Published in final edited form as:

Endocr Metab Immune Disord Drug Targets. 2009 March ; 9(1): 108–112.

METHODOLOGIC ISSUES IN THE VALIDATION OF PUTATIVE BIOMARKERS AND SURROGATE ENDPOINTS IN TREATMENT EVALUATION FOR SYSTEMIC LUPUS ERYTHEMATOSUS

Matthew H. Liang, MD, MPH^{1,2,3,4,6}, Julia F. Simard⁴, Karen Costenbader, MD, MPH^{1,2}, Benjamin T. Dore¹, Michael Ward, MD⁵, Paul R. Fortin, MD, MPH⁷, Gabor G. Illei, MD, PhD⁵, Susan Manzi, MD, MPH⁸, Barbara Mittleman, MD¹⁰, Jill Buyon, MD⁹, Samardeep Gupta, MD^{1,6}, and Michal Abrahamowicz, PhD¹¹

¹Department of Medicine, Division of Rheumatology, Immunology, and Allergy, Brigham and Women's Hospital

²Robert B. Brigham Arthritis and Musculoskeletal Diseases Clinical Research Center

³Center for Advanced Methodological Support for SLE Clinical Trials (ASSIST)

⁴Harvard School of Public Health

⁵Intramural Program, National Institute of Arthritis and Musculoskeletal and Skin Disease

⁶Section of Rheumatology, VA Boston HealthCare System

⁷Division of Rheumatology, University of Toronto

⁸Division of Rheumatology, University of Pittsburgh

⁹Department of Rheumatology; New York University

¹⁰Office of the Director, National Institutes of Health

¹¹Department of Clinical Epidemiology, McGill University

Abstract

No new drugs have been approved for the treatment of systemic lupus erythematosus (SLE) by the FDA for the last 30 years and one barrier has been the lack of validated of biomarkers and surrogate endpoints. Validation of SLE biomarkers in the past have been methodologically flawed. We put forth a conceptual framework and the five critical criterion for validating putative biomarkers and bio-surrogates in this heterogeneous multi-system disease with protean manifestations. Using the example of a putative biomarker for end-stage renal disease from lupus nephritis, we also performed computer simulations for planning a biomarker bio-repository to support the validation process. "Random time window" sampling where a biomarker is obtained in an interval randomly selected from the total follow-up time for that subject yields serious 'survival bias'. This can be avoided by the "fixed calendar window" design, in which biomarkers are measured within the same, pre-specified period for all cohort members who remain at risk during that period. In lupus nephritis where the incidence rate of end-stage renal disease is relatively low, to accumulate 300 instances of end-stage renal disease, at risk patients would have to be followed for about 5,000 person-years, implying 500 subjects followed, on average, for about 10 years. Increasing the number of biomarker determinations per subject from one to five reduces the required number of subjects by 10-15%, while further increases of the number of observations per subject yielded much smaller gains. The large numbers of subjects required for a bio-repository, makes it essential to maximize the efficiency of study designs and analyses and provides the strongest rationale for collaboration and the use of standardized measures to ensure comparability.

Introduction

Systemic lupus erythematosus (SLE) is a major multi-system rheumatic disorder whose treatment could be considerably improved and yet no new drugs have been approved by the FDA for the last 30 years. Meetings since 2002 sponsored by the major charities for support of research in SLE, the NIH, and the FDA and attended by patients, persons from academia, and representatives from the pharmaceutical industry have highlighted the barriers to drug discovery, development, and evaluation in SLE. The lack of well-validated biomarkers and surrogate endpoints is seen as a major challenge and impediment to the conduct of feasible and meaningful evaluation of new agents. The availability of validated and credible biomarkers and surrogate endpoints would have a major impact in shortening the time and expense of evaluating new agents in SLE in the US since a new drug can be approved on the basis of a less than fully established surrogate that is “reasonably likely, based on epidemiologic, therapeutic, pathophysiologic or other evidence to predict clinical benefit” (21 CFR 314.5000 subpart H) especially in serious or life threatening disease. Notable examples of validated biomarkers in other diseases include CD4 cell counts and viral load in HIV drug testing and quantification of plaques in the central nervous system by MRI in the testing of drugs for multiple sclerosis. The use of 1/creatinine for end-stage renal disease and cholesterol for atherosclerotic vascular disease are examples of putative surrogate endpoints, which have not been validated for drug approval.

A number of potential candidates for biomarkers/bio-surrogates in SLE have been described, but their validation has been methodologically flawed or inadequate (1,2). Our ad hoc group made up of (clinical epidemiologists and trialists, SLE clinical investigators, laboratory investigators, a statistician) was organized to put forth a framework and recommendations for conceptual experiments and studies that should be mounted in the validation of any putative biomarker(s) for SLE disease activity and the validation of putative surrogate endpoints, and the specifications of a biomaterials repository that would be needed to support a rigorous evaluation efficiently .

Definitions

In this discussion we use the following definitions (3-9):

A biomarker is any measurable or observable physiologic, or pathologic structural property or activity that is an indicator of normal biologic or pathogenic processes or response to a therapeutic intervention. Biomarkers can be used for diagnosis, assessment of disease activity, prognosis and/or as a therapeutic target. Biomarkers might be measured in tissue, cells, or fluids.

A surrogate endpoint (we prefer this term to “biosurrogate” the other commonly used term) is a measure or defined state that substitutes for a clinically relevant or meaningful endpoint such as morbidity, death, or end-stage organ damage. A surrogate endpoint could be measured in cells, serum, in a body fluid or by an imaging or functional imaging study. Surrogate endpoints are of particular interest when they are highly correlated with a specific, infrequent clinically meaningful endpoint, and when that clinical endpoint takes years to become manifest. Surrogate endpoints, therefore, are frequently applied to manifestations of end-organ damage. Examples in SLE would be a validated surrogate endpoint which could substitute for end-stage renal disease or one which could predict significant atherosclerotic vascular disease such as a myocardial infarction or stroke. In end-stage lupus nephritis doubling of the serum creatinine or double stranded DNA have been proposed but not definitively or formally validated. In atherosclerotic heart disease, carotid ultrasound showing extensive plaque and intimal media thickening, electron beam CT scans showing coronary calcification, and flow-mediated dilatation of the brachial artery assessed by

ultrasound are being used to show the benefits of prevention strategies but have not been directly validated nor are they accepted for drug approval by the FDA.

Although biomarker and surrogate endpoint are used in their singular form throughout, a composite measure is not precluded. The criteria we propose apply as well to the situation where a “surrogate endpoint” is defined as an aggregate of more than one variable.

The validation of a putative biomarker involves the comparison of the biomarker against independent criteria. Given the absence of a gold standard or “hard” measure of overall SLE disease activity, in practice this means either a physician global evaluation, a patient global evaluation, or a validated physician rating scale such as the BILAG, SLAM-R, SLEDAI, SELENA-SLEDAI, ECLAM (10).

We put forth criteria for validating putative biomarkers and surrogate endpoints recognizing that validation is an iterative process proceeding from screening potential candidates and preliminary validation on available clinical material to increasingly definitive studies.

In these criteria, we focus on the validation of the performance of a putative marker, not the characterization of the marker’s properties such as accuracy, precision, assay range, stability, biological rationale, and ease of determination, which are also important (11-13). These parameters are critical and must be established in advance of validating the biomarker.

A valid biomarker of SLE disease activity **must be** (14, 15):

Criterion 1: Reproducible within subjects An analyte that is measured in routine clinical laboratory or a novel one measured by specialized analytical technology must meet good laboratory practices (The National Committee for Clinical Lab Standards, is a framework for this) and ongoing quality control procedures during a trial. This is crucial particularly in multi-site studies.

Criterion 2: Responsive to clinically meaningful changes in disease activity. This criterion is usually pursued by testing candidate biomarkers in available patient biomaterial collected for clinical purposes or by protocol from prospectively followed SLE patients and other populations. In validating a biomarker for assessing disease activity, the first external criterion to be correlated with the biomarker may be dichotomous (i.e., active vs. inactive SLE). After having demonstrated usefulness in separating such subjects, quantitating a continuous relationship between a scale of disease activity and the biomarker might then be carried out.

Criterion 3: Defined with respect to its temporal relationship with disease activity. The biomarker appears or increases or decreases in amount before (or, less usefully, coincides) with changes in disease activity, but cannot follow it. The majority of the biomarkers in SLE have been studied cross-sectionally which can lead to misleading or overoptimistic associations between a clinical state and the biomarker. Studied prospectively, the predictive accuracy may be non-existent or weak. (16,17)

Criterion 4: Change in expected direction with known effective treatments. Biological plausibility is demonstrated when the biomarker changes on the basis of known pathogenetic pathways.

Criterion 5: Defined for a specific SLE phenotype(s) or (subset) and preferably by severity or stage of disease. A potential biomarker for flare of lupus nephritis such as double

stranded DNA and/or complement split products may not be useful in non-renal lupus patients and not in all subjects with lupus nephritis.

Recognizing that validation of biomarkers is an iterative process evolving from laboratory observations to translational work linking these to clinical states or phenomena, the group made recommendations for validation studies appropriate for various stages of discovery or the state of the knowledge at the time. For the purposes of explication these are termed screening, formative validation and definitive (summative) validation. We limit our discussion to biomarkers of endpoints or of disease activity.

SCREENING FOR PUTATIVE BIOMARKERS

In screening a likely candidate might be identified in the course of mechanistic studies of a specific end organ pathogenetic mechanisms and the investigator would wish to confirm the findings. The most incisive test would be to identify existing informative clinical material. Convenience archived specimens or material from consecutive subjects could be used. Convenience is often sacrificed for informative studies. Unless the available validation material is thoroughly characterized, collected longitudinally in contrast to the majority of existent studies done on cross-sectional data - especially with respect to when the material was obtained relative to the onset of a clinical manifestation, spurious findings may result. The most common scenario, unfortunately, is the observation of a strong correlation where none exists because the test material was collected when the SLE is flaring. When patient material is collected routinely independent of visits for symptoms or problems, immunochemical tests appear less useful for prediction (16)

Once a screened marker shows promise in signaling specific organ involvement or a generalized flare, or in predicting a clinically important endpoint, further tests of its specificity are warranted. Again, well-characterized patient material would accelerate this process but access to diverse heterogeneous patients would also be useful.

DEFINITIVE VALIDATION OF BIOMARKERS—Once a biomarker has been screened for potential interest, definitive validation in other patient populations with particular phenotypes is possible. The ACR directed the Ad Hoc Committee on Response Criteria to develop a priori response criteria for selected target organs (the Renal Criteria are the first in a series (18)), and to make specific recommendations for the minimal endpoints and their preferred manner of measurement. The results of these efforts enable one to provide statistical guidance on the size of studies that will be required.

STATISTICAL CONSIDERATIONS FOR SCREENING BIOMARKERS

A putative biomarker should be able to detect the minimal clinically important difference in SLE overall disease activity or in specific target organ involvement. The minimal clinically important difference in SLE overall disease activity has been defined empirically (10) and makes it possible to calculate the number of subjects required for such work based on standard statistical sample size calculations using any one of the commonly used measures of disease activity (SLE-DAI, BILAG, SLAM, etc). In lupus renal disease, consensus criteria for the minimal renal endpoints and criteria for improvement, worsening and no change have also been approved by the ACR (18) permitting estimates of sample size requirements for studies using these endpoints. Sample size calculations for studies that aim at validation of the biomarkers or surrogate outcomes for such clinical endpoints present additional challenges. Whereas considerable research has addressed the methodological issues in the evaluation of biomarkers of cancers, assessment of the predictive ability of biomarkers in SLE raises special analytical challenges. The analytical methods to validate a biomarker as a useful surrogate outcome and the sample size required for such studies

depend critically on the biological relationships between the biomarker and the observed clinical outcome against which the biomarker will be validated. In the SLE, the most promising biomarkers are ones of disease activity in that they reflect current SLE disease activity (1,2). However ‘true’ SLE disease activity is more a concept than something which can be directly observed. Some manifestations such as cutaneous lupus can be observed at the bed side but many manifestations such as cerebritis, serositis, glomerular disease cannot be directly observed and therefore cannot serve as a criterion for validation. On the other hand, the cumulative result of active disease leads to damage. While cumulative damage can be subclinical or latent, it increases the likelihood of observable clinical endpoints such as e.g. end-stage renal disease in lupus nephritis. Thus, we postulate a model of SLE progression, in which an observed biomarker of SLE would be validated against clinical outcomes, observed many years later, or not observed for a large subgroup of patients. If a biomarker could be validated as a predictor of significant clinical outcomes in the future, this provides strong justification that the biomarker is a valid surrogate of outcome and appropriate for use in trials of relatively short duration.

However, this model of SLE progression has major implications for the design and analysis of the validation study which we summarize. Firstly, the fact that observable clinical outcomes may occur, if at all, only several years later, calls for a retrospective design of the validation study. This will presumably involve a retrospective cohort of SLE patients, for whom biomarker values can be measured, based on sera, collected repeatedly in the past, and frozen for future analyses. The cohort will have to be followed up, in the past, for several years, with some cohort members experiencing the clinical event of interest at some point in time and other censored as being event-free until the end of their follow-up time. The presence of censored observations and the variation in the follow-up duration, both suggest that the association between the biomarker and the risk of the event should be analyzed using time-to-event analysis (19). In our context, the survival analysis of retrospective requires that one decide:

- a. the time window from which the biomarker values will be used for the analyses, and
- b. how these values should be represented, to ensure validity of the results and maximize efficiency of the analysis (20).

A computer-intensive simulation study (21) has provided insights about how an informative repository needs to be amassed. For instance, our simulations showed that the “random time window” design, where a biomarker is obtained in an interval randomly selected from the total follow-up time for that subject, so that each subject contributes maximally to the analysis, yields seriously biased results (21). This, seemingly appealing and ‘efficient’ design results in a type of ‘survival bias’ (22), by artificially altering the relationship between the previous disease course before the biomarkers is obtained, and the time to the clinical event used for validation. Our simulations demonstrated that such bias is avoided by an alternative “fixed calendar window” design, in which biomarkers are measured within the same, pre-specified period for all cohort members who remain at risk during that period, even if this causes some reduction of the effective sample size.

Another crucial finding obtained from our simulations regarded the sample sizes required to validate putative biomarkers of current SLE activity as a surrogate outcome for an outcome in the future. The fact that the putative link between such biomarker and observed outcomes is mediated through two unobservable variables (current disease activity and irreversible damage resulting from cumulative activity) increases the necessary sample size. While sample size estimates clearly depend on a number of assumptions, the requirement that a meaningful clinical state be the major validation criterion increases the required sample size

by an order of magnitude. For example, in a theoretical biomarker which is a very good proxy of current activity, with Pearson correlation of +0.8 only 10 patients with informative serial biomaterials would be needed to ensure 80% power to detect a significant association between the biomarker and the “true” current disease activity. In contrast, the simulations show that (for the same +0.8 correlation with ‘true’ disease activity) the power to establish that the biomarker is a valid predictor of an **event** is below 60% even with 300 subjects with the observed event. Using lupus nephritis where the incidence rate of end-stage renal disease is relatively low, to accumulate 300 instances of end-stage renal disease, a group of at risk patients would have to be followed for about 5,000 person-years, implying 500 subjects followed, on average, for about 10 years. The large numbers of subjects required to validate biomarkers of SLE disease activity against long-term outcomes, makes it essential to maximize the efficiency of study design and analysis and also provides the strongest rationale for collaboration and the use of standardized measures to ensure comparability. A consistent finding from the simulations is that increasing the number of biomarker determinations per subject from one to four or five reduces the required number of subjects by 10-15%, while further increases of the number of observations per subject yielded much smaller gains (21).

Finally, even with the number of biomarker measurements kept constant, the statistical power of an analysis could be improved by selecting an efficient statistical model. Specifically, the model of SLE disease progression on which our simulation relied, we used implied that the risk of the clinical endpoint should be a roughly multiplicative function of disease duration and disease activity, with the latter approximated by the mean of the observed biomarker values. This model suggested the use of an interaction (i.e. product) of the disease duration and mean biomarker value as an efficient representation of the biomarker, and preliminary simulations suggest that such an interaction may increase the power, for a fixed sample size, compared to simpler representations of the biomarker history (21).

RECOMMENDATIONS AND CONCLUSIONS

In this paper we have defined the terminology, validation process, and statistical guidance for establishing a validation bio-repository, screening putative biomarkers and definitive validation studies. Standardizing this process and ensuring valid and efficient design and statistical methods would greatly enhance and accelerate the discovery and evaluation process. A major practical barrier to the field is the availability of informative biomaterial (sera, cells, etc.) of sufficient number which are precisely and thoroughly characterized with respect to their phenotype, medication use (which might suppress certain biomarkers), and SLE disease activity at the time of collection. Attention equal to that paid to data quality of the biomarker data must be devoted to the sampling and the numbers of subjects needed to study the validity of biomarkers and the information collected on the phenotypes observed. A consortium of investigators from the intramural program of NIH and academia are in the early stages of organizing this resource and our analysis addresses these concerns. SLE therapeutics seems on the verge of a breakthrough. Advances in biology, clinimetrics and clinical trials methodology have converged propitiously and may usher in the next generation of advances in pursuit of improved treatment.

Acknowledgments

Supported by a Kirkland Scholar Award, NIH Grant AR47782, R13AR4758401, The Center for Advanced Methodological Support for SLE Clinical Trials (ASSIST), and the Office of the Clinical Director, Intramural Program, National Institute of Arthritis and Musculoskeletal and Skin Disease.

References

1. Illei GG, Tackey E, Lapteva L, Lipsky PE. Biomarkers in Systemic Lupus Erythematosus I. General Overview of Biomarkers and their Applicability Arthritis Rheum. 2004; 50:1709–1720.
2. Illei GG, Tackey L, Lapteva L, Lipsky PE. Biomarkers in Systemic Lupus Erythematosus II. Markers of Disease Activity. Arthritis Rheum. 2004; 50:2048–2065.
3. Firestein GS. A Biomarker by any other name. Nature Clinical Practice Rheumatology. 2006; 2:635.
4. Frank R, Hargreaves R. Clinical Biomarkers in Drug Discovery and Development. Nat Rev Drug Discov. 2003; 2:566–580. [PubMed: 12838269]
5. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther. 2001; 69:89–95. [PubMed: 11240971]
6. Fleming TR, Prentice RL, Pepe MS, Glidden D. Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and AIDS research. Stat Med. 1994; 13:955–68. [PubMed: 8047747]
7. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. Stat Med. 1989; 8:431–40. [PubMed: 2727467]
8. Lesko LJ, Atkinson AJ Jr. Use of biomarkers and surrogate endpoints in drug development and regulatory decision making: criteria, validation, strategies. Annu Rev Pharmacol Toxicol. 2001; 41:347–66. [PubMed: 11264461]
9. Wood AJJ. A proposal for Radical Changes in the Drug-Approval Process. NEJM. 2006; 355:618–623. [PubMed: 16899784]
10. ACR Ad Hoc Committee on Response criteria. The American College of Rheumatology Response Criteria for Systemic Lupus Erythematosus Clinical Trials: Overall Disease Activity. Arthritis Rheum. 2004; 50:3418–3426. [PubMed: 15529383]
11. Braggio S, Barnaby RJ, Grossi P, Cugola M. A strategy for validation of bioanalytical Methods. J Pharm Biomed Anal. 1996; 14:375–88. [PubMed: 8729635]
12. Buick AR, Doig MV, Jeal SC, Land GS, McDowall RD. Method validation in the bioanalytical laboratory. J Pharm Biomed Anal. 1990; 8:629–37. [PubMed: 2100599]
13. Findlay JW, Smith WC, Lee JW, et al. Validation of immunoassays for bioanalysis: a pharmaceutical industry perspective. J Pharm Biomed Anal. 2000; 21:1249–73. [PubMed: 10708409]
14. Buyse M, Molenberghs G. Criteria for the validation of surrogate endpoints in randomized experiments. Biometrics. 1998; 54:1014–29. [PubMed: 9840970]
15. Weir CJ, Walley RJ. Statistical evaluation of biomarkers as surrogate endpoints: a literature review. Stat Med. 2006; 25:183–203. [PubMed: 16252272]
16. Esdaile JM, Abrahamowicz M, Joseph L, MacKenzie T, Li Y, Danoff D. Laboratory tests as predictors of disease exacerbations in systemic lupus erythematosus: why some tests fail. Arthritis Rheum. 1996; 39:370–8. [PubMed: 8607885]
17. Liang MH, Simard JF. The large print giveth; the small print taketh away: Pre-emptive treatment of serologically active, clinically quiet SLE. Arthritis Rheum. 2006; 54:3378–80. [PubMed: 17075813]
18. Renal Disease Subcommittee of the American College of Rheumatology Ad Hoc Committee on SLE Response Criteria. The American College of Rheumatology Response Criteria for Proliferative and Membranous Renal Disease in Systemic Lupus Erythematosus Clinical Trials. 2006; 54:421–432.
19. Kalbfleisch, JD.; Prentice, RL. Wiley and Sons; NY: 1980. The Statistical Analysis of Failure Time Data.
20. Leffondre K, Abrahamowicz M, Siemiatycki J. Evaluation of Cox's model and logistic regression for matched case-control data with time-dependent covariates: a simulation study. Statistics in Medicine. 2003; 22:3781–3794. [PubMed: 14673938]
21. Abrahamowicz, M.; du Berger, R.; Esdaile, JM.; Fortin, PR.; Liang, MH. Selected Methodological Issues in Validation of Biomarkers of Latent Disease Activity. Proceedings of the 27th Annual Meeting of the International Society for Clinical Biostatistics (ISCB); Geneva, Switzerland. 2006; Aug.

22. Zhou Z, Rahme E, Abrahamowicz M, Pilote L. Survival Bias Associated with Time-to-Treatment Initiation in Drug Effectiveness Evaluation: A Comparison of Methods. *Am J Epidemiol*. 2005; 162:1016–1023.