Simulation Studies of Surrogate Endpoint Validation Using Single Trial and Multitrial Statistical Approaches

MARISSA LASSERE, KENT JOHNSON, MICHAEL HUGHES, DOUG ALTMAN, MARC BUYSE, SALLY GALBRAITH, and GEORGE WELLS

ABSTRACT. ***Objective.*** A schema was recently proposed for assessing the levels of evidence for surrogate validity that included 4 domains: Target, Study Design, Statistical Strength, and Penalties. This report examines one component of the schema. It surveys the literature on methods of statistical validation of surrogate markers and compares these methods head-to-head using simulated datasets.

Methods. Simulated datasets (continuous, multivariate normal) were generated to capture 3 possible relationships of surrogate (S) and true (T) outcome (none, weakly positive, strongly positive) each applied to 4 treatment effects (effect on both surrogate and true outcome, effect on neither, effect on surrogate only, and effect on true outcome only). These datasets were analyzed using single and multitrial statistical approaches, and the results were provided to participants for discussion.

Results. The multitrial surrogate threshold effect seemed to capture best the requirement that surrogate validation is demonstrated by a treatment-associated change in the surrogate predicting a treatment-associated change in the outcome.

Conclusion. There was general agreement that neither a single trial nor any of the single trial statistical methods was adequate to establish surrogate validity. These exercises also showed that summary statistics developed specifically to establish surrogate validity, such as the proportion of the effect explained, were problematic. A sizable statistical research agenda remains, which includes investigating the additional advantage obtained with modeling subject-level data compared to modeling with only trial-level data; and developing and testing multitrial statistical approaches robust to settings with only a few trials. (J Rheumatol 2007;34:616–9)

Key Indexing Terms:
SURROGATE BIOMARKER LEVELS OF EVIDENCE STATISTICS VALIDATION

At OMERACT 8 a proposed schema for assessing the levels of evidence for surrogate validity was discussed at a workshop[1]. The schema scores markers across 4 domains: Target, Study Design, Statistical Strength, and Penalties. The issues regarding the statistical validation of surrogates are briefly reviewed, and the results of applying several statistical methods to the examination of simulated and real datasets are presented here. We wanted to engender a free discussion between clinicians and statisticians on the fundamental nature of the statistical aspect of analysis of surrogate and outcome data, and on the relative strengths and weaknesses of methods proposed to date. We also wanted to begin a process of formally comparing different methods with the same datasets, an approach that seemed most transparent to nonstatisticians.

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 New South Wales, Sydney, Australia; K.R. Johnson, MD, Senior Lecturer in Medicine, University of Newcastle, Newcastle, and University of New South Wales, Sydney, Australia; 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, PhD, 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

Overview of the literature

There are 2 general approaches to statistical validation of surrogate outcomes: statistical analysis of single trials and statistical analysis of multiple trials (metaanalytic assessments). These are briefly summarized, and we direct interested readers to reviews published in the statistical literature[2]. Both approaches usually require that a change in a surrogate outcome is associated with or predicts a change in the true outcome.

Single-trial statistical approaches. The first formal approach to testing for statistical validity of surrogacy using the single-trial approach was proposed by Prentice in 1989[3], although there had been earlier instances of assessing trial data of clinical outcomes from multiple trials in order to inform registration deliberations for a new drug in the same class, and instances of trial-based modeling of outcome change versus surrogate change[4]. However, Prentice was the first to generalize the process and articulate "all or nothing" criteria for its validity. Additionally, during the 1980s the promotion of surrogates supported only with observational data began to be

questioned. The most prominent example of this was the conduct and publication of the Cardiac Arrhythmia Suppression Trial (CAST)[5]. Such upsets have remained a spur to trialists and methodologists in their quest for criteria for surrogate validity. In the early 1990s the issue was given additional urgency by hopes that the CD4 marker would be a valid surrogate and thus help accelerate drug development in HIV/AIDS.

In contrast to the "all or nothing" approach, Freedman, *et al*[6] in 1992 suggested a graded criterion for surrogacy validity depending on the relation between treatment and outcome (the coefficient of the treatment term in the model reflecting the steepness of the slope) and its change when the surrogate is introduced into the model. A surrogate is more valid to the degree that the treatment coefficient is reduced when the surrogate is added. The Freedman criterion is called the "proportion of treatment effect explained" (PTE). However, others have commented on the conceptual and mathematical difficulties of this approach[7].

Multiple-trial statistical approaches. By the mid 1990s a number of proposals moved beyond the single-trial methods of Prentice and Freedman to methods applicable to the simultaneous analysis of multiple trials[8-10]. These included both frequentist and Bayesian metaanalyses, some with hierarchical modeling. Importantly, some methods now proposed modeling a comparison of the outcome in the treated versus the control arm of randomized trials as a function of a comparison of the surrogate in the treated versus the control arm. This approach captured the essential difference between a validated surrogate marker and a validated prognostic marker, where the latter implies the marker predicts the outcome ("predictive validity") and the former implies a change in the marker associated with treatment predicts a change in the outcome. However, these approaches are not mutually exclusive, and a marker could be both. These methods are based on multivariate regression modeling of trials (or, both trials and patients in hierarchical regression models). Recent work[11-13] has used models and calculated prediction (forecast) bands[14] that show the range (within 95% limits) for the predicted outcome for a particular future trial.

Head-to-head comparison of statistical validation of surrogate data using simulated trial datasets

To begin the process of directly comparing the different statistical methods in head-to-head comparisons, a series of simulated trial data results were generated. The conventional notation for the statistical validation of surrogate outcomes using clinical trial datasets is: S, the "surrogate" variable; T, the "true endpoint" variable; and Z, the "treatment" binary variable (where the control group has the value of 0 and the treatment group the value of 1). For this initial exercise, both the surrogate and true outcome were assumed to be continuous and have a multivariate normal distribution in a 2-arm trial (single treatment arm and single control arm). The setting was designed to mimic rheumatoid arthritis (RA) with the surrogate, S, being a soluble cytokine marker elevated in active RA, and therefore the focus of treatment Z, a monoclonal antibody. S is measured continuously on a 0 (best) to 1000 (worst) scale. The outcome, T, a patient-centered outcome of function, is a scale like the Health Assessment Questionnaire (HAQ), but disaggregated into 24 possible ratings (0–3 for each of its 8 categories), therefore ranging from 0 (best) to 24 (worst). The 24-point interval scale is close enough to a continuous variable for our purposes. The aim of therapy is a reduction in S with the hope that this will translate into a reduction in T.

Simulated trial datasets were generated with random-number simulations using the statistical packages Stata 9 (Stata Corp., College Station, TX, USA) and R (The R Foundation for Statistical Computing, Vienna, Austria. Available from: http://www.r-project.org/). These simulated datasets were designed to capture the various correlations of surrogate and true outcome (3 possible correlations: none, weakly positive, strongly positive), each applied to various treatment effects (4 possibilities: effect on both surrogate and true outcome, effect on neither, effect on surrogate only, and effect on true endpoint only). Twelve single-trial datasets were generated, each with 200 subjects, 100 per treatment arm. Two multitrial data sets, each with 20 trials, again each with sample size 200, were also generated. One multitrial set consisted of results contrived to support surrogacy, and the other intentionally made an admixture of trials with conflicting results. All datasets are available on request. The dataset relationships were "blinded" and analyzed using the following approaches: correlation between S and T unadjusted and adjusted for treatment; coefficient of S on T, adjusted for treatment, and its significance; single-trial R-squared (R^2, proportion of variance explained by the model, a measure of dispersion) of T given S, unadjusted and adjusted for treatment; Proportion of the Treatment effect Explained (PTE), Relative Effect (RE)[9], multitrial R^2_{trial}, $R^2_{individual\ ST}$, R^2_{total}; and Surrogate Threshold Effect (STE)[12,13]. The results were presented to the group and discussed to evaluate which method or methods were the most useful for judging whether S was a valid surrogate for T.

There was a clear preference for multitrial over single-trial approaches. In the single-trial approaches, the participants appeared to rely mainly on whether the treatment coefficient was statistically significant in analyses without and with the surrogate term. The participants also appeared to rely on the R^2 measures of the models, and there was consensus that a surrogate with less dispersion (larger R^2) was preferable to one with more dispersion. Rarely did the PTE appear helpful, and some of the results of this statistic were difficult to interpret. The Relative Effect was also difficult to interpret.

In the multitrial setting, hierarchical models to obtain R^2_{trial}, $R^2_{individual}$ were unsuccessful because of failure of convergence. Most participants found the STE most promising. In this approach the trial rather than the individual is the unit of analysis, and one directly adjusts for Z (the treatment vari-

able). The STE provides (1) the statistical significance of the surrogate–true outcome relationship; (2) the explained variance of the surrogate–true outcome relationship (R^2); and (3) the numerical threshold (in the units of the measure) above which a surrogate benefit reliably estimates a true outcome benefit. This is best illustrated in the following 2 examples. In multitrials 1 and 2 (Figure 1) the horizontal axis is the difference, treatment arm versus control arm, of the cytokine level, and the vertical axis is the difference, treatment arm versus control arm, in physical function. The broken line is the mean regression line (fixed effects model), the shaded narrower band is the 95% confidence band for the mean regression, and the solid-line wider band is the 95% prediction band for an individual trial. The critical point (the surrogate threshold) is

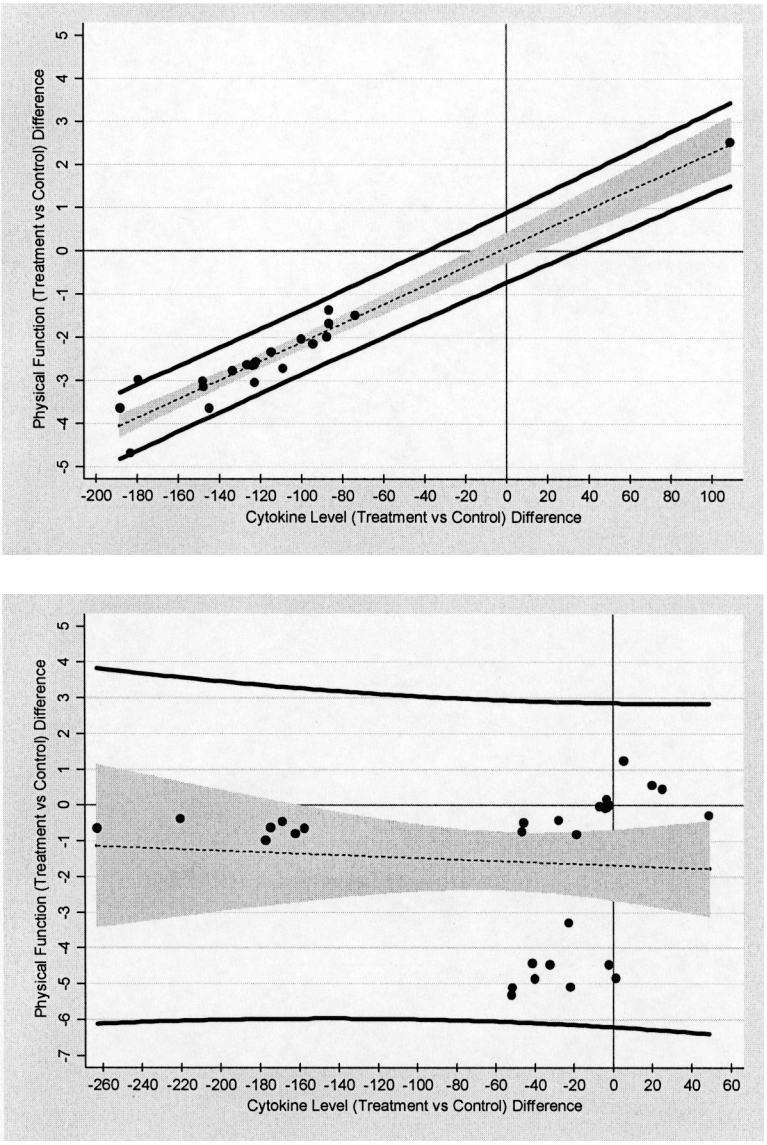

Figure 1. Surrogate threshold effect (STE) in simulated multitrial datasets. Difference (treatment vs control) of soluble cytokine levels (horizontal axis) versus difference (treatment vs control) in physical function (vertical axis). Negative numbers denote treatment better than control. Upper graph: simulated multitrial dataset 1 generated to support surrogate-true outcome relationship. Lower graph: simulated multitrial dataset 2 generated to show no surrogate-true outcome relationship. Broken line, gray band: mean regression line and its 95% confidence band; solid lines: limits of the 95% prediction band for an individual trial. In the upper graph, a difference of cytokine levels between treatment and control of –40 units or more has a 95% CI that excludes a nil difference of change in physical function. In other words, at or above this threshold an effect on the cytokine levels reliably predicts a favorable effect on the true outcome. In the lower graph, no level of cytokine reduction demonstrates a physical function benefit as all y-values are greater than zero.

where the 95% prediction band fully excludes no physical function benefit, i.e., the entire 95% prediction band lies below the x-axis. In multitrial 1 (the upper graph), only trials with mean cytokine reduction of ≥ 40 mmol/l are associated with physical function benefit, whereas in multitrial 2 (lower graph), at no point is cytokine reduction associated with physical function benefit. In multitrial 1 the adjusted R^2 of the regression model was high (0.93) and the model was statistically significant ($p < 0.001$). Even with removal of outliers from this model, the threshold remains at around –40, although the adjusted R^2 is now 0.78. In multitrial 1, the regression line is oriented at about 35 degrees. If the line were steeper, then a wider prediction band could be tolerated, still with the same threshold of –40. By contrast, in multitrial 2 the adjusted R^2 was 0.00 and the model was not significant ($p = 0.7$).

Conclusions and research agenda

In these simulated dataset exercises, where both the surrogate and true outcome assumed a continuous and multivariate normal distribution in a 2-arm trial (single treatment arm and single control arm), there was consensus that none of the single trial approaches [correlation between S and T unadjusted and adjusted for treatment; coefficient of S on T adjusted for treatment and its significance; single-trial R-squared (proportion of variance explained by the model) of T given S unadjusted and adjusted for treatment; Proportion of the Treatment effect Explained, Relative Effect] individually was adequate to establish surrogate validity. These exercises also showed that summary statistics developed specifically to establish surrogate validity, such as the PTE, were problematic. There was general agreement that a single trial was not adequate to establish surrogate validity, not only because there is no appropriate statistical test but also because one should be wary, in principle, of relying on a single trial. Multitrial approaches were preferred to the analysis of single trials, and the multitrial STE approach to the statistical validation of surrogate data, which adjusts directly for Z (the treatment variable). It seemed to us, in principle, that the multitrial approach with the STE may even be superior to multiple single-trial analyses of trials for which there are subject-level data.

The statistical research agenda is sizeable. To date the majority of datasets for the statistical evaluation of surrogate validity come from cardiology, oncology, and HIV/AIDS, all of which have many trials, with many classes of interventions for analysis. Even in these disciplines, subject-level data are not always freely available. Therefore, multitrial approaches involving a smaller number of trials and the incremental advantage of modeling subject-level data above trial-level data alone are important areas for further investigation. In our exercises both the surrogate and true outcome assumed a continuous and multivariate normal distribution in a 2-arm trial. Other scenarios require evaluation: noncontinuous variables (such as binary or time-to-event variables), non-normal distributions, datasets with significant treatment-surrogate interactions; and datasets where treatment effects are partially mediated independent of the surrogate. We anticipate that continuing interdisciplinary dialog will move the research agenda forward.

REFERENCES

1. Lassere MN, Johnson KR, Boers M, et al. Definitions and validation criteria for biomarkers and surrogate endpoints: development and testing of a quantitative hierarchical levels of evidence schema. J Rheumatol 2007;34:xxxxx.
2. Haynes RB, Sackett DL, Guyatt GH, Tugwell P. Clinical epidemiology: How to do clinical practice research. Philadelphia: Lippincott Williams & Wilkins; 2005.
3. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. Stat Med 1989;8:431-40.
4. A'Hern RP, Ebbs SR, Baum MB. Does chemotherapy improve survival in advanced breast cancer? A statistical overview. Br J Cancer 1988;57:625-28.
5. Cardiac Arrhythmia Suppression Trial (CAST) Investigators. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. N Engl J Med 1989;321:406-12.
6. Freedman LS, Graubard BI, Schatzkin A. Statistical validation of intermediate endpoints for chronic diseases. Stat Med 1992; 11:167-78.
7. De Gruttola V, Fleming T, Lin DY, Coombs R. Perspective: validating surrogate markers — are we being naive? J Infect Dis 1997;175:237-46.
8. Daniels MJ, Hughes MD. Meta-analysis for the evaluation of potential surrogate markers. Stat Med 1997;16:1965-82.
9. Buyse M, Molenberghs G. Criteria for the validation of surrogate endpoints in randomized experiments. Biometrics 1998;54:1014-29.
10. Buyse M, Thirion P, Carlson RW, Burzykowski T, Molenberghs G, Piedbois P. Relation between tumour response to first-line chemotherapy and survival in advanced colorectal cancer: a meta-analysis. Meta-analysis Group in Cancer. Lancet 2000;356:373-8.
11. Sargent DJ, Wieand HS, Haller DG, et al. Disease-free survival versus overall survival as a primary end point for adjuvant colon cancer studies: individual patient data from 20,898 patients on 18 randomized trials. J Clin Oncol 2005;23:8564-5.
12. Burzykowski T, Buyse M. An alternative measure for meta-analytic surrogate endpoint validation. Chapter 18. The evaluation of surrogate endpoints. New York: Springer; 2005.
13. Johnson KR, Ringland C, Stokes BJ, et al. Response rate or time to progression as predictors of survival in trials of metastatic colorectal cancer or non-small-cell lung cancer: a meta-analysis. Lancet Oncol 2006;7:741-6.
14. Altman DG. Practical statistics for medical research. London: Chapman and Hall; 1991.