Definitions and Validation Criteria for Biomarkers and Surrogate Endpoints: Development and Testing of a Quantitative Hierarchical Levels of Evidence Schema

MARISSA N. LASSERE, KENT R. JOHNSON, MAARTEN BOERS, PETER TUGWELL, PETER BROOKS, LEE SIMON, VIBEKE STRAND, PHILIP G. CONAGHAN, MIKKEL ØSTERGAARD, WALTER P. MAKSYMOWYCH, ROBERT LANDEWÉ, BARRY BRESNIHAN, PAUL-PETER TAK, RICHARD WAKEFIELD, PHILIP MEASE, CLIFTON O. BINGHAM III, MICHAEL HUGHES, DOUG ALTMAN, MARC BUYSE, SALLY GALBRAITH, and GEORGE WELLS

ABSTRACT. ***Objective.*** There are clear advantages to using biomarkers and surrogate endpoints, but concerns about clinical and statistical validity and systematic methods to evaluate these aspects hinder their efficient application. Our objective was to review the literature on biomarkers and surrogates to develop a hierarchical schema that systematically evaluates and ranks the surrogacy status of biomarkers and surrogates; and to obtain feedback from stakeholders.

Methods. After a systematic search of Medline and Embase on biomarkers, surrogate (outcomes, endpoints, markers, indicators), intermediate endpoints, and leading indicators, a quantitative surrogate validation schema was developed and subsequently evaluated at a stakeholder workshop.

Results. The search identified several classification schema and definitions. Components of these were incorporated into a new quantitative surrogate validation level of evidence schema that evaluates biomarkers along 4 domains: Target, Study Design, Statistical Strength, and Penalties. Scores derived from 3 domains — the Target that the marker is being substituted for, the Design of the (best) evidence, and the Statistical strength — are additive. Penalties are then applied if there is serious counterevidence. A total score (0 to 15) determines the level of evidence, with Level 1 the strongest and Level 5 the weakest. It was proposed that the term "surrogate" be restricted to markers attaining Levels 1 or 2 only. Most stakeholders agreed that this operationalization of the National Institutes of Health definitions of biomarker, surrogate endpoint, and clinical endpoint was useful.

Conclusion. Further development and application of this schema provides incentives and guidance for effective biomarker and surrogate endpoint research, and more efficient drug discovery, development, and approval. (J Rheumatol 2007;34:607–15)

Key Indexing Terms:
SURROGATE | TRIAL ENDPOINT | BIOMARKER
LEVELS OF EVIDENCE | PREDICTIVE FACTORS

From the Department of Rheumatology, St. George Hospital, University of New South Wales, Sydney, Australia.

M.N. Lassere, MB, BS, Grad Dip Epi, PhD, Associate Professor in Medicine, Department of Rheumatology, St. George Hospital, University of NSW, Sydney, Australia; K.R. Johnson, MD, Senior Lecturer in Medicine, University of Newcastle, Newcastle, University of New South Wales, Sydney, Australia; M. Boers, MD, PhD, Professor, Department of Clinical Epidemiology, VU University Medical Centre, Amsterdam, The Netherlands; P. Tugwell, MD, MSc, Professor, Departments of Medicine, and Epidemiology and Community Medicine, Canada Research Chair and Principal Scientist, Institute of Population Health, University of Ottawa, Ottawa, Canada; P. Brooks, MBBS, MD, Executive Dean of Health Sciences, The University of Queensland, Brisbane, Australia; L. Simon, MD, Associate Clinical Professor of Medicine, Harvard Medical School, Boston, Massachusetts, USA; V. Strand, MD, Clinical Professor of Medicine, Department of Immunology, Stanford University, Palo Alto, California, USA; P.G. Conaghan, MB, BS, PhD, Professor of Musculoskeletal Medicine, Academic Unit of Musculoskeletal Disease, University of Leeds, Leeds, United Kingdom; M. Østergaard, MD, PhD, DMSc, Professor in Rheumatology/Arthritis, Copenhagen University Hospitals at Herlev and Hvidovre, Copenhagen, Denmark; W.P. Maksymowych, FRCPC, Professor, Department of Medicine, University of Alberta, Edmonton, Alberta, Canada; R. Landewé, MD, PhD, Associate Professor, Department of Medicine, University of Maastricht, Maastricht, The Netherlands; B. Bresnihan, PhD, Professor, University College Dublin, Dublin, Ireland; P-P. Tak, MD, PhD, Professor, Division of Clinical Immunology and Rheumatology, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands; R. Wakefield, BM, Senior Lecturer in Rheumatology, Academic Department of Musculoskeletal Medicine, Faculty of Medicine and Health, University of Leeds, Leeds, United Kingdom; P. Mease, Clinical Professor, Swedish Medical Center and University of Washington School of Medicine, Seattle, Washington, USA; C.O. Bingham III, MD, Assistant Professor of Medicine, Johns Hopkins University, Baltimore, Maryland, USA; M. Hughes, PhD, Professor of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA; D. Altman, BSc, DSc, Professor of Statistics in Medicine, Centre for Statistics in Medicine, Wolfson College, Oxford, United Kingdom; M. Buyse, ScD, Executive Director, IDDI, Louvain-la-Neuve, Belgium; S. Galbraith, BMaths, DSc, Lecturer, School of Mathematics and Statistics, University of New South Wales, Sydney, Australia; G. Wells, PhD, MSc, Professor, Department of Epidemiology and Community Medicine, University of Ottawa, Ottawa, Ontario, Canada.

Address reprint requests to Prof. M.N. Lassere, Department of Rheumatology, St. George Hospital, Gray Street, Kogarah 2217, Australia. E-mail: marissa.lassere@sesiahs.health.nsw.gov.au

Biomarkers and surrogate outcomes are used throughout the development, testing, and ongoing positioning of medical therapeutics[1-3]. Temple defined a surrogate outcome as "a laboratory measurement or a physical sign used as a substitute for a clinically meaningful end point that measures directly how a patient feels, functions or survives"[4]. The term "surrogate" literally means "a substitute for, replacement, proxy," so some additional outcome is expressed or implied[5]. For example, serum cholesterol and blood pressure are frequently accepted surrogate endpoints for cardiovascular morbidity and mortality. However, often the nature or even the identity of the substituted outcome may go unmentioned, making the relevance of surrogate data to the overall clinical condition difficult to determine. Additionally, some surrogates are surrogates for other surrogates. Many different terms such as biomarkers[6], intermediate outcomes or endpoints[7-9], leading indicators[10], and even surrogate endpoint biomarkers[11], surrogate intermediate endpoint[12], and early markers of response, have arisen and are sometimes used to imply a benefit in some other outcome without being explicit as to what that outcome is or providing the evidence between surrogate and patient-relevant outcome. The relationship between prognostic factors and surrogate markers may also be a source of confusion. This heterogeneity of nomenclature, content, and inference in surrogate therapeutics obscures the meaning of what is being measured and its genuine influence on patients.

In 1963, Donald Mainland, a statistician, asked, "Will the variables that we observe be the variables that we really wish to know about?" If "we are substituting something that is easy to observe for something that is difficult to observe, we have no right to do so unless we know the connection between the two things"[13]. Substitution is frequently employed in medical practice. Disease and pathologic processes manifest as patient symptoms, abnormal findings on physical examination, or abnormal laboratory investigations. However, physical findings and laboratory investigations are often asymptomatic and their importance and meaning are determined by their connections to how a patient feels, functions, or survives, usually at some later point in time. Diagnostic and prognostic test evaluation examines the scientific basis of this connection, and criteria have been developed[14,15] and continue to be developed. Clinicians apply diagnostic and prognostic test evaluation of symptoms, physical signs, and investigations daily in clinical practice, although the application generally is not explicit (few formally calculate pretest and posttest odds) and is often imperfect.

Similarly in clinical practice, we use a combination of symptoms, signs, and test results to evaluate whether treatment is working. Often we initially rely on a physical sign or test result alone to ascertain whether our treatment is effective, particularly if we believe that changes in these indicate that the treatment is working, but before there is a change in how a patient feels, functions, or survives. While substituted outcomes are often easier, cheaper, or quicker to observe, the change in the substituted outcome may not capture the combined benefit-harm of a treatment. How do we establish whether a substituted outcome, i.e., surrogate, effectively captures the combined benefit-harm?

In a users' guide to surrogates, Bucher, *et al*[16] provide examples where the use of surrogate endpoints in clinical trials would have caused more harm than good. An early example in rheumatology was the randomized controlled trial (RCT) of fluoride, which showed more rather than less fractures over placebo despite improved bone density[17]. Therefore, they propose "before offering an intervention on the basis of effects on a surrogate outcome, the clinician should note a consistent relation between surrogate and target in randomized trials; the effect of the intervention on the surrogate must be large, precise and lasting; and the benefit-risk tradeoff must be clear." Their proposal is an important advance because it identifies the central issues in surrogate therapeutics. Building on this, we now offer a formal development of a hierarchical system to provide "levels of evidence" across the spectrum of surrogates. Although levels of evidence have been established to quantify the strength of evidence for treatment itself[18], there are no criteria for quantifying the strength of evidence for using substituted outcomes. A system is needed that explicitly grades the evidence-base of surrogate outcomes, and that, in turn, gives recommendations regarding the application of surrogates. The drug development industry, patients, and consumers also have a strong interest in surrogate outcomes[19-21]. The industry's interest in surrogates is reflected in their language, with terms such as "proof of concept," "leading indicators," "early markers of response," and "intermediate endpoints." Early drug development needs laboratory markers as "proof of concept" and in drug-dose exploration; however, these terms may be unfamiliar to clinicians.

All stakeholders, industry, the medical profession, and patients, clearly benefit from the use of surrogates. Conducting a large longterm trial of a new drug therapy is difficult for both investigators and for enrolled subjects. Using surrogate endpoints rather than how a patient feels, functions, or survives may be easier, more economical, and produce earlier results. There are substantial cost-savings to sponsors and the drug may come to market earlier, thereby benefiting patients. Therefore, the products of surrogate therapeutics have the potential to yield considerable clinical benefits to patients and financial benefits to shareholders. But clinical and financial risks have also been clearly shown by certain high-profile setbacks[22,23]. The schema that follows attempts to address the conceptual, statistical, and pragmatic issues regarding surrogate therapeutics by developing criteria that define and operationalize levels of evidence for biomarkers and surrogate endpoints.

Nomenclature

Begin with the term "variable," a term that has no underlying association regarding its origin or application. This variable

sits on a "variables in medicine" continuum (Figure 1). At one end of this continuum are variables that are disease-centered, reflecting the biology of the disease process and the mechanism of disease. These variables are called biomarkers. The markers may be a biochemical marker, a cellular marker, a cytokine marker, a genetic marker, an imaging marker, a physiological marker[5] (Figure 1). At the other end are patient-centered variables. These are endpoints[5] that reflect how a patient "feels, functions, and survives."

A patient-centered variable has obvious patient relevance, is an end in itself, and requires no further explanation as to its immediate consequence. Although a patient may assign different values on how she or he feels, functions, and survives (that is, there may be a hierarchy of outcomes), a patient-centered variable has intrinsic face validity and needs no further validation. Patient-centered variables are the standard to guide individual clinical decisions and as primary endpoints in clinical trials of therapeutic intervention, but they have limitations — hence the need for biomarkers.

A disease-centered variable (such as blood pressure, LDL-cholesterol, prostatic-specific antigen, rheumatoid erosions on a radiograph) has no immediate or obvious meaning to patients or clinicians. Meaning is determined over time after data are collected from laboratory, epidemiological, and clinical settings, as mechanisms of biology and pathology are understood. Therefore disease-centered variables are not intrinsically valid, and they must undergo a process of validation before they are used in the same clinical contexts as patient-centered variables.

Validation is not an all or nothing event, it is an incremental process. Before a disease-centered variable can be used in clinical contexts (to guide clinical decision-making or be used as a primary endpoint in a therapeutic trial) the variable must meet minimal validation criteria, as proposed in our hierarchical schema (Table 1). We propose reserving the term "surrogates"[5] to disease-centered variables that have been validated in this manner. In Figure 1 the biomarker has moved from its disease-centered position closer towards a patient-centered position. We also propose that the terms intermediate outcomes, leading indicators, early markers of response, and markers are avoided. In this classification system disease-centered variables are called biomarkers, patient-centered variables are called patient outcomes, and biomarkers that have scored above a certain threshold are called surrogates. Being disease-centered in this context does not preclude a biomarker being an essential tool in proof-of-concept studies.

Risk factors, prognostic factors, and surrogate outcomes

Risk factors, prognostic factors, and surrogate outcomes share many properties. Risk factors and prognostic factors are predictive over time, are comparative, require evidence for validity, and are logically equivalent. By convention, individuals without disease have risk factors and individuals with disease have prognostic factors, and both are more likely to develop a certain outcome depending on their risk or prognostic factor status. Risk or prognostic factors may or may not be biomarkers. Just like risk factors or prognostic factors, surrogate outcomes are also predictive over time, comparative, and require

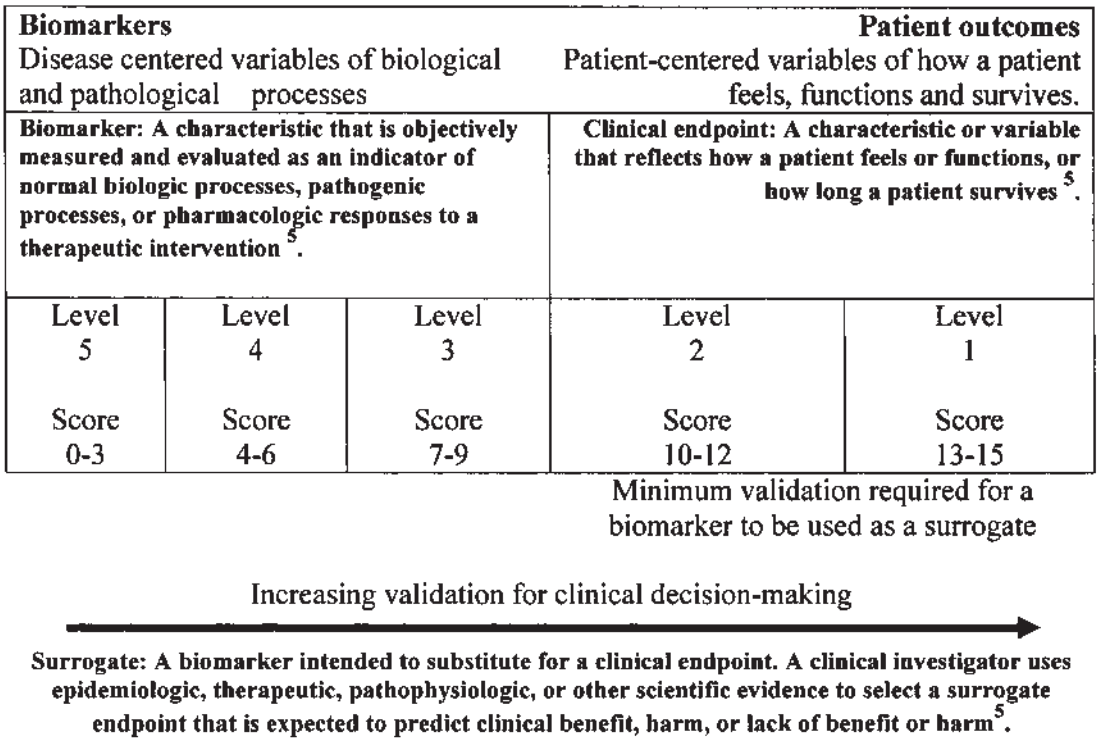

Figure 1. Levels of evidence for biomarkers and surrogates.

Table 1. Ranking surrogate validity: domains, criteria, and ranks.

Domains	Ranks	Criteria
A. Target (for all studies ranked in Domain B)	0	All targets studied are **disease-centred** and **reversible**.
	1	At least one target studied that is **disease-centred** is **irreversible**
	2	At least one **patient-centered** target that is **reversible**
	3	At least one **patient-centered** target of **irreversible minor organ morbidity** or **minor** irreversible **clinical burden of disease**
	4	At least one **patient-centered** target of **irreversible major organ morbidity** or **major** irreversible **clinical burden of disease**
	5	**Death**
B. Study design (Requires as baseline appropriate study quality , study power and study duration)	0	Evidence from ***in vitro*** *OR* **animal studies** *OR* **Case reports** *OR* **Cross-sectional observational** *OR* **Retrospective observational cohorts** studies evaluating the relationship between marker and target.
	1	**At least one prespecified non-population based prospective observational study** with collection of all covariates needed to adjust for known confounding and effect modification evaluating the relationship between marker and target.
	2	**At least one prespecified population-based prospective observational study** with collection of all covariates needed to adjust for known confounding and effect modification evaluating the relationship between marker and target *OR* **One randomized controlled trial** of the **same drug class** of an intervention evaluating the relationship between marker and target.
	3	**At least two randomized controlled trials** of the **same drug class** of an intervention evaluating the relationship between marker and target.
	4	**At least two randomized controlled trials** in each of **two drug classes** of an intervention evaluating the relationship between marker and target.
	5	**At least three randomized controlled trials in each of three known drug classes** of an intervention evaluating the relationship between marker and target *OR* at least three **randomized surrogate objective** trials
C. Statistical Strength	0	**No** association / prediction **OR** no relevant data
	1	**At least fair** association or better between marker change and target change in most single study analyses
	2	**At least fair** association or better between marker change and target change in all single study analyses **OR** fair prediction in an across-study analysis evaluating the effect of treatment on marker change and target change.
	3	**At least good** association or better between marker change and target change in all single study analyses **OR** good prediction in an across-study analysis evaluating the effect of treatment on marker change and target change.
	4	**At least very good** association or better between marker change and target change in all single study analyses **AND** very good prediction in an across-study analysis evaluating the effect of treatment on marker change and target change.
	5	**Excellent** association between marker change and target change in all single study analyses **AND** excellent prediction in an across-study analysis evaluating the effect of treatment on marker change and target change.
D. Penalties due to lack of evidence or evidence to the contrary	-1	**No *in vitro* or animal study** evidence to support surrogacy validity OR no **epidemiological** evidence to support surrogacy validity
	-1	At least one **RCT** that does not demonstrate statistically significant surrogate validity (i.e. **evidence of no effect in at least one adequately powered RCT**)
	-1	At least one **epidemiological** study that supports **opposite assertion**.
	-1	At least one **epidemiological** study that does not demonstrate surrogacy validity (i.e. **evidence of no effect in at least one adequately powered epidemiological study**)
	-1	At least one RCT that demonstrated evidence of significant **clinical heterogeneity**
	-2	At least **one RCT** that supports **opposite assertion**
	-3	At least **one RCT** that demonstrates use of marker **confers patient harm**
	-3	**Does not meet the threshold criterion** of a rank of 3 in at least one domain if score is 7 or more
NB. Marker must meet minimum technical performance criteria as per OMERACT Filter		

evidence for validity, but in a therapeutic context a surrogate is required to "substitute for" the target outcome. This implies that a treatment-associated change in the surrogate for the better will result in a change for the better in the target outcome; therefore, surrogate outcomes must be modifiable. Ideally the change in the target outcome should be of equal magnitude to the change in the surrogate. Prognostic factors also have therapeutic utility, because they can be used to identify subgroups that are more likely to respond to treatment. Finally, because all surrogate endpoints are predictive over time, they are all prognostic factors.

Surrogate outcome validity — sources

The evidence for surrogate outcome validity includes plausible biology regarding disease mechanisms and data supporting a strong association in observational studies between the surrogate and the target outcome. However, one high-profile example of surrogate failure using these sources alone was the Cardiac Arrhythmia Suppression Trial[23]. Prior to 1985, cohort studies showed that certain cardiac arrhythmias were a major risk factor for subsequent mortality. Drugs were developed, approved, marketed, and used based on the assumption that suppressing arrhythmias (surrogate marker) did more good than harm. This hypothesis was tested in the randomized controlled Cardiac Arrhythmia Suppression Trial: 2 active drugs were tested against placebo in patients with post-acute myocardial infarction arrhythmias. The drugs effectively suppressed arrhythmias, but mortality was greater on drugs than on placebo. Therefore, one negative trial relating surrogate to outcome invalidated a decade of cardiac therapeutics based on the surrogate marker results.

This single example highlights why data from biology and well designed prospective observational studies are not sufficient sources for surrogate validity. Biology and observational studies miss the unknown effect of treatment on the surrogate-target outcome relationship. The treatment that modifies the surrogate may have unsuspected patient-relevant toxicity that reduces or even nullifies the surrogate-target outcome link. The surrogate-target outcome relationship is always potentially influenced by the treatment used to induce change in the surrogate. Therefore, evidence of surrogate validity will always require some evaluation of the surrogate-treatment interaction from randomized prospective studies, that is, prospective RCT. Randomization is necessary because it is the only known way to balance, on average, unknown confounders.

Surrogate levels of evidence scheme: Domains and ranks

What follows is a system that assigns a "surrogacy" rank to disease-centered biomarkers. This system essentially operationalizes the definitions developed by a National Institutes of Health (NIH) working group[5]. This system has 4 domains. Within each domain, criteria are ranked: the higher the rank, the better the evidence that domain brings to surrogate validation. The first domain is Target. It is ranked from zero to 5, where zero is "All target outcomes studied are disease-centered and reversible," and 5 is "Death in all studies." Change in a biomarker must agree with change in patient outcome for the biomarker to be a valid replacement for patient outcome in clinical situations. Further, biomarkers that are closer in proximity to patient outcome than others should be captured by the ranking system. Using the example of blood pressure again, in order to establish whether blood pressure can be used as a surrogate, one must show that change in blood pressure is tightly linked to a change in how a patient feels, functions, or survives. Studies that show that a reduction in blood pressure translates into fewer symptoms, improved function, less burden of disease, and reduced mortality clearly provide better evidence of surrogacy status of a biomarker than studies that compare reductions in blood pressure to a change in another biomarker (such as LV hypertrophy on ECG). Successful studies of biomarker validity must choose a target outcome that is patient-centered and irreversible without treatment.

The second domain is Study Design. The study designs that have been used to validate the surrogacy status of a biomarker are ranked from zero, "evidence from *in vitro* or animal studies," to 5, "at least 3 RCT in each of 3 known drug classes of an intervention evaluating the relationship between measure and target OR at least 3 randomized surrogate objective trials." The other scoring categories are listed in Table 1. An example of a study that achieves a rank of 5 is a hypothetical randomized controlled targeted blood pressure trial, in which patients randomized to achieve a target systolic and diastolic level, after stratification by known antihypertensive drug treatments, had better patient outcomes. The study design domain assumes appropriate study quality, study power, and study duration, although these are not explicitly ranked in this proposed schema at this stage. Finally, all biomarkers should be evaluated for technical criteria of reliability and sources of variability, capacity, sensitivity for change, and preliminary appraisal of validity and feasibility before testing in any large and possibly costly RCT.

The third domain is Statistical Strength. This ranks the strength of the association and its statistical significance between change in a surrogate and change in target. It is also ranked from zero to 5, with zero as "No association/prediction OR no relevant data" to 5 as "Excellent association between marker and target in all single-study analyses AND excellent prediction in an across-study analysis evaluating the effect of treatment on marker and target." Clearly, if there is 100% prediction between change in a biomarker and change in its target, this provides statistical evidence of the validity of substituting the biomarker for a target outcome. However, there is currently little consensus on what statistic, or combinations of statistics, is most appropriate here, nor on the precise cutoffs for the rank descriptors (poor, fair, good, very good, and excellent). This domain requires further consideration,

although some of these issues are discussed in greater detail in a companion article[24].

In the fourth domain the biomarker incurs penalties either because of lack of, opposing, or inconsistent supporting evidence from biology, clinical epidemiology, or therapeutic trials, or because there is evidence that use of a biomarker has caused patient harm. Consistency is an important component of surrogate validity. The schema's scoring system incorporates all these areas of concern. Penalties are provided in Table 1.

Surrogate levels of evidence scheme: Overall score

The final step is to determine the overall level of evidence for surrogacy status. An additive scoring system gives a score from 0 to 15 (sum highest score achieved in each domain, then subtract penalties). The levels are ranked to reflect increasing strength of the evidence, with the highest being level 1 (score 13–15) and the lowest level 5 (score 0–3) (Figure 1). To be called a Level 1 or 2 surrogate, a biomarker must meet the rank of at least 3 within the Target Outcome, Study Design, and Statistical Strength domains, and there must not be evidence from a RCT that use of the biomarker caused patient harm. The term "surrogate" should be reserved for levels 1 and 2 only. Otherwise the term "biomarker" is used.

Testing the surrogate levels of evidence scheme: Stakeholder meeting

In rheumatology, as in many areas of medicine, there are incentives to use biomarkers as endpoints in therapeutic trials. At the OMERACT 8 (Outcome Measures in Rheumatology Clinical Trials) workshop meeting in May 2006, stakeholders representing ultrasound (US), magnetic resonance imaging (MRI), synovial tissue assays, soluble biomarkers, and single-joint response were asked to "road test" the nomenclature and surrogate validation schema. Each biomarker stakeholder group was asked to apply their "best" biomarker to the schema to evaluate how well their biomarker performed within the schema and how well the schema performed given their biomarker. The 4 biomarker groups were also asked to apply their chosen biomarker, if possible, in the context of single-joint response assessment, for example, when a single joint is treated with intraarticular biologic or gene therapy. A final group, a "statistical stakeholder group," specifically considered statistical issues.

The workshop began with a plenary session, where the surrogate levels of evidence schema, its rationale, and the supporting literature were described by one author (ML). The basic concept of a 4-domain level of evidence instrument and its scoring system was presented. The plenary was followed by 6 smaller stakeholder workshops. Each breakout group was asked to report: (1) whether the group in principle were comfortable with the schema and/or each of its domains; (2) the application of their "best" biomarker to the schema; the biomarker's rankings for Target, Design, Statistical Results, and Penalties of the selected biomarker; and the biomarker's combined Level of Evidence score; (3) the degree of discussion and degree of agreement/disagreement generated by point (2) above; (4) how well the schema performed given the biomarker and the discussion and agreement/disagreement generated; and (5) general comments and suggestions regarding the schema's validity and feasibility.

Regarding schema construction, most of the breakout groups were comfortable with the schema in principle, and about half were comfortable in principle with combining domains. There was unanimity of comfort with the 4 areas designated as domains and in the use of hierarchy. Regarding the content of the domains themselves, all groups felt "Target Outcome" needed some modification, at least by adding the phrase "clinical burden of disease" to complement "organ morbidity." The soluble biomarker group believed the content of Target Outcome was too strict with its focus on patient-centered outcomes. Three of 6 groups felt the Design domain was satisfactory, 2 wanted it modified to make RCT less dominant compared to longitudinal observational studies, and one wanted more information on power and duration of the RCT. The Statistical Strength domain was not formally discussed by all groups, as many believed they had insufficient expertise within the group. Four of 6 groups felt the Penalty domain was satisfactory, with one asking to add "credit points" as an inverse to penalty points.

The ultrasound group applied the schema to ultrasound "synovitis" and determined this marker to be Level 4. The MRI group applied rheumatoid erosions, finding it to be Level 3. The synovial tissue group scored CD68 macrophage numbers in rheumatoid arthritis, finding it to be Level 3. The soluble biomarker groups attempted to score CRP, but found it difficult to identify a patient-centered endpoint relevant to soluble biomarkers in rheumatoid arthritis and spondyloarthropathy clinical trials. This group used the framework of the OMERACT filter (truth, discrimination, feasibility) to develop validation criteria for soluble biomarkers that could be used to substitute for primary radiological endpoints (imaging biomarkers) in rheumatoid arthritis and spondyloarthropathy clinical trials[24]. All groups agreed that evaluating their biomarker within the context of single-joint response assessment using the schema needed further examination. Results from the breakout workshops are summarized in Table 2.

Following the rapporteurs' feedback from breakouts, there was a general plenary discussion. Three topics were prominent. The first was whether, in the area of rheumatology, Target Domain should have greater discrimination in the lower rankings and whether the highest rank should be downweighted because it is not considered attainable with most clinical trials. The second was the relative overweighting of RCT versus observational data in the Study Design domain. The third topic was what role the Penalty domain should play. The Statistical Results domain, by comparison, generated less discussion, in part because it had been the topic of the previous day's deliberations by a group of statisticians, and the

Table 2. Summary of results from breakout groups on the proposed surrogate validity levels of evidence schema.

	In principle, group comfortable with:			
Group	**Schema**	**Domains**	**Use of Hierarchy**	**Combining Domains**
Generic 1	Yes	Yes	Yes	Yes
Generic 2	Yes	Yes	Yes	No
US	Gp1 Yes GP2. Not sure	Yes	Yes	Yes
MRI	Yes	Yes	Yes	No
Soluble biomarkers	Probably Yes	Yes	Yes	Difficult to say
Synovial tissue	Yes	Yes	Yes	Yes

Group	A. Target Outcome	B. Study Design	C. Statistical Strength	D. Penalties	Example Biomarker
Generic 1	Modify to reflect 'clinical burden of disease'	Good	Good	Good, perhaps include credit points as well	Not applicable
Generic 2	Modify to reflect 'clinical burden of disease'	Good	Good	Rework	Not applicable
US	Modify	Ranking of large observational studies versus small RCTs needs more consideration	No Comments	Good	Rheumatoid Synovitis by ultrasound Level 4
MRI	Modify	Good	No Comments	Good	Rheumatoid Erosions by MRI Level 3
Soluble Biomarkers	Possibly modify	Overweighting of RCTs	Not enough information included	Modify	Serum CRP in RA Difficult to rank
Synovial Tissue	Modify	Need more information on study power & duration	No Comments	Good	Synovial CD68 macrophage numbers (RA) Level 3

content was not finalized at that point. That work involved comparisons of various proposed methods of statistical analysis with simulated and real datasets and is currently under further development[23].

On the last day of the meeting, about 120 participants (clinicians, clinician researchers, researchers, industry researchers, statisticians, and patient participants) voted on the proposed levels of evidence nomenclature and schema. The voting questions and results are reported in Table 3. Consensus was reached on the majority of voting questions. Nearly 90% of participants agreed on the NIH definitions of biomarker, surrogate endpoint, and clinical endpoint, and nearly 80% agreed that other nomenclature and definitions should be avoided where possible. This may be the first time that these concepts had been ratified by a sizable group. The notion of operationalizing the NIH definitions through defin-

Table 3. Results from plenary voting on elements of the proposed surrogate validation levels of evidence schema.

Plenary Voting Questions (N ~ 120 participants)	Results
1. Do you agree we use the NIH definitions of biomarker, clinical endpoint, and surrogate endpoint?	Yes 88% No 5% Don't know 7%
2. Do you agree we should avoid other definitions / nomenclature where possible?	Yes 77% No 13% Don't know 10%
4. In the superworkshop we have tried to operationalize the conceptual NIH definitions by defining criteria that can be used to validate a biomarker as a surrogate. Do you think this is useful?	Extremely useful 25% Probably useful 47% Neutral 18% Probably not useful 8% Not useful at all 2%
5. We have split surrogate validation into several domains, e.g., Target Outcome, Study Design. Do you think this is useful?	Extremely useful 20% Probably useful 45% Neutral 25% Probably not useful 7.5% Not useful at all 2.5%
6. Do you think that there should be a domain that reflects Target Outcome (burden of disease, death)?	Yes 74% No 12% Don't know 14%
7. Do you think that there should be a domain that reflects study design, e.g., animal studies, observational studies, RCT?	Yes 76% No 7% Don't know 16%
8. Do you think that there should be a domain that reflects statistical considerations?	Yes 78% No 6% Don't know 16%
9. Do you think that there is a natural hierarchy of evidence in one or more of these domains?	Yes 79% No 8% Don't know 13%
10. Do you think that sometimes there is evidence that can negatively affect surrogate validity?	Yes 78% No 4% Don't know 18%
11. Do you want to aggregate (lumpers) the domains into a single level of evidence or do you want to report the domains separately (splitters), or both (report separately, then aggregate: lump and split)?	Lump only 5% Split only 20% Both lump and split 71.5% Don't know 3.5%

ing criteria that could be used to validate a biomarker as a surrogate was considered useful by just over 70% of participants. About 75% agreed that domains reflecting Target Outcome, Study Design, and Statistical Results were needed. The majority agreed that Target Outcome should be modified to incorporate "clinical burden of disease" to explicitly reflect functional impairment. Four issues that featured prominently in the plenary are discussed next. These were: whether the schema could be used to address the needs of "early markers of response"; the relative weight of RCT versus longitudinal observational studies in the study design domain; how to integrate risk/benefit and other issues that may bear negatively (or positively) on surrogate validity [as incorporated in the Penalty (+ credit) domain]; and whether the domains should be reported separately only, or whether there was an advantage also to combine the domains in some manner and to report a single level of evidence measure.

Early markers of response and Target Domain. The surrogate validation schema was developed to operationalize the NIH definitions and not as a schema to be highly discriminatory regarding early biomarkers for drug development. However, the schema could be applied to early markers of response if the lower rankings of the Target Domain are expanded. This would allow discrimination and ranking of biomarkers substituting for other biomarkers. But these early biomarkers will not attain a higher level of evidence rank for surrogacy status unless they are subsequently tested against higher ranking patient-centered target outcomes.

The relative weight of RCT versus longitudinal observational studies. The fundamental concept here was that regardless of the strength of prognostic data from epidemiologic settings, application of the surrogate concept to therapeutics cannot, without evidence, ensure that there will be no unexpected toxicity that may undermine what otherwise should be a smooth translation of a surrogate change to an outcome benefit. Given that the treatment must be demonstrated to be effective (on both the surrogate and the final outcome), there is almost always a need for an unbiased control, because effects are usu-

ally at best moderate in size, variability is substantial, and confounders (including unknown confounders) need balancing, all of which necessitates randomized controls. However, well designed observational studies may suffice for validating a prognostic factor or predictive factors, or for validating biomarkers designed to substitute for other biomarkers.

Surrogate validity: penalties and credits. There are issues that bear on surrogate validity that are not captured by the Target Outcome, Study Design, and Statistical Results domains. To capture all these "additional" issues, a Penalties domain was proposed because these issues had a negative influence on surrogate validity, were conceptually diverse, and sometimes had no natural hierarchy. Although the "Penalties" solution was simple and worked, it appeared discretionary to some. Additional confusion may have been due to the –3 penalty if a biomarker did not attain a rank of at least 3 in each of the Target, Design, and Statistical domains. This criterion, one suspects, was also seen as arbitrary. However, this was intentional, because it was based on the conviction that each of these 3 domains is independently important and that a weakness in one cannot be corrected by strength in the other(s). However, there is no one right answer on how these issues can be resolved and other models should be explored.

Reporting a single level of evidence ("lumping"), reporting Domains separately ("splitting"), or reporting both. This is a recurrent issue, with an often almost philosophical tone. The urge to lump is probably driven by the desire to enable simple, cross-venue comparisons, in which a hesitancy to lump is driven by the conviction that the domains are fundamentally different constructs and therefore should defy addition, subtraction, or formal combination in any way. If the goal is to establish levels of evidence for biomarkers and surrogates, then lumping may be needed, and this was agreed upon by 76% of participants.

OMERACT 8 generally, and the workshop specifically, was an extremely valuable venue to explore the generic issues surrounding surrogate validation, to present the surrogate validation levels of evidence schema, to receive critical feedback regarding its rationale and construction, and to test it preliminarily in rheumatology biomarker research settings. These data and feedback will be taken forward along with the results of the statistical methodology workshop[23] to further testing and refinement.

REFERENCES

1. Ellenberg S, Hamilton JM. Surrogate endpoints in clinical trials: cancer. Stat Med 1989;8:405-13.
2. Wittes J, Lakatos E, Probstfield J. Surrogate endpoints in clinical trials: cardiovascular diseases. Stat Med 1989;8:415-25.
3. De Gruttola V, Fleming T, Lin DY, Coombs R. Perspective: validating surrogate markers — are we being naive? J Infect Dis 1997;175:237-46.
4. Temple RJ. A regulatory authority's opinion about surrogate endpoints. In: Nimmo WS, Tucker GT, editors. Clinical measurement in drug evaluation. New York: John Wiley and Sons Inc.; 1995.
5. De Gruttola VG, Clax P, DeMets DL, et al. Considerations in the evaluation of surrogate endpoints in clinical trials. Summary of a National Institutes of Health workshop. Control Clin Trials 2001;22:485-502.
6. Karpetsky TP, Humphrey RL, Levy CC. Influence of renal insufficiency on levels of serum ribonuclease in patients with multiple myeloma. J Natl Cancer Inst 1977;58:875-80.
7. Keys PW, South JC, Duffy MG. Quality of care evaluation applied to assessment of clinical pharmacy services. Am J Hosp Pharm 1975;32:897-902.
8. Freedman LS, Schatzkin A. Sample size for studying intermediate endpoints within intervention trials or observational studies. Am J Epidemiol 1992;136:1148-59.
9. Ylinen J, Takala EP, Nykanen M, et al. Active neck muscle training in the treatment of chronic neck pain in women: a randomized controlled trial. JAMA 2003;289:2509-16.
10. Gion M, Rampazzo A, Mione R, Bruscagnin G. Cost/effectiveness ratio of carcinoembryonic antigen — importance of adequacy of routine requests of tumor markers. Int J Biol Markers 1992;7:179-82.
11. Boone CW, Kelloff GJ. Intraepithelial neoplasia, surrogate endpoint biomarkers, and cancer chemoprevention. J Cell Biochem 1993;Suppl 17F:37-48.
12. Coppin C, Porzsolt F, Kumpf J, Coldman A, Wilt T. Immunotherapy for advanced renal cell cancer. Cochrane Database Syst Rev 2000;CD001425.
13. Mainland D. Elementary medical statistics. 2nd ed. Philadelphia, London: W.B. Saunders Company; 1963.
14. Jaeschke R, Guyatt GH, Sackett DL, et al. Users' guides to the medical literature. 3. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients. JAMA 1994;271:703-7.
15. Laupacis A, Wells G, Richardson WS, et al. Users' guides to the medical literature. 5. How to use an article about prognosis. JAMA 1994;272:234-7.
16. Bucher HC, Guyatt GH, Cook DJ, Holbrook A, McAlister FA. Users' guides to the medical literature: applying clinical trials results. A. How to use an article measuring the effect of an intervention on surrogate endpoints. JAMA 1999;282:771-8.
17. Riggs BL, Hodgson SF, O'Fallon WM, Chao EY, Wahner HW. Effect of fluoride treatment on the fracture rate in postmenopausal women with osteoporosis. N Engl J Med 1990;322:802-9.
18. Sackett D, Haynes R, Guyatt G, Tugwell P. Clinical epidemiology: a basic science for clinical medicine. 2nd ed. Boston: Little, Brown and Company; 1991.
19. Colburn WA. Selecting and validating biologic markers for drug development. J Clin Pharmacol 1997;37:355-62.
20. Lathia CD. Biomarkers and surrogate endpoints: how and when might they impact drug development? Dis Markers 2002;18:83-90.
21. Wagner JA. Overview of biomarkers and surrogate endpoints in drug development. Dis Markers 2002;18:41-6.
22. CAST investigators. Preliminary report: effect of encainide and flecainide on mortality in a randomized controlled trial of arrhythmia suppression after myocardial infarction. The Cardiac Arrhythmia Suppression Trial (CAST) Investigators. N Engl J Med 1989;321:406-12.
23. Lassere MN, Johnson KR, Hughes M, et al. Simulation studies of surrogate endpoint validation using single trial and multi-trial statistical approaches. J Rheumatol 2007;34: (in press).
24. Maksymowych WP, Landewé R, Poole R, et al. Development of draft validation criteria for soluble biomarkers before they can be considered surrogates for structural damage in rheumatoid arthritis and spondyloarthritis. J Rheumatol 2007;34: (in press).