sample size is calculated to cover the expected number of dropout cases. It seems to me that the calculated number of the sample size is just for reference as an appeal to the cautious attitude of the researcher when conducting the trial.

TABLE 2 Necessary Steps to Conduct Trials in Musculoskeletal Disorders

Developing protocol	Aim and goals with using strategies
	Simple pragmatic design
	Minimal amount of data correction
	Random allocation
	Ethical considerations
Patients	Clear inclusion/exclusion criteria
	Minimal burden for patients
	Recruitment strategies
Interventions	Drug, therapeutic exercise, physical agents, orthosis, Education or care package
Outcomes	Assessment of functions and QOL, good questionnaire
	Optimal combining outcome measures
	Data management procedures
Time constrained	Limited core activity
Staff and their training	Minimal demand
	Sufficient preparation and supporting

TABLE 3 Developing a Protocol for Clinical Trial

Title
Summary
Background and rationale for the trials
Hypothesis to be tested
Primary outcome(s)
Secondary outcomes
Inclusion and exclusion criteria
Interventions to be tested
Estimated sample size
Information for patients and consent
Analyses plan, concluding dummy table
How patients will be entered into the study, concealment of allocation
Duration and methods for follow-up
Data collection, including questionnaires
Trial management
Trial supervision
Publication policy
Reference

Source: Modified from Duley and Farrell [23], with permission.

10.8.5.3 Interventions

Treatments in musculoskeletal disorders, apart from surgical interventions, are classified into "drug therapy," "therapeutic exercise," "physical agents," "orthotics and devices," "education or care packages," and "others." When these intervention methods are indicated in the experimental group, it is important how comparative methods are selected in the control group. It is impossible to use the double-blind technique with placebo in musculoskeletal disorders because the comparative control is obviously different from the experimental content for the trial participants. The patients already know which treatments have been allocated and only

the assessors are unaware of the allocation results. This "open-label method" is often indicated in this field.

10.8.5.4 Outcomes

The content of outcomes used in clinical trials should be important to patients, usually the five D's: death, disease in clinical course, discomfort in symptoms, disability in activities of daily living (ADL), and dissatisfaction in QOL. There are many de facto standards for outcome measures in musculoskeletal disorders (Table 4).

Another important point in the use of outcome measures is the timing of application. Follow-up periods for outcome measures are categorized as short term (less than 6 weeks), intermediate term (6 weeks to 1 year), or long term (more than 1 year follow-up).

In an RCT, the results basically underestimate the difference between the comparative groups because the analysis is conducted with a more conservative standpoint for assessment (an intention-to-treat principle), and that will tend to dilute the estimate of the true difference. The content of interventions often fluctuates depending upon its nature, application timing, and patients' adherence. When the patients receive the strict treatment, they can expect more obvious efficacy than that reported.

10.8.5.5 How to Choose Right Outcome Measures

As an example, there is a quick-reference book that summarizes and evaluates more than 150 outcome measures for each joint of the extremities [39]. As Suk et al.'s handbook does not include the outcomes for spinal disorders, we have to have proper outcome measures for those, including neck and back problems [40, 41] (Table 4).

The necessary information for selecting outcome measures is as follows:

1. Goal of measurement
2. Nature of measurement: questionnaire, performance rating, physical properties
3. Specific population for which the instrument was developed
4. Format of measurement: number of items, response options, minimum and maximum score
5. Issues related to feasibility: time needed to perform the measurement, required equipment, and training

For clinical researchers, outcome measurements are essential for the advancement of their studies. To assess the overall quality of an outcome measure, three major elements should be considered: content of construct, psychometric evaluation, and clinical utility.

The conventional methods of studying the dimensional structure of measures are principal component analysis and factor analysis; Cronbach's α is calculated to determine the internal consistency of the dimensions. Outcome measures usually consist of one or more domains that reflect the concept of supposed construct.

TABLE 4 Widely Used or Highly Qualified Outcome Measures for Musculoskeletal Disorders

• Low back pain	
Oswestry low back pain disability questionnaire [31]	
Roland–Morris disability questionnaire (RDQ) [32, 33]	
• Rheumatoid arthritis	
Health assessment questionnaire (HAQ) [34]	
Arthritis impact measurement scale (AIMS) [35–37]	
• Osteoarthritis (hip and knee)	
Western Ontario McMaster University (WOMAC) [38]	
(In this section of joint functions, the outcome measures that scored higher than 8 are shown)	
• Joint functions [39]	
Shoulder	Disabilities of the arm, shoulder and hand (DASH)
	Flexilevel scale of shoulder function (FLEX-SF)
	Oxford shoulder score
	Shoulder instability questionnaire
	Shoulder pain and disability index (SPADI)
	Simple shoulder test (SST)
	Upper extremity function scale
Elbow	Elbow functional assessment scale (EFA)
	Liverpool elbow score
	Upper extremity function scale
Wrist/hand	Boston questionnaire (also known as Brigham and Women's carpal tunnel questionnaire)
	Cochin rheumatoid hand disability scale
	Patients rated wrist evaluation (PRWE)
	Sequential occupational dexterity assessment (SODA)
	Upper extremity function scale
Pelvis	—
Hip	AAOS hip and knee score
	Functional recovery score
	Harris hip score
	Oxford hip score
	Western Ontario and McMaster Universities OA index (WOMAC)
Knee	AAOS hip and knee score
	Activity rating scale
	Fulkerson–Shea patellofemoral joint evaluation score
	Knee outcome survey activities of daily living scale
	Knee injury and osteoarthritis outcome score (KOOS)
	Kujala patellofemoral score (also know as the AKPS—anterior knee pain scale)
	Oxford 12-item knee questionnaire
	Western Ontario and McMaster Universities OA index (WOMAC)
Ankle	Foot health status questionnaire
Calcaneus	—

It is very important to review, integrate, and consolidate the information with measuring instruments for cross-cultural use [6].

10.8.5.6 Questionnaires

The clinical utility of an outcome measure is divided into two parts: patient friendliness (acceptability) and clinician friendliness (feasibility) [42].

If we want a questionnaire to be user friendly, for both patient and clinician, and to improve clinical utility, the questionnaire has to be refined repeatedly. The content of the questionnaire must be limited to the minimum amount of information needed and be designed to be filled out easily. Even in the case of multiple choice or fill-in-the-numeral forms, there are some problems involved in interpreting the data.

Recently, the content of questionnaires that has only been determined by so-called expert staff is no longer regarded as sufficient for assessing content validity. We have to include the opinions of patients during the development process of the questionnaires in order to check content validity and face validity (these are often used interchangeably).

10.8.6 DATA ANALYSIS AND INTERPRETATION

10.8.6.1 Sample Size

The sample size of a trial is calculated based on the estimate of primary outcome [43]. I would like to show an example of how to calculate a sample size from the difference between the two means.

We have to recognize a way of denying null hypothesis, which is always associated with false-positive and false-negative rates. The former (false positive) is known as the type I error, that is, significance level α. The latter (false negative) is the type II error β, and $1 - \beta$ is the power. The sample size is given here in the condition of two groups' comparison, and randomization in equal group numbers.

$$\text{Sample size } N = \frac{4(Z\alpha + Z\beta)^2}{\Delta^2}$$

where $Z\alpha$, $Z\beta$ are standardized normal deviates for given α, β, and Δ is standardized effect size.

$Z\alpha = Z_{1-\alpha/2}$; $\alpha = 0.05$ in two-sided, then $Z_{1-\alpha/2} = Z_{0.975} = 1.96$

$Z\beta = Z_{1-\beta}$; a power of $1 - \beta = 0.9$, then $Z_{1-\beta} = Z_{0.9} = 1.645$

$\Delta = \mu A - \mu B/\sigma$; $(\mu A - \mu B)$ is difference between the two means, and σ is standard deviation of the endpoint

If you set Δ less than 0.1, approximately 0.5, or more than 1, it means a small, moderate, or large standardized effect, respectively.

Sample size is noted in the majority of articles on clinical trials. Sample size depends on whether we want much more power to test the difference among the groups or we want to estimate the precision of the confidence interval. If we have an available working hypothesis, we can calculate the necessary sample size according to other similar studies. The result of the calculations only indicates the number required for the data analysis. Usually, we have to add a few more cases, expecting that there will be some dropouts.

10.8.6.2 Intention to Treat

Intention to treat (ITT) is a conceptual principle and is not defined as a specific technical procedure. One widely used procedure is to review the data in detail when the treatment and follow-up have been completed and all the patient information has been collected, that is, at the time that all the data has been frozen. Once allocation has been done, it is the ITT principle to analyze as if the original allocation was continued, even after the intervention itself was changed. The "last observation carry-forward" method is able to cover the deficit data when the trial has been in progress for some time (more than two check points).

However, even such principles cannot be maintained at analysis in the following cases:

1. Criteria unfit for inclusion, even after allocation
2. Complete loss to follow-up

10.8.6.3 Multiple Comparisons

We must cut down the working hypotheses to one main one. In a usual trial comparing two groups, a statistical test provides a figure of P value as the boundary line for the verdict of hypothesis. If the calculated P value for the primary endpoint is less than the predefined line (usually $P < 0.05$), we will deny the null hypothesis of no difference between the groups. However, if we use more than one endpoint, there are more than two comparisons arising (sometimes this situation becomes much more complex with multiple comparisons). In such a situation, a false-positive rate is no longer $\alpha \times 100\%$ and increases according to the number of multiple comparisons.

10.8.6.4 Post Hoc Analysis and Subgroup Analysis

It is the nature of things that the therapy-responder group has a better prognosis and the nonresponder group has a worse result. That is a typical post hoc analysis. Grouping allocation should be provided at the beginning and after disclosing the result, retrograde grouping is not permitted.

At times, clinicians like to know whether certain types of patients show more obvious benefits from interventions than others, knowing that a sample group of interest is a mixture of various types of patients. This is the subgroup analysis, for example, mild or severe cases, young or old patients. Some subgroups could easily reveal positive effects by chance, even if the overall trial results are negative. A common mistake is to mix the P value of more than 0.05 in all data-combined analysis and less than 0.05 in one or more subgroup(s). This may lead to the false conclusion that a certain subgroup has a favored result by chance, in spite of the fact that the true result is negative. When you plan a subgroup analysis at the design stage, advance adjustment in sample size should be considered.

10.8.6.5 Data Synthesis and Combined Result

If there are several numbers of articles available for review, systematic review, is regarded as an "infrastructure" of the information system supporting EBM practice.

In regards to treatment, the pathway of systematic review is very logical; so the results provide the most accurate and authoritative guidelines for therapy. To avoid potential problems in interventional studies (clinical trials), only RCTs should be included, with complete follow-up information, blinded outcome assessment, and analysis based on intention to treat. This is applicable to an epidemiological research in observational studies, but it is still unable to solve all the problems of biases.

Meta-analysis is defined as a statistical synthesis of the numerical results of some trials (quantitative systematic review) that all examined the same question. It is a type of research that attempts to reanalyze and combine the results already reported, mainly as RCTs.

The assessment of methodological quality for meta-analysis has been recognized to be very important because even the conduct of an RCT is no guarantee of unbiased outcomes from such a study. Even meta-analysis and systematic reviews should be scrutinized carefully and analyses based on small studies should especially be treated with caution.

10.8.6.6 Side Effects

We have to pay sufficient attention to avoid adverse effects. In the case of surgery, examples of adverse side effects are nerve injury, infection, bleeding, secondary osteoarthritis, and even death. Some side effects occur even in conservative therapy.

We cannot say that treatment side effects are out of the question, even if there is no obvious difference in the frequency and degree of adverse effects between the treatment group and the control group in a certain RCT. An RCT is not an appropriate design to assess side effects because an RCT basically suppresses the number of patients to a minimum and the incidence of side effects is usually low. The role of the trial steering committee and the data-monitoring committee should be established and the range of their responsibilities defined. These committees have to take necessary actions such as early stopping, considering the results of the interim analysis, or the frequency and content of side effects. I realize that relatively small size trials do not often have the side effects and other problems that would stop a trial, but large multicentered trials have the latent possibility of such troubles.

10.8.6.7 Reporting Clinical Trials

After completing a clinical trial, an important step is to report the results, whether positive, negative, or equivocal. Selective reporting, in which only positive studies are published, often distorts a true situation; this practice is called "publication bias." When meta-analysis is performed as a clinical trial overview, it is definitely important to include all relevant unpublished trials to gain overall results.

The International Committee of Medical Journal Editors, which consists of several leading medical core journals, decided that clinical trials should be registered in advance [44, 45]. If clinical trials are not registered in advance, journals might refuse to publish a report of a trial. This is an important step in avoiding publication bias and other inappropriate analyses.

Article writing is regarded as a form of communication or "dialog," even in the case of a scientific journal; that is the object of our communication. The writer of an article must be aware of the target audience. Consolidated Standards of

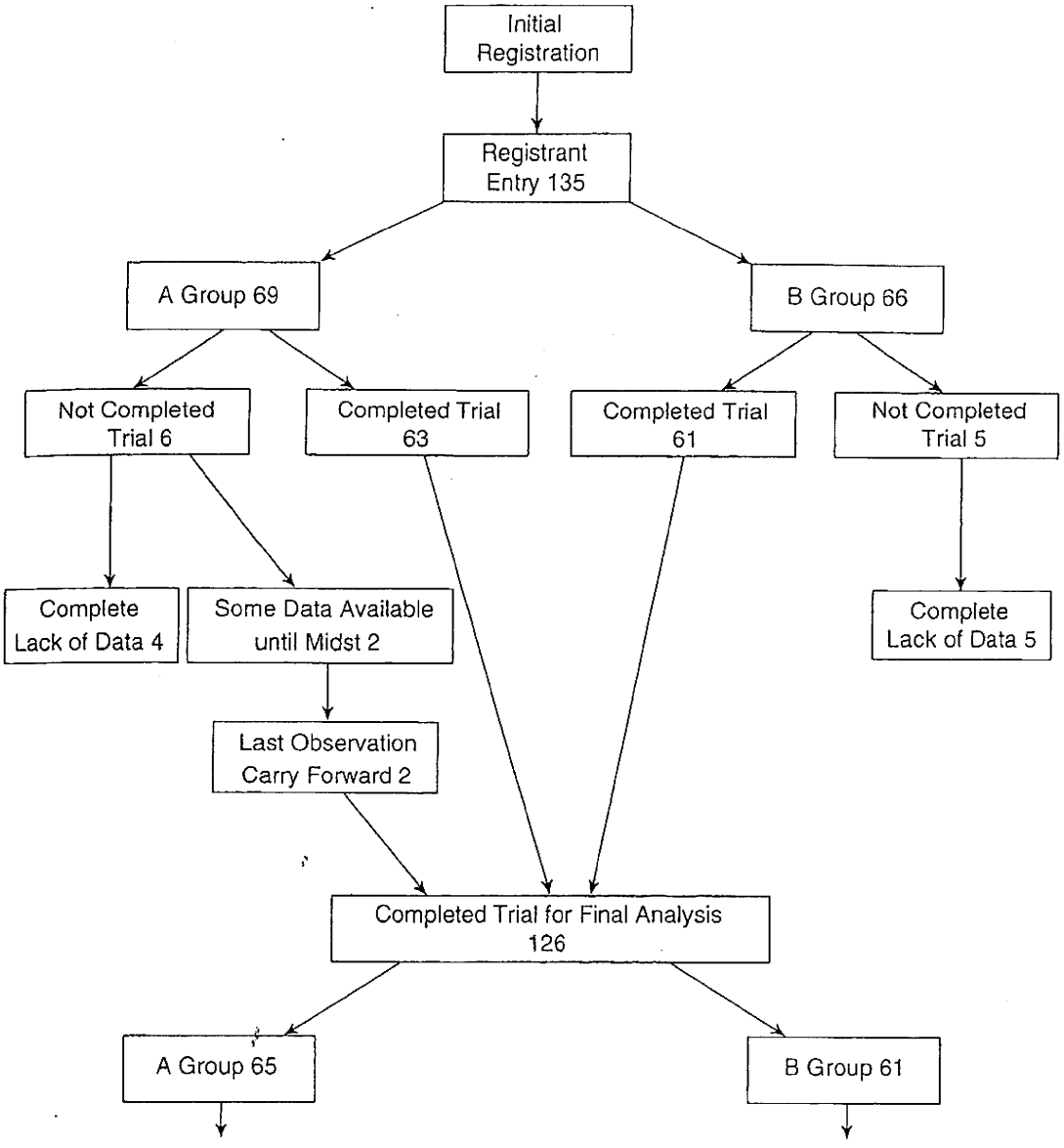


FIGURE 1 *Example flowchart of trial according to CONSORT statement.*

Reporting Trials (CONSORT) is one of the examples of such a checklist or flowchart to be used, as well as the instructions for improving the quality of reporting randomized controlled trials when a paper is to be submitted [46, 47] (Fig. 1). Also suggested through discussions at a conference was a guideline for reporting systematic reviews or meta-analysis referred to as the Quality of Reporting of Meta-analysis (QUOROM). The discussions resulted in the creation of the QUOROM Statement, which consists of a checklist and flowchart [48].

10.8.7 RCT AND FUTURE DIRECTIONS

At the close of this section, I would like to talk about how to apply the results of an RCT to daily practice, as well as mention the potential problems when reaching the limit of a statistical outlook.

10.8.7.1 Impact from RCT to Daily Practice

Medical practice is a life-long, continuous process of self-learning, and, as clinicians, we are required to keep up-to-date on various medical developments. Evidence-based medicine is our way of integrating individual clinical expertise with the best available evidence to assist us in making decisions about each patient's care. Reported results of RCTs make it possible to cover the majority of our activities systematically, from daily practice for patient care to writing and reading scientific papers.

Evidence-based medicine is regarded as a new paradigm in medical practice, equal to the Human Genome Project. However, there is still much confusion and misunderstanding about the concept and content of EBM; it is often limited to searching the literature and reading articles, and to serving cost cutters and suppressing clinical freedom. The use of clinical guidelines or the managed care system is seen as intimidating the discretion of doctors in clinical practice. EBM is also seen as the fashionable trend of a group of medical academics armed with epidemiological and statistical jargon, but there are more serious basic problems behind such a statistical paradigm.

10.8.7.2 Internal Validity and External Validity

In clinical practice we are always surrounded by a condition called the "gray zone" [49]. Of course, there are many factors involved in this obscurity. I would like to focus closer attention on a statistical outlook of the world around us, that is, internal validity and external validity.

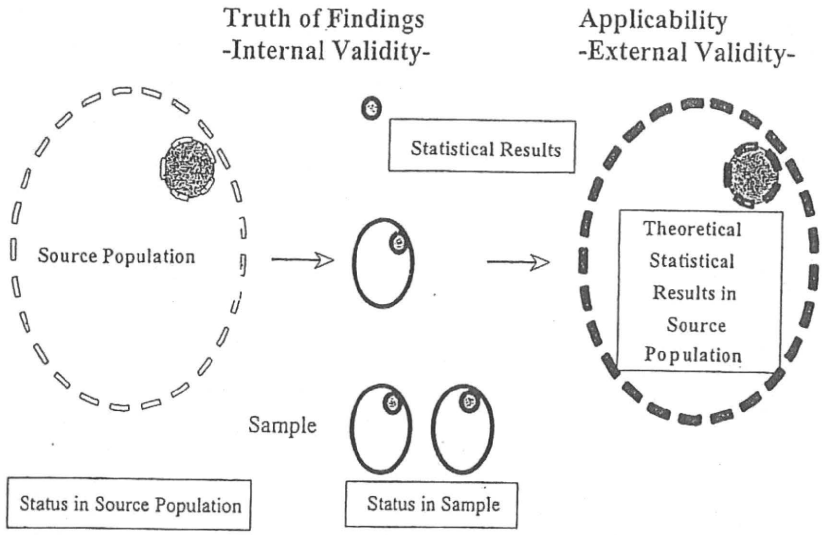
These are concepts relating to the extrapolation of the relationship between sample and source populations. In an experimental trial like an RCT, internal validity is high because the difference between the experimental group and the control group is thoroughly controlled only for the target issue, and by chance. However, external validity is low when the results from the sample with several specific premises are brought into a general finding in the mother population. On the contrary, in an observational study, such as a cohort study, there are many confounding factors involved and internal validity is low in surveyed samples. External validity in an observational study is high because a real situation in a resource population reflects just what it is (Fig. 2).

10.8.7.3 Quantitative Study and Qualitative Study

Clinical trials cover a wide range of medical fields and use very diverse methods. The data used is basically divided into quantitative, such as laboratory data, and qualitative, based on verbal information collected (mainly used in psychology or nursing science at present). Behind the methods of investigation and analysis in these quantitative and qualitative approaches, there is an obvious difference in concepts [50].

The general concept for study design consists of the following characteristics (Table 5):

1. Aims of study
2. Methods of outcome measurements



Observational Study
Cross-sectional Study or Cohort Study — Internal Validity: low; External Validity: high

Interventional Study
Randomized Controlled Trial, Blind & Placebo-controlled — Internal Validity: high; External Validity: low

FIGURE 2 *Relationship between sample and mother population.*

TABLE 5 Comparison Between Quantitative Research and Qualitative Research

	Quantitative Research	Qualitative Research
Study design	Intervention	Observation
Methods	Survey, experiment	Interview
Question setting	Enumeration	Classification
Reasoning	Deductive	Inductive
Sampling method	Statistical model	Theoretical model
Closeness to the truth	Internal validity	Credibility
Applicability/generalizability	External validity	Concept transferability
Consistency	Reliability	Dependability
Neutrality	Objectivity	"Confirmability"

3. Selection of subjects
4. Analytic methods and provided results

In a conventional quantitative study, the main strategy is to make up a working hypothesis and test it. An RCT is a typical example of this type of strategy and is inseparable from biostatistics for data analysis. To generalize the results from samples, such as the rejection of null hypothesis in the majority of confirmatory studies, requires estimating and testing as a basic procedure. The advance of computer technology makes possible the use of multivariate analysis and the ability to search and verify potential factorial structure, even without a working hypothesis.

This technique makes it easy to analyze latent factorial structure, which occurs in some observational studies, from a calculated relationship of how the weight distribution among explanatory variables accounts for the tendency of objective variables. Therefore, this exploratory study makes it possible to investigate and verify risk factors or to produce a model of a certain phenomenon. The distinction between the term "quantitative" study and "qualitative" study will be more obscure once the quantitative studies move toward an exploratory nature from a confirmatory one for testing hypothesis.

On the contrary, a qualitative study does not always follow the deductive logic of applying the result from a sample to the mother population. Researchers try to categorize the verbal information received from subjects to establish a coding system for the data and generalize a new concept in order to build a new theory or model. It is good enough to select the subjects that fit the aim of a study as one likes. It would be possible to produce an epochal qualitative study that could produce a unique, clinically significant theory that would never have been yielded by a confirmatory study using hypotheses testing. A qualitative study is able to make progress in building up a new and significant theoretical system [51]. However, good or bad, theory is totally dependent upon the capability of the researcher.

10.8.7.4 Further Suggestions

If a certain intervention shows obviously successful results for an otherwise lethal condition, we do not require RCTs and do not wait for more studies to be conducted. We can get the evidence from the basic sciences, properly designed follow-up studies, or proper cross-sectional trials.

We need to be careful not to become so skeptical as to assume that if a study is not an RCT it would have no value or use. We also need to recognize the limitations of a statistical approach through RCTs. At each step, we continue our efforts to track down the best evidence to answer our clinical questions and promote cost-effective prevention and treatment for musculoskeletal disorders.

REFERENCES

1. Akai, M. (2002), Evidence-based medicine for orthopedic practice, *J. Orthop. Sci.*, 7, 731–742.
2. Akai, M., Doi, T., Fujino, K., et al. (2005), An outcome measure for Japanese people with knee osteoarthritis, *J. Rheumatol.*, 32, 1524–1532.
3. BJD on line (Bone and Joint Decade's musculoskeletal Portal); http://www.boneandjoint-decade.org/.
4. McDowell, I., and Newell, C. (1996), *The Theoretical and Technical Foundations of Health Measurement. Measuring Health; A Guide to Rating Scales and Questionnaires*, 2nd ed., Oxford University Press, New York, pp. 10–46.
5. Guyatt, G. H., Kirshner, B., and Jaeschke, R. (1992), Measuring health status; What are the necessary measurement properties? *J. Clin. Epidemiol.*, 45, 1341–1345.
6. Dekker, J., Dallmeijer, A. J., and Lankhorst, G. J. (2005), Clinimetrics in rehabilitation medicine; current issues in developing and applying measurement instruments, *J. Rehabil. Med.*, 37, 193–201.

7. Wolfson, A. M., Doctor, J. N., and Burns, S. P. (2000), Clinician judgments of functional outcomes; how bias and perceived accuracy affect rating, *Arch. Phys. Med. Rehabil.*, 81, 1567–1574.
8. Guyatt, G. H., Feeny, D. H., and Patrick, D.L. (1993), Measuring health-related quality of life, *Ann. Int. Med.*, 118, 622–629.
9. Clancy, C. M., and Eisenberg, J. M. (1998), Outcomes research; measuring the end results of health care, *Science*, 282, 245–246.
10. World Health Organization (2001), *International Classification of Functioning, Disability and Health*, WHO, Geneva.
11. Stucki, G., and Grimby, G. (2004), Applying the ICF in medicine, *J. Rehabil. Med.*, 36 (suppl 44), 5–7.
12. Irrgang, J. J., and Anderson, A. F. (2002), Development and validation of health-related quality of life measures for the knee, *Clin. Orthop.*, 402, 95–109.
13. Zelen, M. (1979), A new design for randomized clinical trials, *N. Engl. J. Med.*, 300, 1242–1245.
14. Zelen, M. (1990), A new design for randomized consent designs for clinical trials; an update, *Stat. Med.*, 9, 645–656.
15. ter Riet, G., Kleijnen, J., and Knipschild, P. (1990), Acupuncture and chronic pain; A criteria-based meta-analysis, *J. Clin. Epidemiol.*, 43, 1191–1199.
16. Koes, B. W., Bouter, L. M., Beckerman, H., et al. (1991), Physiotherapy exercises and back pain; A blinded review, *BMJ*, 302, 1572–1576.
17. Koes, B. W., Assendelft, W. J. J., van der Heijden, G. J. M. G., et al. (1991), Spinal manipulation and mobilisation for back and neck pain; A blinded review, *BMJ*, 303, 1298–1303.
18. Koes, B. W., Bouter, L. M., and van der Heijden, G. J. M. G. (1995), Methodological quality of randomized clinical trials on treatment efficacy in low back pain, *Spine*, 20, 228–235.
19. Akai, M., and Hayashi, K. (2002), Effect of electrical stimulation on musculoskeletal systems; A meta-analysis of controlled clinical trials, *Bioelectromagnetics*, 23, 132–143.
20. Akai, M., Kawashima, N., Kimura, T., et al. (2002), Electrical stimulation as an adjunct to spinal fusions; A meta-analysis of clinical controlled trials, *Bioelectromagnetics*, 23, 496–504.
21. Machin, D., Day, S., and Green, S. Eds. (2004), *Textbook of Clinical Trials*, Wiley, West Sussex, UK.
22. Bowling, A. (2001), *Measuring Disease; A Review of Disease-Specific Quality of Life Measurement Scales*, 2nd ed., Open University Press, Buckingham, UK.
23. Duley, L., and Farrell, B. Eds. (2002), *Clinical Trials*, BMJ Books, London.
24. McDowell, I., and Newell, C. (1996), *The Theoretical and Technical Foundations of Health Measurement. Measuring Health; A Guide to Rating Scales and Questionnaires*, 2nd ed., Oxford University Press, New York.
25. Riegelman, R. K. (2005), *Studying a Study and Testing a Test; How to Read the Medical Evidence*, 5th ed., Lippincott Williams and Wilkins, Philadelphia.
26. Suk, M., Hanson, B. P., Norvell, D. C., et al. (2005), *AO Handbook; Musculoskeletal Outcomes Measures and Instruments; 155 Instruments Evaluated and Assessed*, AO Publishers, Thieme.
27. Resource Center for Randomized Trials; http://www.rcrt.ox.ac.uk/.
28. An introduction to clinical trials; http://www.clinicaltrials.gov/.
29. Farrell, B., and Spark, P. (2002), Building resources for randomized trials, in Duley, L., and Farrell, B., Eds., *Clinical Trials*, BMJ Books, London, p. 84.

30. Fayers, P. M., and Machin, D. (1995), Sample size; how many patients are necessary? (editorial review), *Br. J. Cancer*, 72, 1–9.
31. Fairbank, J. C. T., Couper, J., Davies, J. B., et al. (1980), The Oswestry low back pain disability questionnaire, *Physiotherapy*, 66, 271–273.
32. Roland, M., and Morris, A. (1983), A study of the natural history of back pain. Part II. Development of a reliable and sensitive measure of disability in low-back pain, *Spine*, 8, 141–144.
33. Roland, M., and Fairbank, J. (2000), The Roland-Morris disability questionnaire and the Oswestry disability questionnaire, *Spine*, 25, 3115–3124.
34. Fries, J. F., Spitz, P., Kraines, R. G., et al. (1980), Measurement of patient outcome in arthritis, *Arthritis Rheum.*, 23, 137–145.
35. Meenan, R. F., Gertman, P. M., and Mason, J. H. (1980), Measuring health status in arthritis; The arthritis impact measurement scales, *Arthritis Rheum.*, 23, 146–152.
36. Meenan, R. F., Gertman, P. M., Mason, J. H., et al. (1982), The arthritis impact measurement scales; Further investigations of a health status measure, *Arthritis Rheum.*, 25, 1048–1053.
37. Meenan, R. F., Mason, J. H., Anderson, J. J., et al. (1992), AIMS2; The content and properties of a revised and expanded arthritis impact measurement scales health status questionnaire, *Arthritis Rheum.*, 35, 1–10.
38. Bellamy, N., Buchanan, W. W., Goldsmith, C. H., et al. (1988), Validation study of WOMAC; A health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee, *J. Rheumatol.* 15, 1833–1840.
39. Suk, M., Hanson, B. P., Norvell, D. C., et al. (2005), *AO Handbook; Musculoskeletal Outcomes Measures and Instruments; 155 Instruments Evaluated and Assessed*, AO Publishers, Thieme, pp. 37–40, 47–405.
40. Bombardier, C. (2000), Outcome assessments in the evaluation of treatment of spinal disorders; summary and general recommendations, *Spine*, 25, 3100–3103.
41. Stratford, P. W., Binkley, J., Solomon, P., et al. (1994), Assessing change over time in patients with low back pain, *Phys. Ther.*, 74, 528–533.
42. Suk, M., Hanson, B. P., Norvell, D. C., et al. (2005), *AO Handbook; Musculoskeletal Outcomes Measures and Instruments; 155 Instruments Evaluated and Assessed*, AO Publishers, Thieme, pp. 20–24.
43. Machin, D., Campbell, M. J., Fayers, P. M., et al. (1997), *Sample Size Tables for Clinical Studies*, 2nd ed., Blackwell Science, Oxford, pp. 40–78.
44. DeAngelis, C. D., Drazen, J. M., Frizelle, F. A., et al. (2004), Clinical trial registration; a statement from the international committee of medical journal editors. *JAMA*, 292, 1363–1364.
45. DeAngelis, C. D., Drazen, J. M., Frizelle, F. A., et al. (2005), Is this clinical trial fully registered? A statement from the international committee of medical journal editors, *JAMA*, 293, 2927–2929.
46. Begg, C., Cho, M., Eastwood, S., et al. (1996), Improving the quality of reporting of randomized controlled trials; The CONSORT statement, *JAMA*, 276, 637–639.
47. Moher, D., Schulz, K. F., Altman, D. G., for the CONSORT group (2001), The CONSORT statement; revised recommendations for improving the quality of reports of parallel-group randomized trials, *Lancet*, 357, 191–194.
48. Moher, D., Cook, D. J., Eastwood, S., et al. (1999), Improving the quality of reporting of meta-analyses of randomised controlled trials; The QUOROM statement, *Lancet*, 354, 1896–1900.

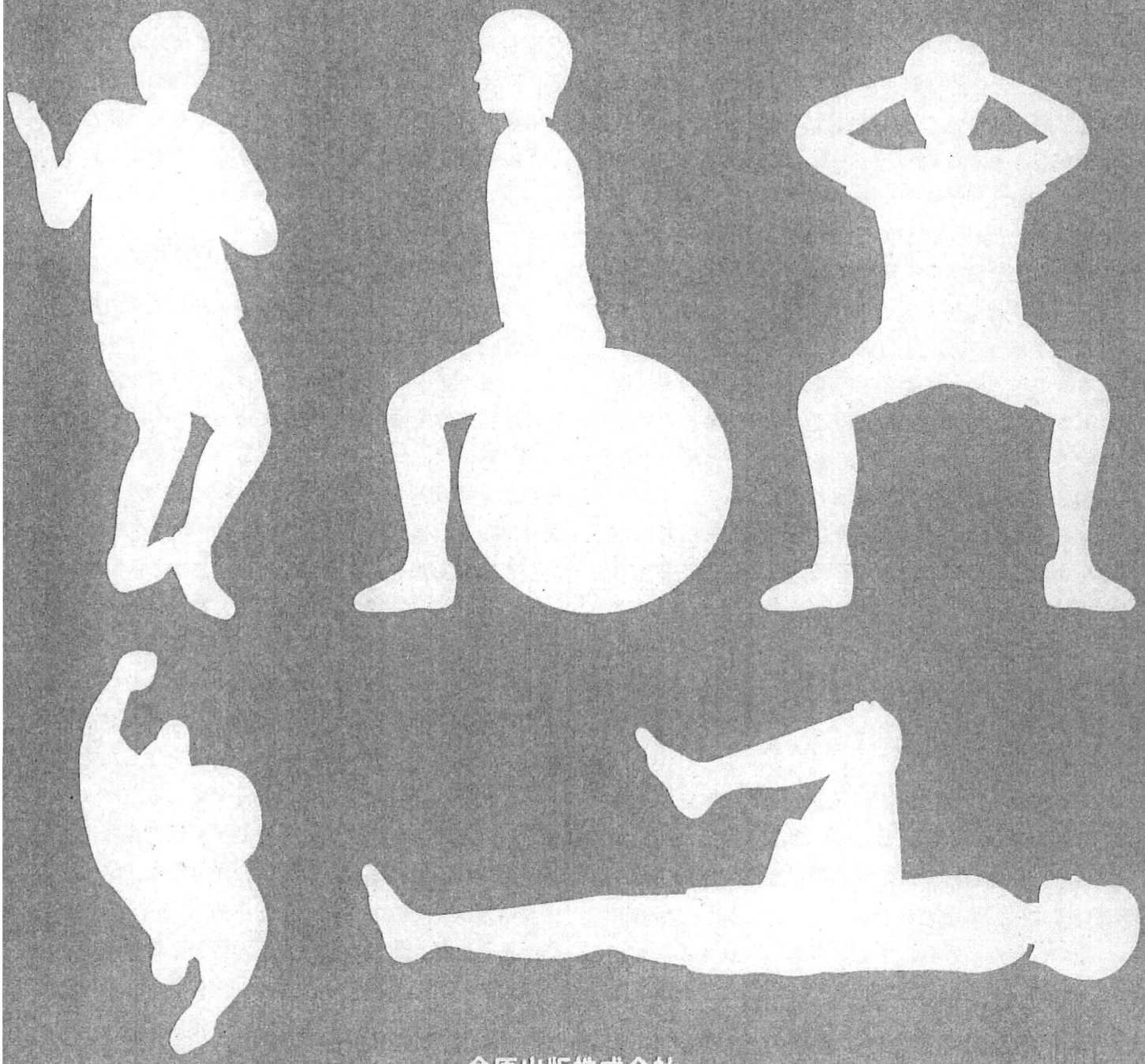
49. Naylor, C. D. (1995), Grey zones of clinical practice; Some limits to evidence-based medicine, *Lancet*, 345, 840–842.

50. Greenhalgh, T., and Taylor, R. (1997), Papers that go beyond numbers (qualitative research), *BMJ*, 315, 740–743.

51. Öhman, A. (2005), Qualitative methodology for rehabilitation research, *J. Rehabil. Med.*, 37, 273–280.

運動器慢性疾患に対する
運動療法

順天堂東京江東高齢者医療センター副院長
黒澤 尚 編



金原出版株式会社

10. 慢性運動器疾患の評価法

赤居正美*1　岩谷　力*2　土肥徳秀*3

骨折や軟部組織損傷などの外傷は急性発症であり，手術治療がそのまま適応されることも多い。これに反し，加齢変化を背景に5～10年をかけて緩徐に進行する慢性運動器疾患の場合には，まず一定症状までは保存療法が適応となる。もちろん，最終的に手術療法が適応されることもあるが，これらの治療選択の前提に，障害の機能面や回復の余地からみた評価分別がなされる。したがって，慢性的な経過をたどる症例での療法の適応と限界についての判断には，何らかの障害度評価が不可欠である。

この場合，対象となる患者の障害度をどのように測るかということが，治療開始時期および介入の有効性を論ずるうえで問題となる。すなわち臨床医は診断に基づき，その時点で最善と判断した治療手段を実行し，良好な反応・経過を期待する。このとき，医療介入の効果を測るものがアウトカムであり，健康状態の違いを介入前後で比較する[1]。また実績のある治療法も，どの段階から適応になるかの時期判断は必要である。

こうしたアウトカム尺度を用いて統一的な扱いをすることで，治療や医学的介入の質，有効性を定量的に評価できる[2]。現在，実行された医療介入のもたらす効果判定，ランダム化比較試験の中心となる考え方である[3)4)]。

Ⅰ. 評価指標

従来からの評価法には罹患率，死亡率，治癒率などがあった。集団としての患者を議論している限り，これら評価法の定義の明確性，重要性については疑問の余地はなかった[5]。しかし，疾病構造の変化が進行し，生活習慣病に代表されるような慢性経過をたどり，基本的に治癒しない病気はどう考えるか疑問が生じる。また，多くの運動器疾患では治療目的は機能向上が中心であって，生死とは直接結びつきにくい場合にはどうするか，との難しい問題もある。

そこで，患者側の健康観，満足度など主観的指標も取り込み，生活の質（QOL）評価を中心にした，患者に中心を置くアウトカムという考え方が登場した[2)6)]。こうした患者立脚型アウトカム評価に用いられる尺度では，基本的に病気ではなく病人を測るという立場をとる。すなわち，集団としての死亡率，罹患率などではなく，患者個々のレベルでの健康状態を評価する[5)6)]。

Ⅱ. 運動・動作・活動の関連

この際，重要なことは何を物差しにしてその量的な変化を測るかである[7]。

運動器疾患の注目対象となる我々の日常生活活動は四肢体幹を動かすさまざまな動作から成り立つが，動作は立つ，歩く，食べるなど目的のある複合的・連続的運動からなっている。この場合，運動は空間座標上の位置変化，位置，時間，方向，力など物理変数として測定できる。しかし，動作の複合体である日常生活は，動作の達成度や自立性，遂行に必要な時間などに注目して数値化を図っている。運動，動作，日常生活活動各レベルの状況をそれぞれ数値化して捉えることができたなら，これら各レベルの関連や，問題を生じた際の原因解明などに役立てられるはずである[2]。

では，動作や活動に関する評価尺度は直接測定が可能なのであろうか。直接測定が可能な対象には，臨床検査で汎用される生理学的，生化学的検査所見がある。新しい検査技術の開発と精度の向上により，次々と新たな情報入手が可能になっている。しかし問題は，これらが患者の求める回答に直接結びついているか疑問な点にある。患者や

*1 Masami AKAI　*2 Tsutomu IWAYA：国立障害者リハビリテーションセンター，*3 Tokuhide DOI：福岡クリニック

家族は「また歩けるようになるのか」「術後には元のような生活に戻れるのか」などの疑問に対する回答を求めているが，検査データはその説明の裏づけにはならない。

Ⅲ. アウトカムと構成概念

上記の数値化の際，測る対象が物理・化学現象であれば，基本的単位とそれらの組み合わせによって，全て直接計測が可能である。ところが，人間の機能，能力の多くは，想定された概念そのものを直接測定することはできない。測定ができなければ，達成度にみられるような仮想的な構成概念（construct）を作って測定することになる。構成概念とは「とりあえずその存在を仮定することによって複雑に込み入った現象を比較的単純に理解することを目的として構成した概念」とされる。

構成概念の内容を日常生活面や社会参加に持ち込んで評価を行う。多くは多元的であり，ある目的に添って操作的に定義されている[8]。したがって，この構成概念測定が意味のある，妥当なものか，測定に信頼が置けるのかといった計量心理学的検証が不可欠になる[5]。

Ⅳ. 評価尺度の計量心理学的検討

このような考え方から，数多くの評価尺度が提案されてきており（表1～表3），計量心理学的立場からおのおの検証のうえ，採点を行っている書籍もある[9]。外国で開発された尺度は単に翻訳するだけでは不十分であり，日本人において標準化，検証が行われなければならない。

米国の食品医薬品局（FDA）がガイドラインを検討中であり，計量心理学上での概念がよりわかりやすい形で整理されつつある[10]。その中心となる枠組みは以下のとおりである。

・信頼性（reliability）

対象概念を変更しなかった際の評価の経時的安定性。

内的整合性：領域中の項目が相関しているか，内的整合性（α係数）などから検討。

検者間再現性：複数の検者（インタビュアー）が評価を実施する際の回答間での一致。

・妥当性（validity）

内容妥当性：項目および回答の選択が妥当であり，領域ないし概念が包括的かをチェック。

構成概念妥当性：おのおのの項目，領域，概念の間にある関連性が評価手法自体や仮説についての概念構造から予測される内容と一致するか。

・反応性（responsiveness）

患者に変化がない場合，評価スコアも変わらないか，効果サイズの計算から患者に明らかな変化を生じた際に予測の方向でスコアが変化しているか。変化検出能力は時間変化に特異的であることが必要。

・解釈性（interpretability）

群間の平均スコア差であり，治療効果について明白なエビデンスが提示される。分布に基づくアプローチ，臨床的または非臨床的な手がかり，経験に基づく規定，または各手法の組み合わせによる評価での経験に基づく。臨床的手がかりを用いたminimal important difference（MID）の定義は，minimal clinically important difference（MCID）と呼ばれることもある。

Ⅴ. 構成概念の構造検討

こうした手法の中心をなす構成概念の構造についての解析手法としては，以下がよく用いられる[1]。

A. 主成分分析（principal component analysis）

測定項目のセットが仮定上の構成概念を定義する程度を計る方法。変数を変換して新しいより少数の総合的特性値を求め，多くは因子数の推定に利用する。

B. 因子分析（factor analysis）

ある多変量データを少数の共通因子の一次結合として表し，変数間の相関関係を規定している潜在因子を探る方法。

表1. JLEQ (Japan Low back pain Evaluation Questionnaire) の質問項目

腰の状態についての質問票 (JLEQ)

Ⅰ. 腰の痛みの程度

次の線は「あなたの腰の痛みの程度」をおたずねするものです。左の端を「痛みなし」，右の端をこれまでに経験した「最も激しい痛み」としたときに，この数日間のあなたの痛みの程度はどのあたりでしょうか。
線の上でこのあたりと思われるところに×印をつけてください。

痛みなし ——— これまでに経験した最も激しい痛み

Ⅱ. 日常生活動作と腰の痛み

この「数日間のあなたの腰の痛み」についてお聞きします。あてはまる回答を1つ選び，□に√をつけてください。

1. この数日間，あお向けで寝ているとき腰が痛みますか。
□ 痛くない □ 少し痛い □ 中程度痛い □ かなり痛い □ ひどく痛い
2. この数日間，朝，起きて動き出すとき腰が痛みますか。
□ 痛くない □ 少し痛い □ 中程度痛い □ かなり痛い □ ひどく痛い
3. この数日間，椅子に腰かけているとき腰が痛みますか。
□ 痛くない □ 少し痛い □ 中程度痛い □ かなり痛い □ ひどく痛い
4. この数日間，立ち上がるときやしゃがみこむとき腰が痛みますか。
□ 痛くない □ 少し痛い □ 中程度痛い □ かなり痛い □ ひどく痛い
5. この数日間，立っているとき腰が痛みますか。
□ 痛くない □ 少し痛い □ 中程度痛い □ かなり痛い □ ひどく痛い
6. この数日間，前かがみになるとき腰が痛みますか。
□ 痛くない □ 少し痛い □ 中程度痛い □ かなり痛い □ ひどく痛い
7. この数日間，腰をそらすとき腰が痛みますか。
□ 痛くない □ 少し痛い □ 中程度痛い □ かなり痛い □ ひどく痛い

Ⅲ. 腰の痛みによる生活上の問題

この「数日間のあなたの腰の痛みによる生活上の問題」についてお聞きします。あてはまる回答を1つ選び，□に√をつけてください。

8. この数日間，同じ姿勢を続けるのはどの程度つらいですか。
□ つらくはない □ 少しつらい □ ときどき姿勢を変えないとつらい □ しばしば姿勢を変えないとつらい □ つねにつらくて，じっとしていられない
9. この数日間，腰痛のため，寝返りはどの程度困難ですか。
□ 困難はない □ 少し困難 □ 中程度困難 □ かなり困難 □ ひどく困難
10. この数日間，腰痛のため，朝，起き上がるのはどの程度困難ですか。
□ 困難はない □ 少し困難 □ 中程度困難 □ かなり困難 □ ひどく困難
11. この数日間，腰痛のため，からだを動かすのはどの程度困難ですか。
□ 困難はない □ 少し困難 □ 中程度困難 □ かなり困難 □ ひどく困難
12. この数日間，腰痛のため，椅子や洋式トイレからの立ち上がりはどの程度困難ですか。
□ 困難はない □ 少し困難 □ 中程度困難 □ かなり困難 □ ひどく困難
13. この数日間，腰痛のため，階段の昇り降りはどの程度困難ですか。
□ 困難はない □ 少し困難 □ 中程度困難 □ かなり困難 □ ひどく困難
14. この数日間，腰痛のため，クツ下やストッキングをはくのはどの程度困難ですか。
□ 困難はない □ 少し困難 □ 中程度困難 □ かなり困難 □ ひどく困難
15. この数日間，腰痛のため，ズボンやパンツの上げ下ろしはどの程度困難ですか。
□ 困難はない □ 少し困難 □ 中程度困難 □ かなり困難 □ ひどく困難

(次頁につづく)

(前頁よりつづき)

16. この数日間，腰痛のため，床にある3～4キログラム(1升ビン2本，または2リットル入りのペットボトル2本)程度のものを持ち上げようとするのはどの程度困難ですか。

□ 困難はない　□ 少し困難　□ 中程度困難　□ かなり困難　□ ひどく困難

17. この数日間，腰痛のため，腰を捻って後ろのものをとろうとするのはどの程度困難ですか。

□ 困難はない　□ 少し困難　□ 中程度困難　□ かなり困難　□ ひどく困難

18. この数日間，腰痛のため，戸外を歩くのがどの程度に制限されていますか。

□ 1時間以上歩ける　□ 30分程度は歩ける　□ 10～15分程度しか歩けない　□ 2，3分程度しか歩けない　□ ほとんど戸外を歩けない

19. この数日間，腰痛のため，簡単な作業や家事(ものを片づける，食事の準備をするなど)はどの程度つらいですか。

□ つらくない　□ 少しつらい　□ 中程度つらい　□ かなりつらい　□ ひどくつらい

20. この数日間，腰痛のため，負担のかかる作業や家事(重いものを運ぶ，家の外の掃除など)はどの程度つらいですか。

□ つらくない　□ 少しつらい　□ 中程度つらい　□ かなりつらい　□ ひどくつらい

21. この数日間，腰痛のため，横になって休みたいと思いましたか

□ 思わなかった　□ たまに思った　□ ときどき思った　□ しばしば思った　□ いつも思っていた

22. この数日間，腰痛のため，仕事や学校，ふだんの作業や家事を差しひかえたいと思いましたか。

□ 思わなかった　□ たまに思った　□ ときどき思った　□ しばしば思った　□ いつも思っていた

23. この数日間，腰痛のため，夜よく眠れないことがありましたか。

□ 腰痛のためによく眠れないことはなかった　□ 一晩ほどよく眠れないことがあった　□ よく眠れるときと眠れないときが半々だった　□ よく眠れない夜の方が多かった　□ 毎晩のようによく眠れなかった

24. この数日間の腰の状態からみて，遠くへの外出はむずかしいと思いますか。

□ むずかしくないと思う　□ 少しむずかしいと思う　□ 中程度むずかしいと思う　□ かなりむずかしいと思う　□ 全く無理だと思う

IV. 健康・精神状態など

この1か月間の状態について，お聞きします。あてはまる回答を1つ選び，□に✓をつけて下さい。

25. この1か月間，腰痛のため，近所への外出を差しひかえたりしましたか。

□ 差しひかえることはなかった　□ 1，2回差しひかえた　□ ときどき差しひかえた　□ しばしば差しひかえた　□ 全く外出しなかった

26. この1か月間，腰痛のため，ふだんしていること(友人とのつきあい，スポーツ活動，趣味活動など)を制限しましたか。

□ 制限しなかった　□ 少し制限した　□ 半分程度制限した　□ かなり制限した　□ 全くやめていた

27. この1か月間，腰痛のため，職場や学校を休日以外に休んだり，ふだんしている家事を休んだりしましたか。

□ 休まなかった　□ 1～3日休んだ　□ 数日以上休んだ　□ 半分程度休んだ　□ ほとんど休んだ

28. この1か月間，腰痛のため気分がすぐれないことがありましたか。

□ 気分がすぐれないことはなかった　□ たまに気分がすぐれなかった　□ ときどき気分がすぐれなかった　□ 気分がすぐれないときが多かった　□ つねに気分がすぐれなかった

29. この1か月間，腰痛はあなたの精神状態に悪く影響していると思いますか。

□ 全く影響はない　□ 少し悪い影響がある　□ 中程度悪い影響がある　□ かなり悪い影響がある　□ ひどく悪い影響がある

30. この1か月間，腰痛はあなたの健康状態に悪く影響していると思いますか。

□ 全く影響はない　□ 少し悪い影響がある　□ 中程度悪い影響がある　□ かなり悪い影響がある　□ ひどく悪い影響がある

ご記入もれがないか，もう一度ご確認ください。

表2. JKOM (Japanese Knee Osteoarthritis Measure) の質問項目

膝の状態についての質問票 (JKOM)

Ⅰ. 膝の痛みの程度

次の線は痛みの程度をおたずねするものです。左の端を「痛みなし」，右の端をこれまでに経験した「最も激しい痛み」としたときに，この数日間のあなたの痛みの程度はどのあたりでしょうか。
線の上でこのあたりと思われるところに×印をつけてください。

痛みなし ―――――――――― これまでに経験した最も激しい痛み

Ⅱ. 膝の痛みやこわばり

この数日間のあなたの膝の状態についてお聞きします。あてはまる回答を1つ選び，□に√をつけてください。

1. この数日間，朝，起きて動き出すとき膝がこわばりますか。
□ こわばりはない □ 少しこわばる □ 中程度こわばる □ かなりこわばる □ ひどくこわばる
2. この数日間，朝，起きて動き出すとき膝が痛みますか。
□ 全く痛くない □ 少し痛い □ 中程度痛い □ かなり痛い □ ひどく痛い
3. この数日間，夜間，睡眠中に膝が痛くて目がさめることがありますか。
□ 全くない □ たまにある □ ときどきある □ しばしばある □ 毎晩ある
4. この数日間，平らなところを歩くとき膝が痛みますか。
□ 全く痛くない □ 少し痛い □ 中程度痛い □ かなり痛い □ ひどく痛い
5. この数日間，階段を昇るときに膝が痛みますか。
□ 全く痛くない □ 少し痛い □ 中程度痛い □ かなり痛い □ ひどく痛い
6. この数日間，階段を降りるときに膝が痛みますか。
□ 全く痛くない □ 少し痛い □ 中程度痛い □ かなり痛い □ ひどく痛い
7. この数日間，しゃがみこみや立ち上がりのとき膝が痛みますか。
□ 全く痛くない □ 少し痛い □ 中程度痛い □ かなり痛い □ ひどく痛い
8. この数日間，ずっと立っているとき膝が痛みますか。
□ 全く痛くない □ 少し痛い □ 中程度痛い □ かなり痛い □ ひどく痛い

Ⅲ. 日常生活の状態

この数日間のあなたの日常生活の状態についてお聞きします。あてはまる回答を1つ選び，□に√をつけてください。

9. この数日間，階段の昇りや降りはどの程度困難ですか。
□ 困難はない □ 少し困難 □ 中程度困難 □ かなり困難 □ 非常に困難
10. この数日間，しゃがみこみや立ち上がりはどの程度困難ですか。
□ 困難はない □ 少し困難 □ 中程度困難 □ かなり困難 □ 非常に困難
11. この数日間，洋式トイレからの立ち上がりはどの程度困難ですか。
□ 困難はない □ 少し困難 □ 中程度困難 □ かなり困難 □ 非常に困難
12. この数日間，ズボン，スカート，パンツなどの着替えはどの程度困難ですか。
□ 困難はない □ 少し困難 □ 中程度困難 □ かなり困難 □ 非常に困難
13. この数日間，靴下をはいたり脱いだりすることはどの程度困難ですか。
□ 困難はない □ 少し困難 □ 中程度困難 □ かなり困難 □ 非常に困難
14. この数日間，平らなところを休まずにどれくらい歩けますか。
□ 30分以上歩ける □ 15分ぐらい歩ける □ 家のまわりを歩ける程度 □ 家の中を歩ける程度 □ ほとんど歩けない
15. この数日間，杖を使っていますか。
□ 全く使わない □ たまに使う □ ときどき使う □ しばしば使う □ 必ず使う
16. この数日間，日用品などの買い物はどの程度困難ですか。
□ 困難はない □ 少し困難 □ 中程度困難 □ かなり困難 □ 非常に困難

(次頁につづく)