Computational Statistics and Data Analysis 78 (2014) 1–20

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Center-within-trial versus trial-level evaluation of surrogate endpoints

Lindsay A. Renfro [a,*], Qian Shi [a], Yuan Xue [b], Junlong Li [c], Hongwei Shang [d], Daniel J. Sargent [a]

[a] Division of Biomedical Statistics and Informatics, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, United States
[b] Department of Statistics, University of Virginia, United States
[c] Department of Biostatistics, Harvard School of Public Health, United States
[d] HP Labs, United States

ARTICLE INFO

Article history:
Received 4 December 2013
Accepted 21 March 2014
Available online 8 April 2014

Keywords:
Clinical trials
Meta-analysis
Surrogate endpoints
Survival analysis

ABSTRACT

Evaluation of candidate surrogate endpoints using individual patient data from multiple clinical trials is considered the gold standard approach to validate surrogates at both patient and trial levels. However, this approach assumes the availability of patient-level data from a relatively large collection of similar trials, which may not be possible to achieve for a given disease application. One common solution to the problem of too few similar trials involves performing trial-level surrogacy analyses on trial sub-units (e.g., centers within trials), thereby artificially increasing the trial-level sample size for feasibility of the multi-trial analysis. To date, the practical impact of treating trial sub-units (centers) identically to trials in multi-trial surrogacy analyses remains unexplored, and conditions under which this ad hoc solution may in fact be reasonable have not been identified. We perform a simulation study to identify such conditions, and demonstrate practical implications using a multi-trial dataset of patients with early stage colon cancer.

1. Introduction

1.1. Background and motivation

Surrogate endpoints are desired in clinical trials where traditional endpoints are too expensive or difficult to obtain, or where substantial follow-up would be required to observe the clinical endpoint (e.g., survival) in a sufficient number of patients to draw meaningful trial conclusions. While numerous methods for evaluating and validating surrogate endpoints have been proposed, recent consensus has supported evaluation of potential surrogates based on patient-level data from multiple similar trials, where surrogate performance is assessed both within trials (i.e., at the patient level) and across trials (trial level). A surrogate endpoint is considered to be *validated* for use in future clinical trials of the same disease setting when both strong patient-level surrogacy and strong trial-level surrogacy are present.

Central to this multi-trial surrogacy evaluation paradigm is the availability of patient-level data from a relatively large number of randomized clinical trials within the disease setting where the surrogate endpoint is proposed for use. Within a comprehensive surrogacy analysis, a strong association between a candidate surrogate endpoint S and a true

* Corresponding author. Tel.: +1 507 284 3202.
E-mail address: renfro.lindsay@mayo.edu (L.A. Renfro).

http://dx.doi.org/10.1016/j.csda.2014.03.011
0167-9473/

clinical endpoint T must be present, where this *patient-level surrogacy* is traditionally quantified as a simple correlation where possible, or evaluated through a multi-trial joint model for S and T, such as a copula model (Burzykowski et al., 2001), otherwise. Arguably of equal or greater importance is *trial-level surrogacy*, which may be demonstrated by a strong predictive relationship (e.g., correlation) between treatment effects on S and treatment effects on T. That is, the experimental treatment's observed effect on a valid surrogate endpoint should provide a strong indication of the experimental treatment's (unobserved) effect on the clinical endpoint. In theory, patient-level surrogacy may be established using patient-level data from one or more historically similar clinical trials, through straightforward joint modeling or correlation (where censoring is not an issue) of the two endpoints S and T. On the other hand, a trial-level surrogacy analysis necessitates access to a *collection* of historical clinical trials, the number of which should be sufficient to compute – at a minimum – the correlation of the (estimated) treatment effects on S and T across trials, along with an associated measure of uncertainty (e.g., standard error). For a surrogacy analysis to be truly informative for clinical decision-making, the standard error associated with an estimate of trial-level surrogacy should be sufficiently small to distinguish a strong surrogate from a weak surrogate, which in turn requires a relatively large "trial-level" sample size (see, e.g., Shi et al. (2011)).

In many practical applications, patient-level data from a large number of comparable randomized trials are difficult or impossible to obtain. Challenges may include: reluctance by data owners to share patient-level data with other parties, lack of time, resources, or expertise to successfully define and pool data elements from a large number of disparate trials into a single database, or non-existence of a large number of similar trials within a specified disease setting and class of treatments. Where a surrogacy analysis is desired but one or more of these issues cause only a few (say, one to five) trials to be available for analysis, a common ad-hoc solution is to perform trial-level surrogacy analyses on trial sub-units, such as centers, investigators, or geographic regions within trials, as if these sub-units were themselves unique trials.

1.2. Published uses and explorations of trial sub-units in surrogacy evaluation

Published examples estimating trial-level surrogacy using trial sub-units for analysis include: evaluation of time to progression and progression-free survival as surrogates for overall survival in advanced ovarian cancer, where centers within trials are treated as the trial unit (Buyse et al., 2000; Burzykowski et al., 2001; Molenberghs et al., 2002; Tibaldi et al., 2003; Burzykowski and Buyse, 2006); change in visual acuity at 6 months after treatment as a surrogate for change in visual acuity at 12 months in age-related macular degeneration, where centers are treated as trial units (Buyse et al., 2000; Molenberghs et al., 2001, 2002; Tibaldi et al., 2003; Alonso et al., 2004b, 2006; Pryseley et al., 2007; Abrahantes et al., 2008; Molenberghs et al., 2008); progression-free survival as a surrogate for overall survival in advanced colorectal cancer, with centers as trial units (Burzykowski et al., 2001; Molenberghs et al., 2002; Tibaldi et al., 2003; Burzykowski and Buyse, 2006; Abrahantes et al., 2008); outcomes of the Positive and Negative Syndrome Scale (PANSS) as a surrogates for the Clinician's Global Impression (CGI) scale in schizophrenia, where treating physicians, main investigators, or countries were considered as trial-level replicates (Molenberghs et al., 2002; Renard et al., 2002; Alonso et al., 2002, 2003, 2004a, 2006; Tilahun et al., 2007; Alonso and Molenberghs, 2007; Abrahantes et al., 2008; Molenberghs et al., 2008, 2010); prostate specific antigen (PSA) as a surrogate for overall survival in advanced prostate cancer, where country was used as the trial unit (Renard et al., 2003; Molenberghs et al., 2004); recurrence-free survival as a surrogate for overall survival in colon cancer, with grouped centers treated as the trial unit (Sertdemir and Burgut, 2009); leukemia-free survival as a surrogate for overall survival in maintenance therapy trials for patients with acute myeloid leukemia in complete remission, where countries within a single trial were treated similarly to trials (Buyse et al., 2011); pathologic complete response and local control as surrogates for overall survival in advanced rectal cancer, where grouped centers were treated as trial units (Bonnetain et al., 2012); and progression-free survival as a surrogate for overall survival in advanced non-small-cell lung cancer, where centers within trials was the unit of assessment of trial-level surrogacy (Laporte et al., 2013).

Although use of trial sub-units in place of trials is commonplace among published trial-level surrogacy analyses, the impact of disregarding the subunit-within-trial hierarchy in these convenient substitutions is relatively unexplored. For the case of two normal endpoints S and T, Abrahantes et al. (2004) performed a simulation study to compare trial-level versus center-level surrogacy estimation as a function of other key factors, such as number of trials, equal versus unequal association of treatment effects at the trial and center levels, and relative variability of trial versus center-level effects. They found that when data contains both trial-specific and center-specific treatment effects, and when these treatment effects truly have the same association across trials as within trials, using center as the unit of measurement to assess surrogacy (rather than trial) does not adversely influence results. However, when unequal association of treatment effects across trials versus within trials is assumed, center-level estimation often over-estimated moderate surrogacy and under-estimated high surrogacy. This observed weakness of naive center-level surrogacy estimation is alleviated when the variability of treatment effects among centers within trials is constrained to a small fraction ($1/100$) of the variability of treatment effects across trials—a scenario the authors argue is desirable, but seems unlikely to be observed in practice. The practical effects of naive center-level versus trial-level surrogacy evaluation have not been explored to date with non-normal endpoints, such as time-to-event endpoints, which are of particular relevance and importance in settings where surrogates are desired specifically because they occur earlier or more often in the population of interest. In addition, the effect of unit choice on patient-level surrogacy estimation is previously unexplored.

In this paper, we compare the performance of common surrogacy estimation methods when applied to trials versus application to sub-units within trials, and focus on the case of two time-to-event endpoints S and T. In Section 2, we present