Clinimetrics

Alvan R. Feinstein

Yale University Press
New Haven and London

ISBN 0-300-03806-2

Clinimetrics

Alvan R. Feinstein, M.D.

Yale University Press
New Haven and London

To students, colleagues, and friends
who stimulate the search for knowledge
and who nurture the searchers

Contents

Chapter 15

Summary and Conclusions

For readers who like to have a synopsis of what they have read, this chapter contains a summary of the preceding fourteen chapters, together with some additional conclusions.

Nomenclature and Functional Classification

Clinimetric indexes are arbitrary ratings for the diverse phenomena of clinical care that are observed subjectively and that cannot be expressed in dimensional numbers. The indexes have diverse roles in clinical medicine. Indexes of status contain diagnostic criteria for existence, or graded ratings for relative magnitude or severity of symptoms, diseases, functional capacity, and other clinical conditions. Indexes of change describe alterations that are determined from successive ratings of a condition's status, or from special transition scales aimed at describing change itself. Indexes of prediction are used for prognostic estimations; and guideline indexes, which are sometimes called "protocols" or "algorithms," offer instructions for decisions that lead to diagnostic and therapeutic actions. The data produced by clinimetric indexes are used for diverse activities in observing, planning, evaluating, and statistically analyzing the results of patient care.

The indexes are important because they describe human sensations, reactions, and judgments that are often regarded as too "soft" to be included in statistical or other "scientific" analyses of patient care. The omission of the soft clinical information is detrimental for both humanism and science in modern medicine. Humanistically, the focus on "hard" data does not include suitable attention to distress, relief, comfort, gratification, and other distinc-

tively human desires and goals. Scientifically, in the absence of suitable soft data, the patients who are the observed "material" are not identified reproducibly or effectively. The identifications are inadequate because some of the most cogent distinctions in diagnosis, prognosis, and therapy depend on patterns of symptoms, severity of illness, effects of co-morbidity, functional abilities, and other clinical phenomena that must be described with soft data, and that will not be accounted for in the hard information.

A common reason why soft clinical information is often scientifically disdained is that the information comes from subjective observations and is expressed without the formal standards applied for measuring hard data in laboratory work. The problem of subjective observation is actually less of a scientific obstacle than the lack of standardized expression because other forms of subjectively observed data—obtained via histopathology, radiology, and electron microscopy—have received widespread scientific acceptance. The standardization problem, which has been a prime topic in this book, can be resolved if a formal process is established for the development of clinimetric indexes. The process will have to differ substantially, however, from what is used in other forms of measurement because the indexes have so many different structures, purposes, and functions.

Principles of Structure

Clinimetric indexes are constructed from component variables that are often expressed in nominal, binary, ordinal, or quasi-dimensional rating scales that differ from dimensional measurements. In some instances, the original scale of a component variable is transformed into a different scale before the component enters the index. A single index often contains many component variables, which are sometimes organized into axes (or "subscales") before being arranged into the output scale of the index.

The output scale formed by a combination of component variables can be cited as a tandem profile or as an aggregation. In a tandem profile, a separate rating appears for each component axis. In an aggregation, the component ratings are joined into a single set of output categories that form a new scale. The aggregations can be arranged as additive scores or Boolean clusters, and the arrangements can be organized as summations or in a ranked hierarchical format.

The diverse scales of expression, multiple component variables, and composite patterns of aggregation create major complexities in construction and evaluation. The strategies used to deal with these complexities will differ substantially from the simpler methods that can be applied when conventional laboratory measurements are cited in standard dimensional scales for single

variables. In a composite clinimetric index, such attributes as the scope, discrimination, coherence, and stipulated aggregation of the output scale are crucial features that seldom need consideration in conventional laboratory measurements.

Choice of Component Variables

Because of the need to consider the specific clinical purpose and setting for which each index is intended, the component variables for clinimetric indexes are chosen in a manner that differs from other forms of measurement. In laboratory work, each measurement is expressed in the same way no matter how the data may be used, but clinimetric indexes may require different components for different usages. In psychosocial measurements, the components that enter each index are often selected after a long list of candidate "items" has been evaluated and tested. In clinimetric indexes, however, the component variables are often chosen directly from a review of past clinical experience, followed by a "dissection" of the "intuition" with which the pertinent phenomena were appraised. The process is eased by the fact that each phenomenon or "construct" being described in clinical work has usually become well recognized and established *before* attempts are made to express it with an index. After being selected, the component variables are further refined to include the appropriate forms of intrinsic or extrinsic evidence and to aim at an appropriate focus for the interpersonal exchange.

When many candidates are considered as possible component variables for a new index, the most important ones are often selected mainly by ordinary clinical judgment, but "importance" can also be determined with various mathematical models or with combinations of clinico-statistical judgment.

Organization of Output Scale

When the component variables are combined to form an output scale, the scale should contain an exhaustive scope of mutually exclusive categories, a suitable location for all the component parts and combinations, and realistic values. The output scale will be defective if it produces excessive, inadequate, or obscure discrimination, or if it does not offer an effective method of discerning changes in state. Some of these problems can be avoided if the index focuses on a smaller number of crucial components or is constructed as a separate transition index, intended specifically to identify changes.

Biometric coherence, another valuable feature of a multi-component index, is achieved if the attributes combined in the composite categories have

a plausible biologic meaning and also have similar statistical properties. Another desirable attribute is "transparency," which allows the distinctive contents of the composite categories to be easily discerned from the output expression. Transparency is impaired if too many variables are combined in a single citation.

The three attributes of biologic plausibility, statistical isometry, and scientific transparency are all desirable, but they often involve conflicts that keep all three from being attainable in a single index. Consequently, the investigator will often have to choose which of the competing goals should be given priority. The problems that arise during these choices will re-appear in the subsequent discussion of conflicts in statistical, psychosocial, and clinical goals for clinimetric indexes.

Role of Patients' Collaboration

Regardless of how the component variables are chosen and combined, the results of many indexes are often unsatisfactory because an important feature has been neglected: the collaboration of the patient. To give accurate or appropriate responses either to an interviewer or to a self-administered questionnaire, the patient must be sufficiently cooperative, alert, and comprehending. In some important instances, however, the responses are accepted without suitable evaluation of the patient's mental condition.

If the index requires performance of a physical task—such as climbing stairs, breathing in a special way, or working at a particular job—the magnitude of the task should be clearly specified and the task should be clearly understood by the patient. If the task is not accomplished, the reasons for non-accomplishment should be differentiated into clinical, phobic, economic, or other reasons; and a distinction should be made between threshold levels or maximum tolerance levels for magnitude of performance. Since the magnitude of a performed task can usually be raised if effort is decreased or if social support is increased, an assessment of magnitude alone will be inadequate if effort and other sources of support are not suitably considered. For example, angina pectoris can easily be "cured" if the patient works much more slowly at the same task that formerly evoked the angina when done rapidly.

Another important aspect of collaboration is the patient's choice of a preferred focus when multiple manifestations or disabilities are the target of therapy. The patient's selection of such a focus can help orient a suitable direction for both the treatment itself and the index with which the treatment is evaluated.

Global Indexes and Scales

A unique feature that differentiates clinimetric from laboratory measurements is the frequent use of global indexes and scales. These indexes are expressed in visual analog, ordinal, binary, or nominal categories, but operational criteria are not provided to demarcate the choice of each category. The indexes are most commonly used for broad "global" ratings of complex phenomena such as severity of illness, post-therapeutic improvement, or health status, but can also be applied for more specific clinical entities such as severity of dyspnea, severity of pain, or degree of cardiac enlargement.

The main advantage of the global indexes is their simplicity and their direct focus on the selected phenomenon; their main disadvantage is the absence of the stipulations needed for scientific reproducibility. The latter disadvantage is often reduced if the same person repeatedly uses the same index, if the global index describes transitions rather than single states, if certain precautions (such as "blinding") are used to enhance objectivity, or if comprehension of categorical ratings is aided by graphic displays.

Because of convenience and ease, global indexes are constantly used in clinical communication, but they are often shunned in scientific analyses because the data are regarded as soft and unreliable. The hard data that are used instead, however, may have high scientific quality but may not offer suitable descriptions of the selected phenomena. The achievement of both sensibility and standardization in the same index is often difficult, and the desire of attaining both goals may lead to major unrecognized conflicts when mathematical, psychometric, and clinical strategies are used in constructing and evaluating clinimetric indexes. The collaborators from the different disciplines may not realize that they have different goals, and may not perceive the need to specify priorities or make suitable compromises when the disparate goals are pursued.

Goals of Statistical Methods

Statistical strategies can be used in at least two different ways in the construction of clinimetric indexes. In one approach, if the index is formulated by judgmental combination, statistical procedures can review a collection of existing data and can identify (or confirm) the variables that seem most important for inclusion as components of the index. In a second approach, the index itself is constructed from the data by a statistical technique that chooses the important variables, assigns weights to their coefficients, and combines them into a score or cluster.

With either approach, the strategies depend on mathematical models that

summarize the data in a selected (usually "linear") pattern and that evaluate the variables according to their apparent effect in reducing statistical variance. Although the results may indicate the "important" variables, the importance depends on basic judgments that are mathematical rather than clinical or scientific. The judgments involve arbitrary decisions in choosing a mathematical model, reacting to violations of basic assumptions that underlie the model, selecting variables to be explored for "interactions," establishing guidelines for the computer operations, identifying elemental variables and "unions" of variables, designating systems to code the data, and interpreting the results. Because the results are usually displayed as coefficients of association and expressions for reduction in variance, the investigator may not see direct tabulations for the actual impact of the individual variables or combinations of variables.

The statistical procedures have been highly successful in circumstances where only a few variables are being analyzed. When many variables are under consideration, the results have the advantage of coming from a standardized mathematical procedure. The disadvantage is that the procedure makes all its decisions with a mathematical orientation. The orientation does not consider biologic plausibility or clinical connotations; it is often strongly affected by the amount (rather than scientific content) of the data; and it is often difficult for clinicians to understand and apply.

Goals of Psychosocial Methods

The psychosocial strategies for constructing indexes are often aimed at specific scientific goals, such as achieving a unidimensional variable, with a consistently monotonic pattern, and equal intervals between categories on the output scale. Although desirable for developing indexes about personal attitudes and beliefs, these goals may differ sharply from what is desired in a clinimetric index. The clinician may deliberately want to combine multiple different variables, knowing that the result will not form a consistently monotonic pattern, and seeking an ordinal set of graded ratings that will not have equal intervals between categories. In addition, when psychosocial methods are used to develop a hierarchical ("Gutmann scale") arrangement, the mathematical formulations may differ substantially from the judgmental approaches with which a clinician might organize the rankings.

In certain psychosocial approaches, personal opinions may be solicited with indirect or general questions such as "Do you like doctors?" rather than with direct, targeted questions such as "Do you like *your* doctor?" The indirect strategy may be desirable for certain types of interrogation, but the answer may not always reflect the particular opinions that were sought.

Perhaps the main advantage of the psychosocial approach is that it is accompanied by an established background of "theory" to justify the methods and goals. Lacking an articulated set of principles and "theory" for their own aims in constructing indexes, clinicians may not recognize the occasional or frequent disparities between psychosocial and clinical goals, and may not provide a suitable orientation for the psychosocial-clinical collaboration.

Theory and Evaluation of Sensibility

A theory can be developed for the goal of sensibility in clinimetric indexes by considering five main features of each index: its purpose and framework, overt format, face validity, content validity, and ease of usage.

The purpose and framework are evaluated to check that the index is suitable for its desired function (in describing status, change, etc.); that its novel features or improvements are justified; and that it is applicable for the clinical setting in which its use is planned.

The overt format of the index should be comprehensible, with a relatively small number of coherent variables; the results should be replicable via operating instructions that are clear and carried out in an unbiased manner; and the output scale should be suitably comprehensive in its scope of categories and suitably discriminating among the categories.

The "face validity" of the index requires an appropriate direct focus for the interpersonal exchange, a suitable emphasis in the choice of basic evidence, components that are biologically coherent, and proper attention to the role of the patient's personal collaboration. The "content validity" of the index is determined by checking that it has not omitted important variables or included unimportant ones; that the component variables have been given suitable weights; that the elemental scales are satisfactory for each variable; and that the basic data have good quality.

The ease of usage depends on the time, effort, potential hazards, and type of personnel needed to employ the index.

The various features just cited can be arranged into a list of 21 principles that can be used, somewhat like a "review of systems" in a patient's clinical examination, as a background strategy or theory for the evaluation of sensibility in clinimetric indexes.

Evaluation of Consistency

An essential feature of quality in scientific data is the achievement of similar results when the same measurement is repeated by the same or another observer. Because the similar results may all be wrong, the term *reliability* is

not a good name for this attribute of the data. The word *consistency* seems preferable to *repeatability* or *reproducibility,* particularly at times when the exact circumstances in which the measurement is made can be neither repeated nor reproduced.

Consistency of a measurement process has paramount scientific importance because inconsistent results will seldom warrant further appraisal. The consistency of the process is usually checked externally for observer variability within or among the users, but composite indexes can also be checked for internal consistency in the inter-relationship among component variables. The different components can be checked for "split-half reliability" or for the generalized inter-correlations calculated with statistical procedures such as Cronbach's alpha.

The most customary and conventional assessments of consistency are devoted to observer variability. Before any formal tests begin, efforts can be made to reduce observer variability by developing clear instructions for use of the index, and by improving the operational instructions after exploration of the disagreements noted in pilot studies. The disagreements can arise from variability in the input, the procedure, or the users. The operating instructions must provide both ingredient criteria for the basic observations and conversion criteria for their transformation into the available ratings.

When tested in formal field trials, observer variability can be noted with several statistical expressions. They can cite the average magnitudes of disagreement, or the percentages of absolute agreement or weighted disagreement. An alternative approach, which contains a correction for agreement that might occur by chance in categorical data, is to cite the concordance associations by calculating kappa or weighted kappa. In all of these statistical expressions, arbitrary numerical boundaries must be established to delineate gradations of good or poor agreement.

Field trials of observer variability have often been omitted when an index seemed to perform well first in pilot studies and later in practical applications. The trials are particularly desirable if substantial acts of subjective judgment are involved in the basic observations or their subsequent conversions.

Evaluation of Validity

Because a "gold standard" measurement seldom exists to offer a definitive result, the accuracy of clinimetric indexes can seldom be evaluated. The appraisal is usually devoted to validity, which is examined in two different ways. In one approach, called "criterion validity," the new index is checked against a gold standard if it is available; otherwise some other indexes, which

are well accepted as measurements of the same or an analogous phenomenon, are used as the substitute criterion.

In a second type of appraisal, called "construct validity," the index is checked for its suitability in describing the concept (or "construct") of the selected phenomenon. These phenomena can range from clinical ideas about congestive heart failure, to general classifications of health status, and to psychometric beliefs about intelligence. Construct validity can be appraised in at least three ways. In the first method, the purpose of the analysis is to help demonstrate the existence of a construct created by mathematical procedures such as factor analysis. In the second method, which is sometimes called "discriminant validity," the goal is to show that the new index does something different from existing indexes, by demonstrating its dissociation (rather than correlation) with those indexes. In the third method, which is particularly common for clinimetric indexes, the index receives a "validation by application." In this procedure, the index is evaluated according to its actual performance in an appropriate research study. For example, an index for rating pain would be validated by application if its results suitably differentiate a known effective analgesic agent from a placebo in a clinical trial of pain relief.

The statistical expressions for validity often employ measurements of association, but additional tactics (such as the calculation of sensitivity, specificity, and likelihood ratios) may be used if the index is being checked for accuracy in diagnostic identifications or prognostic predictions. Two important policy decisions for these evaluations involve the management of uncertain results and the choices of boundaries to demarcate abnormality, success, or other categorical zones.

Design and Evaluation of Field Trials

Although field trials are needed to produce the data that are statistically evaluated for consistency and validity, the trials themselves seldom receive the intensive attention given to the statistical results. Nevertheless, because the trials are designed to test a group of people for a particular purpose, the structure of the trials is crucially important for deciding whether the statistical results are pertinent and generalizable.

In a study of consistency, the choice of the observed group, the participating observers, and the methods of comparison must be unbiased, suitable for the main objective, and applicable to the subsequent groups to which the results would pertain. In a study of validity, the assembled evidence should also be aimed at the kinds of patients and clinical settings in which the index would subsequently be applied.

Field trials are also used for a type of validation that serves to confirm the original studies of "validity." The index is often developed from a previous set of data called the *training, developmental,* or *generating* set. The results are then checked and confirmed in a different group of people, who are called the *challenge* or *test* set. Because this type of confirmation is often essential but often omitted, many indexes have been validated within their generating set, but have not received the additional cogent validation that would come from a challenge set.

Regardless of whether the field trial is concerned with consistency or validity (or both), the input to the trial should contain a suitably broad and representative spectrum of patients, who may need further analysis in appropriate subgroups. The indexes being compared should receive a similar input, avoiding problems that may arise from unstable phenomena, sequential-ordering bias, or changes in clinical setting. The contrasted operating procedures should each be carried out properly and compared in an unbiased manner.

Inadequately structured field trials are accompanied by three main hazards. A false negative conclusion, in which a potentially useful index is dismissed as valueless, can arise if the field trial is done prematurely, before the index has been improved in pilot studies, or if the trial contains unsuitable groups or biased comparisons. A false positive conclusion, which claims value for a relatively useless index, is particularly likely to occur if the comparative observational processes were biased, but can also occur when large sample sizes produce "statistical significance" for relatively unimpressive quantitative distinctions. Perhaps the most common hazard, however, is the over-interpretation of high statistical results for "reliability" and "validity." Investigators seeking an index for a new research project may check the field-trial statistics, be impressed with them, use the index in the new study, and only then discover that the index had not been checked for its sensibility. The index may have been insensitive to change, aimed at the wrong focus, intended for patients in different clinical settings, or otherwise inadequate for its assigned role.

These problems can be avoided if investigators recognize that statistical coefficients of reliability and validity should not be accepted without evaluation of the field trial from which they emanated. In some instances, the investigators may be better off in making the effort to construct a new index, rather than using a "validated" index that works improperly.

Taxonomy of Clinimetric Indexes

For comparative evaluations and classifications, clinimetric indexes can be divided into two main groups: the *ailment-oriented* and the *general.* General indexes refer to general health and functional states that are not distinctive for a particular clinical disease or condition. Ailment-oriented indexes refer to specific diseases, conditions, or other clinical manifestations that are the conventional phenomena observed in patient care. The indexes can be ailment-specific (producing diagnostic criteria for a particular disease or descriptions of symptoms distinctive to that disease), or disorder-specific (describing clinical manifestations that are not unique to a single disease). Ailment-oriented indexes can also describe systemic symptoms (such as anorexia and weight loss), the temporal attributes of clinical manifestations, the co-morbidity of associated diseases, or such ancillary features as additional therapy.

Since all of the ailment-oriented indexes can be used to describe patients in the different parts of the spectrum of an ailment, the term *spectral index* can be applied for combinations that contain several different types of ailment-oriented indexes. The term *hybrid index* can be used when an ailment-oriented index is combined with a general index or with paraclinical data from laboratory tests, radiography, or other technologic procedures.

Additional taxonomic issues involve classification for the format of uncalibrated or global scales that can be applied for many different topics, for the management of demarcated indexes produced as aggregates of global scales, and other variations in conventional structure.

Conclusions

For new technology to be applied with human values and goals, the data under analysis cannot be confined to information obtained exclusively with technologic procedures. The data must include accounts of clinical phenomena that reflect the human values and goals. Among such phenomena are the symptoms and other overt manifestations of disease, the associated disabilities and other functional impairments, and various reactions by patients, families, and physicians to these phenomena. This soft information can be noted and hardened if suitable clinimetric indexes are developed for the activities.

The strategies and structures of clinimetric rating scales can readily be improved, although the procedures will be different from those used to improve scientific quality in conventional laboratory methods of measurement. During the clinimetric activities in construction and evaluation of

indexes, conflicts may sometimes arise between two competing goals. One of these goals is to achieve a sensibility that makes the index easy to use and well suited for its distinctive purpose. The other goal is to have a standardization that makes the index produce consistent, validated results.

Although standardization and sensibility are both desirable goals, they may not always be readily achieved in the same index. The ease of usage desired for clinical sensibility may not be compatible with the extensive details often required for standardization. When the relatively few components of a sensible index are chosen according to clinical judgment, the choice may not have the standardization provided by a mathematical analysis of multiple candidate variables. When the face validity and content validity of a clinimetric index are qualitatively appraised for sensibility, the result will lack the quantitative prestige provided by standard statistical formulas for measuring criterion validity and construct validity. The direct, on-target simplicity of a global scale for describing a complex phenomenon may be sensible but unstandardized, whereas a reliable, validated composite combination of demarcated ratings for multiple component variables may be standardized but not sensible.

The potential conflict between the goals of sensibility and standardization will therefore often require decisions in setting priorities. Certain strategies of statistics and psychosocial science can offer valuable guidance in the quest for standardization, but the clinician's knowledge of purpose, setting, and content is invaluable for achieving sensibility and for giving it priority when crucial decisions are made. The likelihood of achieving both goals will be increased if the clinical investigators and their collaborating consultants become thoughtfully familiar with both sets of goals, with the methods of approaching them, and with the need for both flexibility and ingenuity in the approaches.

Like clinical practice itself, the development of clinimetric indexes requires an intricate combination of artful humanism and rigorous science. The attainment of that combination can help restore, preserve, and augment the role of personal and clinical phenomena as the center of attention when technologic clinical science is used in humanistic patient care.