The hard data creed in current clinical practice: Its spurious validity and the challenge to define meaningful clinical variables

Prof. Alvan R. Feinstein
Yale University School of Medicine, New Haven/USA

Until the technologic advances of the past century, the main information available to physicians was the «soft» clinical data of symptoms and physical signs. During the 20th century, this information was augmented and usually replaced by the scientific and statistical appeal of «hard data» from technologic «para-clinical» tests. The para-clinical focus, however, has led to many adverse scientific, statistical, and humanistic problems. Scientifically, the «clinical material» under study is not precisely or reproducibly identified with para-clinical data alone. Statistically, the results of randomized trials are pertinent for efficacy in average patients, but cannot be applied for decisions about pertinent clinical subgroups and outcomes. Humanistically, patients are distressed that their most urgent concerns (in symptoms, function, and quality of life) do not receive adequate attention. Solutions to these problems will require that «soft» data be identified with new «instruments» and new taxonomic classifications in which patients and clinical care are the main focus of intellectual attention.

The kind of data that were available for medical analysis before the 19th century consisted of patients' symptoms such as *pain*, *dyspnea*, *discomfort*, *distress*, etc. There was also, of course, the demographic information of *age*, *sex*, *occupation* etc.There were also certain physical signs determined by inspection and palpation. The ancient physicians could see things like the *colour* of the skin; they could feel the *size* of a lump; and they could examine the *appearance* of urine and stools as excreta that came out of the body.

During the 19th century additional physical signs became available as data from *percussion*, introduced by *AUENBRUGGER*, and *auscultation*, introduced mainly by *LAËNNEC*, which brought us things such as pulmonary dullness and rales, and cardiac murmurs. Primitive tests began to be done with laboratory examination of excreta and blood. Instead of tasting the urine to make the diagnosis of diabetes mellitus, physicians began to do certain chemical tests and began to boil the urine to determine the existence of protein.

During the 19th and first half of the 20th century, the main scientific challenge in ordinary medicine was *diagnostic*. The challenge was to identify during life what the pathologists would find *post mortem*, after death. The consequence of this scientific challenge was a focus on morphologic diseases that could be seen and identified under the microscope.The primacy of science was given to *pathology* rather than *therapy*. Ordinary conventional physicians began to think about pathology rather than therapy, because therapy was regarded as a useless activity: the science was in pathology. There was also a failure to consider the various functional ailments that were not explained by morphology. My generation of medical students initially grew up with the idea that something like functional bowel distress was not a real disease, because it did not have an abnormal morphology. A headache that was not associated with a brain tumour or other lesions was very often not a real disease. Fatigue could not be a real disease because it was not explained morphologically.With the *morphologic focus*, relatively little attention was given to patients' symptoms and their *ordinary functional capacity*.

As a result of 20th century para-clinical technologic procedures, we now have a *phantasmagoria* of data available for medical analysis of *para-clinical* phenomena. The information comes from: microbiology; chemistry; the gross anatomy that ranges from the early Röntgenogrammes to the modern procedures of ultrasound, radionuclide scans, computerized tomography, and magnetic-resonance-imaging; and microscopic anatomy is now observed during life from biopsies and pap-smears and not just at post mortem examination. We also get information from electro-encephalography, electro-cardiography, and many other electro-physiologic tactics such as electro-myography. There are also data from various intubations, extending from endoscopy above, colonoscopy below, and laparoscopy from the side and front. There are various forms of vascular catheterisation, and the newest information describing immunobiology, DNA, genetics, and chromosomes.

The subsequent 20th century events led to an increased availability and dissemination of the magnificent para-clinical procedures, but also led to the development of a powerful new therapeutic armamentarium in: pharmaceuticals; surgery; various biologic substances such as blood and various prosthetic devices; renal dialysis, improved artifical limbs; and so on.

The opportunities and consequences of these new events were that precise morphologic, microbial and chemical diagnoses of disease could be made during life. Doctors no longer had to wait until the pathologists' post mortem examination to be able to determine an accurate diagnosis. The result is what might be called the «autopsy in vivo». We do all the tests that will let us determine, without the death of the patient, what the exact diagnosis may be. The consequence of these activities was the gradual decline of anatomic post-mortem pathology (because the precise scientific evidence was increasingly acquired during life), an emphasis on testing rather than talking, and, of course, the high costs of the tests that have made modern medicine so very expensive.

In major scientific developments of clinical activities themselves, there have been some early advances during the past thirty or forty years in clinical taxonomy for classifing patterns of symptoms and co-morbidity. There was also the construction of some early methods to analyse what might be called *clinical cohorts* with strategies that can remove or reduce a biased evaluation. In the early 1960's the *thalidomide* disaster occurred, leading to demands for randomized experimental trials to show the efficacy of all new pharmaceutical agents. Like most innovations in medicine, the randomized trials were initially greeted unhappily. There were complaints about the ethics of placebo or inferior treatment. There were complaints about the rigidity of the protocols used for conducting the trials. There were complaints that the doctors' clinical judgment was being removed from the treatment of the patient. These complaints were offered when randomized trials were first introduced about 45 years ago, and the complaints still occur from time to time today, sometimes justifiably. Nevertheless, there was a gradual acceptance of the trials. They began to receive wide-spread usage after the development of an appropriate informed consent arrangement for patients, and after approval by various ethics committees that were established in almost every institution that did randomized trials. The trials, which have now been established as the *gold standard* for evaluating therapy, have brought magnificent scientific progress in our current era. For the first time in medical history, patients can be assured that the pharmaceutical agents tested with randomized trial protocols are indeed efficacious, and that the agents are more than the nostrums of the past.

The challenges and consequences of the need to evaluate and confirm the efficacy of new therapeutic agents has led to many other activities. One of them has been the development and the primacy of the statistical methods often used for the design and analysis of the randomized trial data. In some ways, the statistician has now replaced the pathologist as the doctors' doctor. Randomized trials have also led to a focus on so called reliable or *hard* data. What is reliable? Death, the hardest of all data, and the various para-clinical tests. There has also, of course, been the high costs of doing many of the trials.

I will yield to no one in my admiration for randomized clinical trials. I have done many of them myself; and I have been a director of a co-ordinating center for randomized trials. When they can be done and when it is appropriate to do them, they are a magnificent scientific advance and advantage. It is also quite appropriate, however, to bear in mind that randomized trials are not the only activity in human life or in human medical science.

There are some very definite limitations to randomized trials. The customary goal of a randomized trial is to show that treatment A is, on average, more effective than treatment B or placebo.The randomized trials provide that evidence splendidly, thus demonstrating that, on average, the pharmaceutical agents do indeed have efficacy.The results of the trials, however, may not always be applicable in ordinary clinical practice, because the results for an average patient may not always pertain to the clinical distinctions of individual patients. A second problem is that the admission criteria for the trial are usually set up to give a relatively restricted spectrum of clinical conditions. A further problem is that the end points of the trial may not always represent the totality of clinical phenomena in which the doctor and the patient may be interested. Finally, the results of the trial may not always be specified for pertinent clinical subgroups.

The hard data in randomized trials and in other aspects of current medical science usually refer to data obtained with objective technologic observation, expressed in dimensions such as milligrams, metres, and liters. We want preservable specimens to allow repeated observation for checking reproducibility. All of the para-clinical information is obtained and reported that way. We can save the blood; we can save the urine; we can save the films; we can save the slides; and we can repeat the observations.

We sometimes forget, however, that metres, litres and milligrams became hard data via the consensual agreement of appropriate experts. If you ask: How did a metre get to be this long? the answer is not that someone went up to the mountain and came down with a metre. How did a litre get to be about this much? It was not obtained on a mountain. In both instances, a group of people sat down to agree on just how long a meter is going to be. The group of experts reached an agreement that made the data hard. I make this point, because if we were to decide that soft clinical data warranted intellectual attention, scientific concern, and the development of scientific reproducibility, we could meet and establish standards for that procedure also.

It so happens, however, that the soft phenomena of human events cannot be expressed in technologic measurements. The events require human observers and appropriate citation in rating scales. We could develop those scales, however, and we could arrange for their standardization and usage; and we could harden the soft data if we simply expanded our focus to include soft information about human beings and not just hard data. It's also important to bear in mind that without the soft data, the «clinical material» that we call patients is not reproducibly defined.

Much of our science in contemporary clinical medicine is inadequate, because although we define the disease very well, and although we define and demarcate the para-clinical data very well, we do not demarcate the soft information that distinguishes a patient with one kind of illness from a patient with another kind of illness within the spectrum of the same disease. Our current scientific problem is that satisfactory observations and classifications are missing for soft data entities that determine clinical decisions and outcomes in many diseases. To illustrate the omissions, you can think of whatever disease happens to be your focus of interest, and I shall use a solid cancer.

The *pattern of symptoms* is important. Practicing physicians know that a patient who is asymptomatic is much better off than a patient who has distinct symptoms due to the cancer.

The severity of symptoms and illness is also important. A patient who has lost a great deal of weight is much worse off than a patient who has the same cancer in the same anatomic stage with the same histologic cell type, but who has not lost weight.

The chronometry of events and auxometry, or rate of progress of the illness, are also important. We know very well that a woman who has had the lump in her breast for 20 years, without any change having taken place in that lump, is much better off because she has a slow growing cancer, in contrast to another woman, also with breast cancer, who has the same mammogram, and the same stage I adenocarcinoma, whose breast has shown nothing on repeated examinations and then suddenly in the past month developed a large lump. Nevertheless, that very important distinction in auxometry, or the rate of growth of illness, is almost totally ignored by the oncologists today as they pursue various treatments of cancer.

The severity and other effects of co-morbidity are also important. Co-morbidity refers to the diseases that can be present in addition to the main disease under study. Two patients may have exactly the same cancer, exactly the same age, exactly the same sex, exactly the same morphology and cell type for the cancer, and yet one has a much better prognosis than the other because the second one, in addition to the cancer, has decompensated liver disease, repeated myocardial infarctions, or far advanced chronic pulmonary disease.

A person's *functional capacity* is also important. What people can do, and their limitations in daily life are sometimes a better guide than almost anything else to predicting what's going to happen in certain diseases; and certainly, when we look at the outcome of treatment, functional capacity is an immensely important thing to individual persons. Many aspects of mental and psychosocial status have prognostic implications, and also important effects on observing outcomes. Social and familial support can sometimes make the difference between a patient being able to work or not work. I am thinking, for example, of a paraplegic musician, who in ordinary kinds of functional classifications might be regarded as in dire difficulty. Yet, if he happens to be ITZHAK PERLMAN, he does very well in his concert life because of support that he had from his family when he was growing up. I know a patient who has no arms, who, given help in the morning to get his artifical arms on, can work and function quite effectively. Another patient with no arms and no one to help him put them on would be listed as functionally unable to do anything. The difference of a family member made that distinction.

Finally, there are issues in the *doctor's style and patient's preferences*. We often fail to take into account that some patients like one kind of doctor, and other patients like another. I often divide the medical world into doctors who are *«treaters»* - they like to treat - and those who are *«non-treaters»* - they don't like to give a lot of treatment. You can also divide patients into *«treatees»* - those who love getting medicine - and *«non-treatees»* - those who do not. If you have a *treater doctor* and a *treatee patient*, they often have a wonderful relationship. If you have a *non-treater doctor* and a *non-treatee patient*, they also may do very well. It is the cross-overs where the difficulties often occur. Nevertheless, in looking at the way people feel after they are treated, this is another aspect of style and preferences that is often omitted when we evaluate the results of ordinary clinical therapy.

The problem that we have now distinctly run into is the *deification* of randomized trials. Because some people would argue that randomized trials are the only way to evaluate therapy, a kind of *statistical hegemony* has developed for the design and analysis of the trials. With this hegemony, everything that is critical and important and that requires great intelligence is the statistics. Consequently, a focus is often placed on statistical procedures rather than clinical distinctions. Inadequate attention is given to the need for other scientific methods and to the various non-statistical challenges that are involved in ordinary daily clinical practice. In our current scientific problems, suitable taxonomies exist for classifying and analysing demographic data such as age, race, sex, socio-economic status, and occupation. A splendid taxonomy exists for the para-clinical tests of morphology, electro-physiology, blood, urine, etc. A suitable taxonomy does not exist, however, for the clinical phenomena, expressed in soft data, that differentiate patients and diverse forms of illness within the same disease. One of the great current ironies at the ending of the 20th century as we prepare to enter a new century - a century in which we have molecular biology and majestic technology - is that the oldest form of medical information, symptoms and clinical conditions, still needs major scientific attention for observation and classification. We need suitable *clinimetric* methods to observe and express the phenomena discerned only by patients and clinicians.

What has often happened today is that clinicians have evaded and escaped the challenge of trying to identify their own data, and have often turned the challenge over to psychometric methods. Persons from the world of psychology or sociology, who have developed such tactics as psychometrics and sociometry, have been asked to use their techniques to develop clinical expressions for things like *quality of life*, *health status*, and *satisfaction with care*. Alas, the methods are often unsuitable. One of the difficulties is the development of multi-item instruments that may not capture important personal distinctions. Furthermore, the multi-item instruments are often insensitive to change because they are intended to measure a single state in time. They do not necessarily capture the nuances of changes from one time to another. Sometimes, the multiple items involved in the instrument may obscure the main emphasis on what the patient and doctor are interested in. The multi-item instruments are often aimed at statistical ratings for homogeneity. The goal is to show that all the items are measuring the same thing. The psychometric measurements of reliability and diverse forms of validity are then expressed statistically while ignoring what might be called *face validity* or *common sense*. The idea of important topics in these instruments is often determined by statistical calculations or authoritative proclamations, not by asking patients directly what is important.

Let me tell you one short story. How many of you are familiar with the *Apgar score*? It ranges from 0 to 10, and is used for rating the condition of a newborn baby. How many of you know the first name of the person who created the *Apgar score*? For the women in the audience I would like particularly to point out that the name was VIRGINIA APGAR. I am regularly surprised that women looking for feminine heroes in the medical world are not aware of VIRGINIA APGAR, whom I regard as the founding parent and patron saint of clinimetrics. She was an anesthesiologist working in the birth room helping to deliver babies. One day, she decided that she was tired of using words like «excellent», «good», «fair», and «poor», to rate the conditions of the babies. So she used her own good sense to say: «What do I look at? There are five things: Heart rate, respiratory rate, colour, muscle tone, and various reflex responses.» She took those five things and gave them each a rating of 0, 1 or 2, with 2 for the best. She then added the five ratings together to produce the Apgar score, ranging from 0 to 10, that is now used all over the world. I often wonder what would happen if VIRGINIA APGAR, who did this work fifty years ago in 1943, were trying today! She would have a series of psychometric consultants, and we would not have an Apgar score. Instead, we would have an *Apgar* ***instrument***. It would have a hundred items. The first item might be, «I think newborn babies are cute - strongly agree/strongly disagree». The second item might be, «When a newborn baby turns blue, I get nervous - strongly agree/strongly disagree». The instrument would have all kinds of statistical blessings attached to it for reliability and validity, but it would be utterly useless. Working with her own clinical intelligence, and taking on a challenge that she saw directly in her own clinical work, VIRGINIA APGAR created an index that works very well. That is the kind of activity that we clinicians should be doing for many other kinds of our own observations in ordinary clinical practice.

A prominent problem in the past decade is that patient care is often regarded as dehumanized. The problem arises not because we physicians are brutal beasts, or unethical, or immoral. The problem is created by a scientific decision to deliberately disregard the pertinent human soft data. The difficulty is a scientific principle, not a moral one. Unhappy with orthodox medicine, many patients seek alternative approaches whose efficacy is not documented, but the patients are very happy to find in many alternative approaches a clinician who may actually at times listen. The orthodox medical attempt to improve soft data has been delegated to psychosocial scientists who develop complex instruments for assessing health status and quality of life. Unfortunately, the psychometric methods are usually based on authoritative opinions and mathematical analyses in which individual patients are not asked or allowed to state their own concepts of importance.

Several years ago, working with an orthopedic surgeon who wanted to get a better rating scale for the effects of replacement of the hip, I taught him about sex. How did I teach an orthopedic surgeon about sex? I told him to ask the patients why they wanted their hip replaced, instead of using the ordinary scale, that has things like: «I have pain during the day, I have pain during the night, I can walk up and down steps with ease or with difficulty». Just ask the patients, I said, Why do you want your hip replaced? The surgeon then discovered that a substantial number of patients were having difficulty with certain sexual positions that they might have wanted to be in. They hoped for improvement with the new hip. The simple but highly effective approach is to ask the patient: What do you want and why do you want it?

Our new scientific and humanistic challenges today are to develop suitable clinical measurements and taxonomy for soft data in the spectrum of an index disease. The soft data include things like the patterns and durations of symptoms, the severity of symptoms, the diagnostic and prognostic effects of co-morbidity, functional capacity, and severity of the total illness itself. The scientific and humanistic challenges are to improve the soft data used in general clinical care, in information about health status and the quality of life, in the doctor-patient relationship, and in the way patients are satisfied with care. We also need to recognise that sometimes, simple *visual-analog scales* may be more effective than multi-item statistically validated instruments. We particularly need to restore patients, rather than doctors or para-clinical tests or methodological authorities, to the centre of the clinical universe. I suspect that *PARACELSUS, were he alive today, would also be urging attention to those same challenges.*

Another important set of challenges is to develop suitable observational substitutes for the many circumstances where randomized trials cannot be done.

Finally, I would point out that as we go into the 21st century, we need to give scientific and humanistic analysis to asking patients and getting answers to four old questions that go back, not merely to PARACELSUS, but even to HIPPOCRATES. These questions reflect the oldest and most profound information in our clinical heritage.Those four questions to the patient are: (1) *How are you?* (2) *What would you like done?* Then later, after we have done our act of healing or whatever it might have been, we ask (3) *How are you now?* and (4) *How well was it done?* To the extent that we pay attention to that kind of information, we will be good healers.To the extent that we incorporate the information into the science used for evaluating our healing, we will also be healing scientists.

Further reading

FEINSTEIN AR. Clinical judgment revisited: The distraction of quantitative models. *Annals of Internal Medicine* 1994; 120: 799 - 805

WRIGHT JG, FEINSTEIN AR. A comparative contrast of clinimetric and psychometric methods for constructing indexes and rating scales.*Journal of Clinical Epidemiology* 1992; 45: 1201 - 1218

FEINSTEIN AR. Hard science, soft data, and the challenges of choosing clinical variables in research. *Clinical Pharmacology & Therapeutics* 1977; 22: 485 - 498

FEINSTEIN AR. Clinical Judgment. Williams & Wilkins Co., Baltimore 1967