T. DUCKETT JONES MEMORIAL LECTURE

The Jones Criteria and the Challenges of Clinimetrics

ALVAN R. FEINSTEIN, M.D.

THE MEDICAL IMMORTALITY of being commemorated with an eponym has been achieved by physicians and pathologists who discover or describe some ailment, such as Addison’s disease and von Recklinghausen’s disease, and by surgeons who have invented operative procedures and instruments. For T. Duckett Jones, an eponymic title was attained with a different achievement: He created a set of rules for making an intellectual decision.

It was not the first time that a medical eponym had been given to a decision rule. Robert Koch is well known not only for having found the tubercle bacillus, but also for proposing a strategy of etiologic inferential reasoning, borrowed from Henle, that we now call Koch’s postulates. The work of Duckett Jones was unique in providing a formal rule for diagnostic designation rather than for etiologic inference. Realizing that many clinical ailments could not be identified with a specific histologic pattern or with other forms of pathognomonic evidence, Jones in 1944 published a set of criteria for diagnosing acute rheumatic fever. He is known as the source of the first medical eponym for a set of rules for making a diagnostic decision: the Jones criteria.

Like all scientific innovations, the criteria were not perfect; they have been improved during two revisions. Their basic ideas and contents, however, are intact; and the criteria have had a profound, fundamental impact in clinical science. Many people believe that our major progress of the past few decades in controlling rheumatic heart disease has been primarily a result of the immunologic antibody techniques that allowed the group A streptococcus to be confirmed as the cause and the antibiotics that allowed streptococcal infections to be eradicated or prevented. Nevertheless, just as the control of poliomyelitis required the more basic development of tissue culture techniques, the control of rheumatic fever also required a more basic accomplishment.

That basic accomplishment was the Jones criteria, which allowed clinicians to agree on the process of diagnosing and on how to treat rheumatic fever. This consistency in diagnostic identification of the disease has allowed us to investigate it precisely, to treat it effectively and to prevent it. By demarcating a clearly defined spectrum of disease, the Jones criteria excluded a series of minor ailments that formerly masqueraded under the rheumatic title, and thereby produced a consistent selection of clinical entities that would be diagnosed as acute rheumatic fever. Without this high scientific quality in the basic identification of the disease, clinical investigators would not have had precise correlations for our streptococcal antibody data or clear targets for our antibiotic prescriptions.

Jones’s pioneering idea has been applied in other fields to establish criteria for the diagnostic challenges that occur in clinical circumstances where pathognomonic evidence of disease does not exist or cannot be readily attained. Specific diagnostic criteria have been created for rheumatoid arthritis, systemic lupus erythematosus, gout, scleroderma, alcoholism, tuberculosis, schizophrenia, ischemic heart disease, a variety of cardiographic designations, and many other diseases for which such criteria have been desperately needed in clinical trials, in clinical practice, and in other types of clinical investigation. Despite this progress, however, criteriologic improvements have been long overdue and scant; huge numbers of other clinical phenomena still await the same kind of scientific attention and criteria for designation. I shall discuss the existing problems, the intellectual challenges, and the need for modern investigators to expand the scientific legacy of T. Duckett Jones.

If we contemplate the different kinds of information that must be designated, classified, or measured in medical activities, the variables can be catalogued as demographic, paraclinical, therapeutic and clinical. Demographic data refer to a person’s age, race, gender or occupation. Paraclinical data arise from radiog-
raphy, laboratory tests, biopsies, electrocardiography, and the rest of the phantasmagoria of modern technologic procedures. Therapeutic data describe drug regimens, surgical procedures and other treatments. Clinical data refer to the patient’s physical findings and particularly, to the symptoms and other discomforts that are described when we listen to the patient’s history.

Of these four kinds of variables, the first three can regularly be cited as “hard data” — information regarded as precise, reproducible and trustworthy. The clinical data, however, are usually regarded as soft and unreliable, and so are generally ignored or deliberately omitted when we try to perform scientific studies of patient management. Because of these omissions, the accusation that medical scientists have become dehumanized can often be justified from the absence of human clinical data in the statistics with which we express the results of therapeutic science. If I report that digitalis therapeutically was given demographically to a 14-year-old male with paraclinical evidence of left ventricular enlargement and Q waves, you do not know whether I have treated a dog or a person. If I add the clinical data that chest pain vanished, breathing became comfortable, the recipient of treatment returned to work, and the family was pleased, I have given a humanized report.

To illustrate the amount of human clinical information that is excluded, I shall list some of the many important clinical findings that are regularly noted by good clinicians as part of clinical art and clinical judgment, but that are regularly omitted from the statistics of therapeutic science. We customarily analyze data for patients with a particular disease, but not for the different types of symptom groups that can be encountered in the spectrum of that disease. Even when we distinguish types of symptoms, we do not classify severity of symptoms by separating those that are mild or moderate from those that are severe and incapacitating. We do not indicate the chronometry of symptoms by specifying their duration and sequence. Thus, if one patient has had chest pain and dyspnea for 10 years and another for 10 days, we seldom consider this distinction when we analyze their data. We also do not distinguish a patient who had chest pain for many years and who then developed dyspnea from a patient with a reversal of this sequence.

We regularly omit auxometry — the rate of growth or progression of the disease. Only in recent years, for example, have we begun to differentiate stable angina pectoris from crescendo angina; and in patients with cancer, our anatomic staging system fails to discriminate those with rapid growing tumors from those with slow growth, even though this distinction can be made readily from using the patient’s history.

Many other clinical phenomena could be cited to illustrate the important human events that are neglected in our current statistics, but because of time constraints, the only other one I shall note is performance status or functional capacity. Indexes of performance or function have been used to classify patients with cancer or to note, in cardiac patients, the accomplishments of surgical or medical therapy. Most of these functional indexes, however, are inadequate and clinically naive. For example, when we list a patient as not working, we fail to distinguish whether the occupational incapacitation is caused by the pathophysiologic ravages of the disease, the physician’s advice to stop work prophylactically, the patient’s fear that work may be harmful, or a profound psychic depression induced in the patient by knowledge of the diagnosis and by fears about prognosis. When a patient who is incapacitated only for prophylactic or psychic reasons later encounters a charismatic doctor who gets him to go back to work, we may then attribute the success, erroneously, to whatever therapy the doctor was also using.

The reason for all these errors of omission and commission is that clinical investigators have not given suitable respect and scientific attention to clinical data. We insist on quality control for the dimensional measurements used in paraclinical data and for accounts of therapeutic regimens, but we have not created quality control in expressing the clinical observations that are our own unique clinical responsibility.

The important clinical observations are often omitted from reported analyses because specific rating scales have not been created for expressing the observations as data; or the data cited in a rating scale may be scientifically inadequate because they lack suitable criteria to make them reproducible. Thus, the severity of a particular symptom may not be listed because there is no scale to list it in such terms as mild, moderate, severe or excruciating. Alternatively, such a scale may exist but may lack operational criteria to denote exactly what is meant by mild, moderate, and so on. Consequently, the categories of the scale are applied inconsistently and nonreproducibly.

To improve this situation, we must create new scales or new criteria for old scales that deal with the different roles of clinical data in identifications, classifications, and temporal distinctions. In identification, criteria of existence provide clinical rules for denoting the presence of such entities as arthritis, carditis, or rheumatic fever. In classification, we need criteria of gradation to demarcate such entities as the severity of arthritis, the degree of cardiac enlargement, or the rating of functional impairment. In other types of classification, criteria of quality would specify features of desirability, such as excellent, good, fair, or poor, and features of correctness, such as ratings of right or wrong.

Each of these types of criteria can be applied in different temporal circumstances, referring to a single state, a transition or a prediction. Single-state criteria, based on a single point in time, are used to express phenomena as being present or absent; large, medium, or small; or excellent, good, or poor. Criteria of transition denote a change in two states, expressed with such phrases as appeared or disappeared; larger, unchanged, or smaller; and better, same, or worse. Criteria of prediction are needed when we prognosticate a
future state, using such expressions as good risk or moribund.

If the scales and criteria are suitably constructed, the results can be regarded as measurements regardless of whether they are cited in the kinds of dimensional numbers that so enchanted Lord Kelvin,29 or in categorical expressions. Thus, the kind of scale that is usually regarded as a measurement provides ranked dimensions, such as 0, 1, 2, 3,…, 99, 100,…, with equal distances between any two adjacent categories for such scales as age in years, height and serum cholesterol. Many other useful scales, however, are expressed in categories that can be ranked or unranked.

An ordinal scale, for example, has arbitrarily ranked categories that do not have measurably equal distances between adjacent ranks. This is the kind of scale used for expressing severity of pain as none, mild, moderate, severe, or excruciating; and for citing briskness of reflexes as 0, 1+, 2+ or 3+. Physicians are accustomed to ordinal scales, and we regularly use them for such clinical communications as a statement that “the patient is 4+ sick.”

Another type of categorical scale contains a pair of binary, dichotomous, or existential categories that describe something as being present or absent, or that answer a particular question as yes or no. Such a scale may seem relatively simple because it has only two categories, but the criteria for using the scale may be complicated. For example, to decide whether rheumatic fever is present or absent, we require the elaborate strategic rules provided by the Jones diagnostic criteria. A two-category existential scale can be regarded as semiordinal, because the scale can readily be expanded into a ranked array of several categories if we added such terms as definitely present, probably present, uncertain whether present or absent, probably absent, and definitely absent.

A third type of categorical scale contains unranked or nominal categories such as male or female in the scale for gender; doctor, lawyer, or other in a scale for occupation; and hepatitis, cirrhosis, extrahepatic obstruction, and so on in a nominal scale for diagnoses of hepatobiliary disease.

Our main problem in creating such scales is not that we lack ideas for doing them, but that we lack criteria for denoting the rules that stipulate each category. We have not established a clinical discipline of criteriology, analogous to the operating procedures that lead to quality control in laboratory data. Consequently, we generally lack ideas about how to establish the criteria and to make decisions about what are the criteria for criteria.27

If we construct a scale and set of criteria, we can begin by deciding on their objective. What is the role or function that they are to play? Next, we decide upon a frame of reference. What are the particular kinds of persons or clinical situations in which the criteria should be used? Next, we identify the particular input components that are used to enter data into the criteria. Thus, we look first at the observed evidence, and then convert the observed evidence into the salient elements. In the Jones diagnostic criteria, for example, the observed evidence may consist of such things as pain and swelling or redness of a joint, and the characteristics noted in cardiac murmurs or in an ECG. The salient elements into which we convert this evidence are described in such terms as arthritis, carditis, chorea or erythema marginatum.

Our next step is to decide how these salient elements will be arranged or aggregated into what might be called principal axes. Duckett Jones arranged his salient elements into two principal axes, the major and the minor axes. We then decide how these principal axes are to be converted or arranged into the particular expression that serves as the focal index. In the Jones diagnostic criteria, the focal index is an expression such as yes or no for the diagnosis of rheumatic fever. The instructions for converting the principal axes into that focal index require such things as two elements in a major axis and one element in a minor axis.

We must also engage in the process of justification for the criteria. For the type of activity I am describing, the process of justification has its counterpart in the calibration or standardization that is used for the procedures that provide dimensional data in laboratory work. The justification methods for a scale and its associated criteria can involve three different kinds of activity. One of them is used to demonstrate consistency — the results noted when different observers apply the criteria and the scale in the same situation. The second activity requires external validation, with or without external evidence. For example, a clinical diagnosis of coronary artery disease — using clinical, electrocardiographic and laboratory data — can be externally validated with roentgenographic evidence during coronary arteriography. If no external evidence can be acquired, perhaps the best way of validating a scale is through the consensual agreement of a panel of experts or other authorities. The assembled panel will use Delphi28 or other techniques to arrive at agreement on the expression, qualification and value of the criteria. In many other circumstances, a form of external validation can be obtained to show the “efficacy” of the scale. For example, a set of criteria that performs a diagnostic function can be checked for its sensitivity and specificity. A set of criteria that provides a prognostic prediction can be checked for its predictive accuracy. A set of criteria that provides classification or descriptions of quality can be checked for its impact on the associated clinical conditions in which the criteria are customarily used.

The focal index that expresses the final result of a set of criteria can be created in two main ways: as an additive score or as a Boolean cluster. An excellent example of the additive score technique is the Apgar score. When Virginia Apgar decided that she was tired of using gestalt observations to refer to the condition of a newborn baby as excellent, good, fair, and so on, she created a specific scoring system. She first identified five different components to be observed. They consisted of such things as heart rate, respiration and color. For each of these five entities, she set up a simple
rating scale of 0, 1 or 2. She then added together the ratings for each of the five sets of observations to arrive at an Apgar score, ranging from 0 to 10, that is now universally disseminated and accepted.

Boolean clusters can be created in at least three ways. One of them is a demarcated combination of the type used to arrive at such designations as TNM stage III. In this kind of expression, individual elements such as metastasis to brain, metastasis to liver, or metastasis to the lung are assembled into a specific combination, called distant metastasis, that is marked as stage III. Patients who lack any of the items in that combination are then eligible to be cited as stage II or stage I. Another tactic for a Boolean cluster is what might be called an enumerated union. In this situation, a group of elements are listed as the individual items that can constitute the principal axis of a union. When enough of these items can be counted as present, the answer to the basic question is yes. Thus, the diagnosis of systemic lupus erythematosus can be made if a patient has or has had four of 14 cited manifestations of the disease. Finally, a Boolean cluster can be formed as an enumerated combination. This procedure is illustrated by the Jones diagnostic criteria. To establish the diagnosis of acute rheumatic fever, we look at a combination of two principal axes, and then demand that the patient exhibit two items from a major axis and one item from a minor axis; or one item from a major axis and two or more items from the minor axis.

Several basic intellectual barriers have kept clinicians from achieving more frequent success in our development of clinimetric scales:

1. **The scales are not needed.** Removal of this barrier requires only the recognition that the scales are needed.

2. **The results will be inconsistent.** If suitable criteria are established and if the use of the criteria is standardized, the results will be consistent.

3. **We do not know how to create the scales.** The foregoing illustrations of additive scores and Boolean clusters demonstrate that we do know how to create the scales. The main problem is either that we do not try to do so or that our efforts have been superficial because suitable criteria and standardizations were not developed.

4. **We do not know how to validate the scales.** The validation can be achieved with techniques noted earlier. Certain scales can be checked against external evidence, and others can be validated by the same type of consensual agreement that was used to establish the length of a meter, the weight of a gram, or the volume of a liter. If we establish a clear, unequivocal purpose for which each scale is to be used and if we achieve consistency in usage, the validation process need not be difficult.

With the removal of these barriers, clinicians can proceed with the development of a scientific scholarship of our own — a scholarship that can provide a “hardening” and respectability for the important, crucial, “soft” data of clinical medicine. We must improve our direct methods of clinical observation. We must establish scales and rating systems for clinical phenomena that are currently unidentified; and we must create criteria for the use of those scales. The results will lead to the development of an intellectual domain that might be called clinimetrics. We must make this domain grow and take its place as a worthy, intellectually vigorous member of the rest of the “metric” family.

Biometrics is one of the earliest members of this family. It was created in the late 19th century when Sir Francis Galton fused the methods of biology with the methods of quantification and mensuration. With time, however, the leaders of the world of biometry began to turn away from their primary concern with biology and biologic data. Biometry began to take the more theoretical and analytic issues of mathematical statistics as its prime creative challenge. Consequently, when other people began to look for help in their own problems of measurement, they did not find it in biometry and they eventually established their own “metric” domains. Thus, it was that psychologists, developing rating scales for anxiety and intelligence, began to create a domain called psychometrics, a domain also occupied by people who do educational testing and evaluating of examination procedures. In the world of sociology, the need to establish rating scales for social and familial interactions led to the creation of sociometry. In economics, the need for quantitative analysis and quantitative models of the economic process led to the creation of econometrics. (One of the founders of this domain was recently awarded the Nobel Prize for his work.) Engineers, despite the dimensional data available to them, have created both an intellectual domain and a journal called Technometrics. Even in the field of the humanities, historians who have used the quantitative analysis of data to interpret historical phenomena have taken the name of Clio, the muse of history, to establish cliometrics.

As clinicians enter the “metric” family, we shall find many humanistically oriented friends and relatives awaiting us. They joined the family because they wanted to give intellectual respect and creative scholarship to their important activities. The methods and strategies our “relatives” have already developed can be quite helpful, but we shall have to make careful decisions about which techniques can be usefully transplanted and which ones we need to develop.

The challenges of clinimetrics arise because we have improved our methods of precisely identifying all the data of patient care except the uniquely clinical phenomena that distinguish sick people from animals or molecules. We measure survival but not the quality of life; we report accomplishments in curing disease, but not in helping sick people by relieving symptoms, providing comfort, or enhancing functional capacity. We may then complain that the accomplishments of medical care are evaluated with inappropriate data, such as length of hospitalization, infant mortality rates, and statistics on life expectancy. The blame, however, is really ours. As clinicians and investigators, we have made very few efforts to expand the
scope of the hard and objective but deliberately de-humanized data that are the main information available and encouraged for statistical analysis. If we believe that the human art of clinical medicine is worth preserving and worth analyzing, we must specify the ingredients of that art and express them as documentary evidence. The development of clinimetrics allows us to maintain and improve clinical art while advancing the state of clinical science.

For scholarly clinical investigators and for practicing physicians who want to contribute to clinical scholarship, clinimetrics offers a fundamental creative challenge and an opportunity to achieve eponymic memorials. If we recall the Apgar score for newborn children, the Bayley scale of childhood development, and the Killip classes and Norris-Peel of severity in acute myocardial infarction, we can recognize the eponymic descendants of the Jones diagnostic criteria; we can see the evolutionary intellectual advances that have followed Jones’s original ideas; and we can also be stimulated by how much more work remains to be done.

References

2. Jones TD: The diagnosis of rheumatic fever. JAMA 126: 481, 1944
4. American Heart Association: Ad hoc Committee to Revise the Jones Criteria (Modified) of the Council on Rheumatic Fever and Congenital Heart Disease: Jones criteria (revised) for guidance in the diagnosis of rheumatic fever. Circulation 32: 664, 1965
T. Duckett Jones Memorial Lecture. The Jones criteria and the challenges of clinimetrics.
A R Feinstein

doi: 10.1161/01.CIR.66.1.1

_Circulation_ is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 1982 American Heart Association, Inc. All rights reserved.
Print ISSN: 0009-7322. Online ISSN: 1524-4539

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circ.ahajournals.org/content/66/1/1.citation

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in _Circulation_ can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to _Circulation_ is online at:
http://circ.ahajournals.org//subscriptions/