Disease Classification/Diagnosis Criteria

Hasan Yazici and Yusuf Yazici

The science and practice of rheumatology rely heavy on criteria. This is true for both clinical practice and research. This chapter will focus on disease classification and diagnostic criteria only. Outcome and remission criteria are handled in chapter "Outcome measures in rheumatoid arthritis".

The prevailing view is that we should have separate criteria sets for research and diagnosis, the former requiring *classification* and the latter *diagnostic* criteria. We propose this is not only unnecessary but unfounded. The main aim of this chapter is to discuss why we indeed have to put heavy emphasis on disease criteria in rheumatology and how we should we go about it and why it is wrong to have separate classification and diagnostic criteria for any one disease. In doing this we will resort to specific examples from the recent attempts in criteria making for rheumatoid arthritis (RA) and vasculitides, especially Behçet's syndrome (BS).

H. Yazici, MD (✉)
Division of Rheumatology, Department of Medicine, Cerrahpasa Medical Faculty, Cerrahpasa Hospital, University of Istanbul, Cerrah PaÅŸa Mh No 53, Istanbul 34098, Turkey
e-mail: hasan@yazici.net

Y. Yazici, MD
Division of Rheumatology, Department of Medicine, Seligman Center for Advanced, Therapeutics and Behcet's Syndrome Center, NYU Hospital for Joint Diseases, New York University School of Medicine, New York University, 333 East 38th Street, New York, NY 10016, USA
e-mail: yusuf.yazici@nyumc.org

H. Yazici et al. (eds.), *Understanding Evidence-Based Rheumatology*,
DOI 10.1007/978-3-319-08374-2_3

A Brief History of Disease Criteria in Rheumatology

Generations of physicians around the globe have been and still are taught the Jones criteria to diagnose rheumatic fever [1]. These criteria were developed by Dr. Jones in 1944 by a strictly ad hoc and eminence based approach and we are afraid this legacy continued in several later updates. The last attempt to better these criteria was a consensus conference in 1992 [2]. In this last conference the issue of potential geographical differences in the utility of these criteria were brought up, as if this was something unique to rheumatic fever. As we will bring up once more below, the utility of *any* diagnostic criteria is strictly dependent in the setting in which the criteria are applied. It is no surprise then that more recent formal surveys keep on showing that the sensitivity of the Jones criteria for diagnosing rheumatic fever is only around 30 % in endemic areas like India [3].

In 1974 Dr. Desmond O'Duffy proposed his Behcet's Disease criteria [4]. The more senior author of this chapter (HY) was in the audience as a young fellow when this set of criteria was presented in a rheumatology meeting. At the end of the presentation he got up and had the courage to ask the presenter " How successfully do your criteria tell Behcet's from ingrown toe nails, particularly since I saw no attempts to prospectively test these criteria in a real setting nor a control group in your exercise?" There was little discussion and few heated exchanges, but this outburst was probably the initial stimulus for the latter work related to the formulation of the International Study Group Criteria for Behcet's Disease (ISGC) [5] currently in use.

Perhaps a new era began in criteria making when American College of Rheumatology (ACR) began publishing criteria for many of the vasculitides [6]. These criteria were no longer ad hoc. A survey was conducted seeking formal sensitivity and specificity for many of the common primary vasculitides. Granted they were based on retrospective analyses and lacked prospective testing, they were an important step away from sheer eminence.

Then came the realization that these ACR vasculitis criteria were not useful for diagnostic purposes [7]. In a formal sensitivity and specificity study it was shown that these criteria had limited (17 % to 29 %) positive predictive values when applied to 198 patients with various vasculitides and connective tissue diseases. Two years later, a further study showed that Chapel Hill Consensus Conference (CHC) criteria, another widely recognized vasculitis criteria set mainly based on the size of the vessel involved, correctly identified only 8 of 27 patients with Wegener's granulomatosis and 4 of 12 patients with microscopic polyangiitis [8]. The response to this issue was that these two sets of criteria were not intended for diagnostic use but were strictly classification criteria for research and educational purposes [9]. This contention sounded very reasonable when first heard and over the years it became the standard to call all disease criteria classification criteria. This was followed by a new desire and the promise to prepare diagnostic criteria in addition to classification criteria for our diseases, an exercise, which we are afraid, might be likened to constructing a perpetual motion machine.

Why do we Need Disease Criteria?

As with other major and hotly debated issues the "why" of an exercise is often a neglected caveat. Apart from preparing for boards and other such evils there are some very good reasons for having disease criteria. These include:

1. To diagnose diseases to help our patients. This encompasses both managing and explaining to the patient the nature of his/her illness.
2. To conduct valid clinical or basic research about these conditions.
3. To explain to the public, health authorities, third party payers, research supporters and financial source allocators the nature of our patients' illnesses.

The presence of separate reasons, at first sight, might be taken to indicate that we might actually need separate classification criteria for research and diagnostic criteria for our patients but perhaps other sets of criteria for still other purposes and even perhaps one for Brussels and another for Washington, as well. We, however, propose that one set of criteria should be good enough for diagnosis, research and public awareness as long we explain, first to ourselves, then to our patients and rest of the public what we intend to with these criteria openly, frankly admitting we cannot diagnose every ill we see. This explanation should obviously continue with the explanation that we can manage some of these ills rather effectively even when we do not know what the exact diagnosis is.

Rheumatologic Diseases as Constructs

As we emphasized in the previous chapter, many rheumatologic diseases do not have specific clinical, histologic, laboratory or radiologic features. Hence we have to come up with constructs to specify what we mean by a "disease". For example if we have a shoulder which is swollen in the shape of a shoulder pad and when we biopsy it we find amyloidosis, we do not have to come up with a construct to tell us and the patient that he/she has amyloidosis. The same is true for a painful, hot and swollen knee from which you isolate staphylococci. On the other hand, in a patient with chronic mouth ulcers, attacks of diarrhea and episodes of uveitis you have to build up a construct to identify Behçet's syndrome and another to identify Crohn's. Still yet, you have to build up a construct to tell one from the other. Why do you have to resort to constructs? Simply because neither Behçet's nor Crohn's can be identified by a specific appearance, histology or a laboratory finding. So you need to build up a concept composed of specific features, in other words *a construct*. Surely the need for such constructs in rheumatology is not as extensive as in psychiatry with their voluminous standard reference manual, DSM (Diagnostic and Statistical Manual of Mental Disorders) of the American Psychiatric Association which includes definitions of over 400 different mental disorders [10] but we still need them. A set of criteria in turn is nothing more or less than the declaration of the

components of a construct with some hierarchy (more commonly called weighing) of these components. It is to be underlined that elements which we decide to exclude from this construct make up the exclusions of our criteria. We propose that the first step in understanding disease criteria is to realize that they are constructs put together for specific purposes to explain and convey departures from the normal. In addition such constructs are needed not only for diseases of unknown origin. Sometimes we resort to constructs in handling diseases we know in depth the etiology and/or the pathogenesis of. For example in a tuberculosis endemic area we can justifiably begin treating a patient with a cough for so many weeks and a chest radiograph according to a well built up construct for diagnosis of tuberculosis or admit a patient with a chest pain for a suspected myocardial infarction if he/she fulfills the Cook County criteria for chest pain [11]. The main message then is that in the science and practice of medicine a diagnosis is needed mainly after we consider what we do with it.

The Basic Elements of Criteria Making

We have emphasized that we need disease criteria especially when our disease is a construct. The 3 basic elements of criteria making all have to do with concepts in probability. They are sensitivity, specificity and the pretest probability.

Sensitivity: Sensitivity is an easy concept. It is simply the percentage of true positives. If 95 % of patients with systemic lupus erythematosus (SLE) are positive for antinuclear antibodies (ANA) then the sensitivity of ANA for SLE is 95 %. Alternatively if 85 % percent of patients fulfilling a particular set of disease criteria for SLE then it is said that this set of criteria is 85 % sensitive in detecting SLE.

Specificity: Specificity is a more difficult concept. It is the percentage of true negatives. Following the example we gave in defining sensitivity, if 80 % of all people *not* having SLE are *not* positive for ANA, then the specificity of ANA for SLE is 80 %. Similarly if a particular set of criteria is negative in 85 % among a group of individuals without SLE then we say that this set of criteria is 85 % specific in detecting SLE. Why is specificity more difficult [12]? We propose two reasons. First, before defining either the sensitivity or the specificity of any finding or a set of criteria for any disease, we have to first define what we mean by individuals with and without the disease. This is intuitively easier in sensitivity where our job is to only define what we mean by the disease we are interested in. If we are trying to assess the sensitivity of laboratory finding we are only concerned with one disease, SLE. We can surely also specifically want to assess the sensitivity among a subset of SLE patients like early, mild or severe disease. Whichever is the case, when at the end we say that "The sensitivity of the test A is 75 % in SLE we say practically all that needs to be said. With specificity, however the situation is more involved. When we declare that "The test A is 70 % specific for SLE." the information we convey is incomplete. What we need to define here is not SLE but *what is not SLE*. On the one

hand, we can make our test very specific if we test it among healthy people only or we can make it noticeably less specific for SLE if we test it among patients with a particular disease with a known propensity for having a positive test A. In brief, the definition of specificity of any finding for any disease is incomplete unless we also clearly define the population without having the disease of our concern. What needs to be said is "The test A is 70 % specific for SLE when tested among x number of healthy individuals, y number of patients with disease B and z number of patients with disease C".

The second reason we propose for what makes specificity more difficult to grasp than sensitivity is the way we verbalize either concept. When we say "Among 100 patients with SLE 95 patients were ANA positive. Therefore the sensitivity of having a positive ANA test is 95 % sensitive for SLE.", three positive bits of information follow each other. On the other hand, when we declare "Among 100 patients without SLE, 70 were negative for ANA. Therefore the specificity of having a positive ANA is 70 % specific for SLE." we again verbalize three consecutive bits of fact however, now, the first two of these are negative while the 3rd is a positive bit of information. We propose that this mental incongruity is the second reason why specificity is a relatively more difficult concept to remember.

Confidence Intervals Around Sensitivity and Specificity

As we will repeatedly see in this book some of the evidence behind evidence-based medicine is surprisingly new. Recall that when we defined sensitivity above, we only gave a percentage. It does not require a great insight to realize that the quality of information coming from 700/1000 = 70 % and 7/10 = 70 % differ substantially.

It is also sobering to note that confidence intervals are still not popular with criteria makers of our day. On the other hand this should not be surprising in that it was as late as 1995 that the science of medicine was introduced to confidence intervals around sensitivity and specificity [13].

The Inverse Relation Between the Sensitivity and Specificity – The ROC

A further important point to be discussed about sensitivity and specificity is their inverse relationship. The graphic description of this relationship is the so-called ROC (receiver operating characteristics) curve. The term comes from signal detection used by engineers for military purposes during World War II [14]. A graph is constructed by plotting the sensitivity (the so called true-positives) against 1- specificity (the so – called false negatives) for a series of hypothetical criteria to diagnose a disease. The criteria set A with a 90 % sensitivity and 85 % specificity will correctly pick up 90 % of the patients with the disease while it will also falsely

designate 15 % of the individuals without the disease as having the disease. On the other hand the criteria set B with 95 % sensitivity but this time with 75 % specificity will identify 95 % of all the patients with the disease, however this time a considerably more portion, 25 %, of the individuals without the disease will be incorrectly labeled.

It can be said that a substantial portion of medical decision making is, or more realistically should be, based on constantly working with mostly conceptual ROCs. For example, when confronted with a patient with chest pain you want have criteria as sensitive as possible to put him in a coronary care unit for observation. You can afford to be not very specific for diagnosing him/her as having a myocardial infarction. A short time later when you are debating whether to put a coronary stent in you have to be more specific with a trade off in sensitivity. The decision for a coronary bypass is again another point on the curve, etc. In brief, the relation between what you want to do and where you are on the ROC is all important and without the appreciation of its importance all exercise related to criteria making is in vain.

Importance of Pretest Probabilities and Likelihood Ratios in Making Criteria

It is intuitive that more frequent a disease is, more likely it will be diagnosed and vice versa. Bayes' theorem (BT) expresses this numerically. The importance of disease frequency (pretest probability in Bayesian terms) in making a diagnosis is not well appreciated in that the usefulness of any disease criteria ultimately depends on this theorem. BT states that given a set of disease criteria is positive in an individual, the probability of that individual having the sought disease is the product of the positive likelihood ratio (LR^+) multiplied by the pretest probability (PrP) of that disease in the setting where the patient is seen [15]. Briefly A (the probability of disease being present if the criteria are positive) = B (the PrP) X C (the LR^+ as defined by the disease criteria at hand). The formula is usually given in odds but it works with probabilities as well. Since physicians are more used to probabilities we suggest they use these, remembering that a probability is the likelihood of an event happening against the sum of the probabilities of its happening *and* not happening and thus always expressed as a fraction of unity. The odds, on the other hand, is the ratio of the number of times an event can happen versus the number of times it cannot happen. For example if an event has a 80 % probability of happening then the odds of that event happening versus not happening would be 4:1.

A different type of LR, LR^- also helps us in decision making. While a LR^+ is expressed is sensitivity/1-specificity or more simply the ratio of the %'s of true positives to false positives while a LR^- is expressed as 1-sensitivity/specificity or more simply the ratio of false negatives to the true negatives.

To give an example, we know that the sensitivity of the most popular criteria for Behcet's syndrome (BS), ISGC criteria [5] is 93 % while its specificity is 97 %.

With these specifications it means that the LR^+ of the ICBD criteria is = 31 while its LR^- is = 0.07. So if we apply the ISGC set to 100 consecutive patients in an outpatient clinic and the pretest probability of having BS in this clinic is 1.0 %, then the probability that any one patient fulfilling the ISBD criteria would have BS in this clinic would be 23.7 %. Conversely the probability of any one patient not fulfilling these criteria to have BS in this clinic would be 0.07 %. In this example it is clear that ISGC were considerably more useful in ruling out than ruling in BS in this clinic. One also sees from this example how important the pretest probability is to judge the usefulness of any criteria set.

It should be intuitively apparent to the reader from the above discussion that the diagnostic usefulness of both the LR^+ and the LR^- of criteria set depends very much on the PrP of the presence and the absence of the disease in the setting the disease is being sought. There are two additional arithmetic indices that help us here. The first is the "positive predictive value" which is the ratio of true positives to all (true + false) positives and the second is the "negative predictive value" which is the ratio of true negatives to all (true + false) negatives. The arithmetic formula for the first is:

Positive predictive value = sensitivity X PrPd/sensitivity X PrP + (1-specificity) X (1-PrP)

and for the second is:

Negative predictive value = specificity X PrPnd/specificity X PrPnd + (sensitivity) X (PrPnd)

where PrPd represents the prevalence of the disease in the population we are concerned with and the PrPnd stands for the prevalence of nondiseased (including those individuals with diseases other than the one we are trying to diagnose) in the same population.

Going back to the LRs two more important uses should be underlined. They are used to devise disease criteria themselves and they can also be used to find the inherent prevalence of the disease (PrP) we are seeking to diagnose in the setting we practice or for comparing prevalence between different settings in many situations where we do not know the differing inherent frequencies.

In either instance the basic method is the same. What needs to be done is to collect a large group of patients and suitable controls from diseases which come into the differential diagnosis. We then numerically compare the frequency of the individual clinical and laboratory findings of the diseases that come into the differential diagnosis. This process is commonly known as making "a clinical prediction" rule. The usual arithmetic involved is a step down logistic regression to identify which clinical and/or laboratory findings independently contributed to a diagnosis already established.

For example in order to prepare the ISGC set already alluded to a group of BS patients already diagnosed as such were taken [5, 16]. The frequencies of a group of selected clinical findings (since this syndrome has no specific laboratory or histologic findings) of the BS group were compared to the frequencies of the same clinical

findings among, again, an already diagnosed patients' diseases that usually come into the differential diagnosis of BS. Thus for each clinical feature a sensitivity and a specificity were calculated. These made up the LR^+ and the LR^- for that symptom. Following this a step down logistic regression was made, to see which clinical features weighed the most in the differential diagnosis. In a hierarchical scheme from high to low only those features that added up to an increased ability to differentiate BS from the control group made up the disease criteria. Once made, this set of criteria had its own sensitivity, specificity and LRs to tell BS from other conditions. So, in brief, a diagnostic criteria set is nothing more or less than a LR, which has positive and negative components.

As said the LRs can also be used to estimate the disease prevalence (the PrP) in a practice setting. An excellent example for this is how cardiologists used it in the past to determine the PrP of coronary artery disease (CAD) in 2 cardiology and 2 general medicine settings [17]. They wanted to know this to better judge whether a patient presenting with chest pain had a higher, different chance of having CAD when he/she presented to a cardiology clinic versus a general medicine facility. In this exercise they used how various clinical and laboratory features at the time of presentation, through the LRs similarly calculated told, whether they were eventually diagnosed or not as having CAD. In short the clinical decision rule thus prepared from a list of separate LR's was the LR or C in the Bayes' formula as given above. The A was the probability of CAD as observed and the B was the prevalence eventually estimated in the 4 different settings. It indeed turned out that the two general medicine settings had lower PrPs than the two cardiology settings.

One final word before we leave the discussion about LRs is the rather confusing statements in many expert sources is that LRs do not depend on disease prevalence [18]. The issue is that they do not depend on any frequency once the disease criteria or a clinical prediction rule is formulated but they are very much dependent on disease or a disease feature frequency when they are initially formulated and this directly takes us to the next item to be discussed about disease criteria.

Circularity in Criteria Making

A master of quantitation in rheumatology James Fries had once said [19]: *Presence of disease "criteria" affirms our ignorance of the essence of disease. If we understand a disease, we can ascribe the elements that are necessary and sufficient for its diagnosis. One can so define gouty arthritis, in which joint fluid crystals serve as a "gold standard" against which to measure the usefulness of other observations. No other major rheumatic disease, including SLE, has such a standard. Thus, criteria must be constructed in a circular manner, by testing variables against a diagnosis based on intuition. The 'best' criteria therefore only describe the current conventional wisdom in an efficient manner.*

We believe, especially for the practicing rheumatologist, rather lengthy discussion in the previous section about the Bayes' theorem and the LRs were helpful regarding how circular indeed is all criteria making.

However, here we have to make a distinction between a *circular manner* and a *circular reasoning*. When Humpty Dumpty used a word it meant exactly what he chose it to mean "neither more nor less". This is rather similar to the exercise of making criteria to identify a disease X . Once we make it, when the next patient presents with this construct we say this patient has that disease. This identification has surely been made in a *circular manner* as it was ought to be. Now let us assume we had included in our disease criteria the positivity of the laboratory findings y or z as a prerequisite for a diagnosis and tabulate our experience in time the characteristics of the patients we had seen with this disease. If we then say "We saw 100 patients within the last 3 months with the disease X and very noticeably all of them were y or z positive . This is *circular logic* par excellence. While our readers will find many definitions of circular logic in sources starting from ancient Greece, a quite workable definition of circular logic is "coming to a conclusion *unaware* (italics added) that the conclusion reached was inescapable" [20].

Disease Criteria: Classification Versus Diagnostic Criteria

Many of the current disease criteria we use include a sentence to the effect *this set of criteria are useful after other disease are excluded* [5, 21] which indeed deserves the naughty reply, *If so why do I need these criteria to start with?* Another common statement is "*These are classification criteria and we hope to follow up with diagnostic criteria soon."* As recently admitted in an otherwise excellent review [22] the authors acknowledged there were no diagnostic criteria at hand for vasculitis. We agree, however , the authors continued to give the old promise of diagnostic criteria to come. We are afraid that this urge to prepare universal diagnostic criteria for many of our diseases is rather like the ancient hopes of the alchemists or the zealots of perpetual motion machines.

Why is this so? There are several possible explanations:

First, as brought up in the first chapter and reiterated here many of our diseases are constructs without any specific causes and known pathogenic mechanisms. As such their definition is almost solely dependent on how we define them. Even slight differences in these constructs can make their subsequent identification rather difficult. We will return to this more in the next section on the new ACR/EULAR criteria for rheumatoid arthritis. Second, physicians [23] and their patients are not well trained in probabilities. The all-important Bayesian approach with its pivotal PrP is still not widely appreciated. LR is not a frequently heard term in everyday medical parlance. The Bayesian probabilities dictate that when the PrP of a disease is small even with diagnostic criteria with high $^{+}$LRs the chances of false positive diagnoses too high to be useful in the individual patient. Third, as we have noted in the previous 2 sections the LRs of a diagnostic criteria set are very much dependent on the setting in which the criteria was made. So the true validation of any criteria set should be made in the real clinical setting and this is seldom available.

Next, the well intentioned truism that to classify is somewhat different from to diagnose has compounded the problem. To diagnose is nothing more or less than to classify in the individual patient [12]. The cerebral process between the two is virtually the same. It is also self-deceiving to say, as is frequently done, "We classify when we do research and diagnose when we treat a patient." Of note is the not much appreciated hidden implication of this statement. How can we convince ourselves that we can be more objective and more scientific when we do research and be less scientific and less objective when we try to manage an illness? Such a contention may not even be ethical. Lastly, the patients, the health authorities including the third party payers and most importantly us physicians almost always expect a formal and tangible diagnosis rather than a mere classification from physicians. It is important to note that the historically much older word to "diagnose" goes back to ancient Greece and means to discern including "*knowing the nature of*". The word "classify", however, is 17th century and simply means allocating to different classes. Thus a mere classification implies less precision and even less of an attempt "to better comprehend the nature of" [20]. When confronted with a patient, we prefer to diagnose and when we do research, we like to classify. However, in many instances we do not openly admit we cannot make a specific diagnosis in many of our patients. There are also many instances, while we do not exactly know what a patient has, we do much better in recognizing what he/she does not have. This consideration is particularly relevant in situations where we are confronted with a patient with a hitherto undefined disease. It is indeed puzzling why, as physicians, we do not more frequently admit that most of what we diagnose are based on probability, rather than certainty. Perhaps as the proverbial healer we do want to play down our image.

What was Wrong with the 2010 ACR/EULAR Criteria for Early RA?

We chose to give a special place to the 2010 ACR/EULAR criteria for early RA [24–26] in this chapter for the main reason that some of the shortcomings in the design, execution and interpretation of this major effort were rather representative of how the rheumatology discipline, we suggest, inadequately addressed the whole issue of criteria making

The main drive behind the creation of such a criteria set was that within the preceding two decades rheumatologists understood that earlier we treated RA the better the patient outcomes were. It was clear that methotrexate (MTX) was the anchor drug in managing RA but the new biologic agents were also quite promising. Nevertheless, this needed to be officially announced not only to all the rheumatologists but also to the public where the patients and the third party payers came from. In Fries' words about all criteria quoted above [19] the aim was to *describe the current conventional wisdom in an efficient manner.* Finally, the criteria then at hand to classify RA [27] did this in an inefficient way because it identified the disease late in the game, mainly among patients with already serious morbidities.

However the definition of "early RA" was difficult in that in any setting there were (and are) differing views about what constituted early RA. So a construct was needed to define this. The authors decided to select a group of patients from 9 different early arthritis cohorts from either side of the Atlantic. The main 2 main inclusion criteria were that patients had to have inflammatory arthritis (synovitis/swelling) at least in 1 joint and they had to be prescribed MTX at most within a year of their initial presentation. On the other hand the main exclusion criterium was that patients who had an apparent diagnosis, other than RA.

The authors chose MTX initiation within a year of arthritis as the golden inclusion rule for this exercise. They reasoned that a good rheumatologist invariably prescribed MTX to those patients whom he/she thought was developing or have already developed within a certain time, the erosive, deforming bad disease (construct) which we call classic RA today. They did not want to consider fulfilling the older criteria as the golden rule since, they reasoned, this would cause circularity as this set of criteria particularly identified patients with advanced disease, the outcome they wanted to particularly avoid with the use of the new data set. They emphasized it was important to avoid circularity [24, 26].

The exercise had 3 phases. First, the data driven phase, where chiefly by a factor analysis the main elements that prompted a rheumatologist to start MTX therapy in a patient with early inflammatory arthritis were determined. The second phase was a *consensus-based, decision science–informed approach* with the purpose of deriving a clinician based judgement as to which clinical, laboratory, radiographic clues were determinants of eventually developing bad disease in RA. These clinicians not only used their expertise but were also "informed" of the results of Phase 1 in this undertaking. Finally, the third phase was integration of the information from Phases 1 and 2 with a final validation of the criteria set among 3 of the 9 cohorts not analyzed in Phases 1 and 2.

So what are the outstanding problems with this exercise?

A. There was no intention at formulating specificity of these criteria. It followed that there were no control groups with other diseases that come in the differential diagnosis of RA.
B. The authors reiterated that they wanted to avoid circularity by their design. However as highlighted above, circularity is an essential component of criteria making. We find it hard to understand why the authors did not decide to delineate which factors were more important to recognize in a patient with early arthritis to cause severe disease later on. In such a scheme the control groups for the eventual LRs would naturally be those patients who would not eventually develop the disease construct as defined by the older 1987 criteria. An additional end point would have been to add a construct of a milder RA at, say, one year, i.e. increasing number of joints involved, more seropositivity etc.
C. They said that "One limitation of the new criteria is that they are based on current knowledge." [24] This statement is superflous. *All* criteria are based on current knowledge.

D. After their exercise the authors made the default promise of diagnostic criteria to come [24].
E. The whole exercise was about what prompts a rheumatologist to start MTX in early undifferentiated arthritis. It is unfortunate that the authors did not call the exercise just this.
F. Finally, it was indeed curious why the authors chose to give the gender distribution of the rheumatologists involved. Are there gender differences in decision making in rheumatology, particularly as related to RA?

Since there were no comperator groups with any other diseases at any phase of development of these criteria, many such studies about these criteria followed. These showed in brief that 18 % of patients with early arthritis filfulling these criteria in Leiden had a different diagnosis at the end of one year. Of the 198 patients who were classified differently, 46 developed psoriatic arthritis while 6 turned out to have arthritis associated with cancer [28]. In another cross sectional study among patients receiving routine clinical care in an university outpatient clinic in New York, the sensitivity and specificity of the 2010 criteria were 97 vs 55 % respectively while the corresponding values for the 1987 criteria were 93 vs 73 %. More specifically 67 % of patients with SLE and 50 % of patients with osteoarthritis could be classified as having RA by the 2010 criteria [29]. A recent systemic literature review of publications assessing the preformance of the 2010 criteria came up with similar figures for sensitivity and specificity [30]. It was also interesting to note that the authors of this systemic review concluded that the 2010 criteria was more for classification rather than diagnosis.

What to do?

We must first reconcile ourselves that unless the specific cause/ pathogenesis/ histology of a condition is known a fool proof diagnosis is almost impossible. We always have to deal with probabilities especially when we have to manage diseases we recognize only as constructs. Then we have to teach both our patients and the health authorities what we first have convince ourselves. The patient has to know that the basis of most of our medical interventions are based on probabilities. Similarly, and especially, the health authorities and the third party payers should come to grips with the same.

Finally, after first admitting that diagnostic and classification criteria are one and the same, we must begin formulating how we can make a classification scheme more useful for diagnosis. We propose several approaches:

1. To popularize the understanding that many of our existing disease criteria are much more useful to exclude diseases as we gave the example for BS. Devising criteria specific for excluding diseases can be a novel approach.
2. Tailoring disease criteria to a practice setting is another approach. These, if you will, setting-specific criteria will have the potential to be much more useful in

that the PrP will remarkably increase, and the number of conditions that come into the differential diagnosis will decrease, increasing the specificity of the criteria set. To give an example [12] we recognize that the prevalence of BS is may be 1000 fold greater in Japan [31], as compared to North America [32]. On the other hand, if you go to a dedicated uveitis clinic in either country, you will find that the proportion of BS patients that seen in either setting differs only by several fold, 2.5 % in North America and 6.2 % in Japan. This certainly increases the PrP of BS in North America, while the number of conditions to be differentiated from BS as far as eye involvement is considered will be comparatively few. Similarly, a simple disease criteria set to differentiate BS from inflammatory bowel disease for the gastroenterologist would be most useful.

3. Including family history of the disease being sought, surely very important in making a diagnosis, is for some reason, frequently omitted from disease criteria [33]. This needs to change.
4. The scientific journals might consider always requesting confidence intervals around the LRs whenever we devise new criteria. Similarly a systematic effort can be made to add confidence intervals to criteria commonly used. It is disconcerting to note that confidence intervals are not provided in the final criteria set in any of the classification criteria sets published by the ACR.

References

1. Jones TD. The diagnosis of rheumatic fever. JAMA. 1944;126:481–4.
2. Ferrieri P. Jones Criteria Working Group. Proceedings of the Jones Criteria workshop. Circulation. 2002;106:2521–3.
3. Pereira BA, da Silva NA, Andrade LE, et al.Jones criteria and underdiagnosis of rheumatic fever. Indian J Pediatr. 2007;74(2):117–21
4. O'Duffy JD. Suggested criteria for diagnosis of Behçet's disease. J Rheumatol. 1974;1 suppl 1:18 (abstr).
5. International Study Group for Behçet's Disease. Criteria for diagnosis of Behçet's disease. Lancet. 1990;335:1070–80.
6. Hunder GG, Arend WP, Bloch DA, et al. The American College of Rheumatology 1990 criteria for the classification of vasculitis: introduction. Arthritis Rheum. 1990;33:1065–7.
7. Rao JK, Allen NB, Pincus T. Limitations of the 1990 American College of Cardiology classification criteria in the diagnosis of vasculitis. Ann Intern Med. 1998;129:345–52.
8. Sorensen SF, Slot O, Tvede N, Petersen J. A prospective study of vasculitis patients collected in a five year period: evaluation of the Chapel Hill nomenclature. Ann Rheum Dis. 2000;59:478–82.
9. Hunder GG. The use and misuse of classification and diagnostic criteria for complex diseases. Ann Intern Med. 1998;129:417–8.
10. American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 5th ed. Arlington: American Psychiatric Publishing; 2013.
11. Reilly BM, Evans AT, Schaider JJ, et al. Impact of a clinical decision rule on hospital triage of patients with suspected acute cardiac ischemia in the emergency department. JAMA. 2002;288(3):342–50.
12. Yazici H. Diagnostic versus classification criteria – a continuum. Bull NYU Hosp Jt Dis. 2009;67:206–8.

13. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. JAMA. 1995;274(8):645–51.
14. Fan J, Upadhye S, Worster A. Understanding receiver operating characteristic (ROC) curves. CJEM. 2006;8(1):19–20.
15. Max MB, Lynn J (editors). Symptom research: methods and opportunities. Bethesda: National Institutes of Health, Department of Health and Human Services. Available at: http://symptom-research.nih.gov/tablecontents.htm. Accessed 4 Apr 2009.
16. International Study Group for Behçet's Disease. Evaluation of diagnostic ('Classification') criteria in Behçet's disease. Br J Rheumatol. 1992;31:299–308.
17. Sox HC, Hickam DH, Marton KI, et al. Am J Med . 1990; 89:7-14.
18. Greenberg RS, Flanders WD, W, Eley JW, Boring, III, JR. Diagnostic testing *in* Medical Epidemiology, 4th Edition; Lange Medical Books 2004, p.9.
19. Fries JF. Disease criteria for systemic lupus erythematosus. Arch Intern Med. 1984;144: 252–3.
20. Yazici H. A critical look at diagnostic criteria: time for a change? Bull NYU Hosp Jt Dis. 2011;69:101–3.
21. Dasgupta B, Cimmino MA, Kremers HM, et al.. Provisional classification criteria for polymyalgia rheumatica: a European League Against Rheumatism/American College of Rheumatology collaborative initiative. Arthritis Rheum. 2012;64(4):943–54.
22. Waller R, Ahmed A, Patel I, Luqmani R. Update on the classification of vasculitis. Best Pract Res Clin Rheumatol. 2013;27(1):3–17.
23. Cahan A, Gilon D, Manor O, Paltiel O. Probabilistic reasoning and clinical decision-making: do doctors overestimate diagnostic probabilities? QJM. 2003;96(10):763–9.
24. Aletaha D, Neogi T, Silman AJ, et al. 2010 rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. Ann Rheum Dis. 2010;69(9):1580–8.
25. Funovits J, Aletaha D, Bykerk V, et al. The 2010 American College of Rheumatology/European League Against Rheumatism classification criteria for rheumatoid arthritis: methodological report phase I. Ann Rheum Dis. 2010;69:1589–95.
26. Neogi T, Aletaha D, Silman AJ, et al. The 2010 American College of Rheumatology/European League Against Rheumatism Classification Criteria for Rheumatoid Arthritis: phase 2 methodological report. Arthritis Rheum. 2010;62:2582–91.
27. Arnett FC, Edworthy SM, Bloch DA, et al. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum. 1988;31: 315–24.
28. van der Linden MM, Knevel R, Huizinga TWJ, van der Helm-van Mil AHM. Classification of rheumatoid arthritis. Comparison of the 1987 America College of Rheumatology criteria and the 2010 American College of Rheumatology/American League Against Rheumatism criteria. Arthritis Rheum. 2011;63:37–42.
29. Kennish L, Labitigan M, Budoff S, et al. Utility of the new rheumatoid arthritis 2010 ACR/EULAR classification criteria in routine clinical care. BMJ Open. 2012;2(5):pii: e001117.
30. Sakellariou G, Scirè CA, Zambon A, Roberto Caporali R, Montecucco C. Performance of the 2010 Classification Criteria for Rheumatoid Arthritis: a systematic literature review and a meta-analysis. PLoS One. 2013;8:e56528.
31. Rodriguez A, Calonge M, Pedroza-Seres M, et al. Referral patterns of uveitis in a tertiary eye care center. Arch Ophthalmol. 1996;114:593–9.
32. Goto H, Mochizuki M, Yamaki K, et al. Epidemiological survey of intraocular inflammation in Japan. Jpn J Ophthalmol. 2007;51:41–4.
33. Yazici H, Yazici Y. Criteria for Behçet's disease with reflections on all disease criteria. J Aotoimmun. 20014; 48-40: 104-7.