Review: Clinical Trial Methodology

CLINICAL INVESTIGATION

Surrogate end points: when should they be used?

Clin. Invest. (2013) 3(12), 1147–1155

Surrogate end point evaluation has received a lot of attention for approximately a quarter of a century. Throughout this time, the topic has been surrounded with both hopes and perils. The history of the evaluation of surrogate end point evaluation is described, from Prentice's definition, via single-trial evaluation to meta-analytic approaches. While based on technical developments and advances in statistical methodology, the treatment of the topic here is largely nontechnical. A perspective is given as to what might be a sensible way forward. It is clear that statistical evaluation can offer a contribution to the discussion as to whether a surrogate will be adopted and in what form, but can never be seen as the sole decision maker. While there is room for the adoption of surrogates, it is very important that pitfalls and drawbacks be kept in mind at all times.

Keywords: adjusted association • biomarker • causal inference • information theory • meta-analytic framework • proportion explained • surrogate end point • surrogate marker • surrogate threshold effect

Geert Molenberghs*[1,2], Ariel Alonso Abad[3], Wim van der Elst[1], Tomasz Burzykowski[1,4] & Marc Buyse[1,4]

[1]I-BioStat, Hasselt University, Hasselt, Belgium
[2]I-BioStat, University of Leuven, Leuven, Belgium
[3]Methodology and Statistics, Maastricht University, Maastricht, the Netherlands
[4]International Drug Development Institute, Louvain-la-Neuve, Belgium
*Author for correspondence:
Tel.: +32 11 268 238
E-mail: geert.molenberghs@uhasselt.be

A quarter of a century ago, the seminal papers on surrogate end points in clinical trials by Prentice on the one hand [1] and by Freedman and colleagues on the other [2], instigated a large and relatively new research line. It is fair to say that surrogate end points did not become an instant success. Problems with the logical and formal framework surrounding surrogate end points were compounded by issues relating to the lack of data and, to a lesser extent, the lack of promising candidate surrogates. There were even unfortunate instances that one could classify as 'accidents'; for example, the increased mortality caused by antiarrhythmic drugs. In spite of this, surrogate end points have always enjoyed a certain amount of interest, owing to the compelling prospects of important savings in terms of trial duration and size. Evidently, such gains are to be understood in ethical as well as economic terms.

In the meantime, a large body of research has been developed. A state-of-the-art snapshot, as it was perceived a decade ago, has been laid out in the edited volume of Burzykowski, Molenberghs and Buyse [3]. As is often the case, a few seemingly competing schools of thought have emerged, most commonly referred to as the meta-analysis and the causal schools.

Currently, surrogate end points are gaining clout, not only because more and more candidate surrogates are becoming available, but also because contemporary methodological developments are accompanied by efforts of unification.

Terminology & concepts

The key variables in a controlled clinical trial are the treatment, denoted by 'Z', and the clinical (or true) end point, denoted by 'T'. It is convenient to assume that Z is dichotomous; for example, active treatment (Z = 1) versus placebo (Z = 0), or experimental treatment (Z = 1) versus control treatment (Z = 0). It is possible for Z to have more than two modalities, and even for it to be accompanied by additional predictor variables (such as age, gender, baseline measurements, and so forth). This will not

FUTURE SCIENCE part of fsg

add substantial difficulties to the line of reasoning followed here and therefore attention will be confined to dichotomous Z, for simplicity's sake. The clinical end point T can be a continuous, binary, count, or ordinal variable, or it can be a (possibly censored) time-to-event. Furthermore, T can be measured once, usually at the end of the trial, or it can be measured repeatedly over time, for any one of the data types mentioned earlier. The following will be as generic as possible.

Next to Z and T, the third variable on the field is the surrogate end point, denoted by 'S'. It is useful to lay out the concepts and terminology, used by the Biomarker Definitions Working Group [4]. The clinical end point has been mentioned already. A biomarker is "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacological responses to a therapeutic intervention" [4]. As such, a biomarker is very broad. A surrogate marker or surrogate end point is defined as "a biomarker that is intended to substitute for a clinical end point; a surrogate end point is expected to predict clinical benefit (or harm, or lack of benefit or harm)" [4]. One might distinguish between a surrogate end point, broadly one used in lieu of the true end point in a trial, under conditions laid out in what follows, and a surrogate marker, which is a surrogate end point of the biomarker type. This indicates that there are surrogate end points that are not biomarkers. One such example is the use of one psychiatric rating scale (e.g., Brief Psychiatric Rating Scale) as surrogate S instead of another one that is considered the true end point T (e.g., Positive and Negative Symptoms Scale or Clinicians) or instead of a true end point 'T' that is of a different nature altogether (e.g., the 7-point ordinal Clinician's Global Impression).

In conclusion, the assignment is to study the triple (Z, T, S), where Z is binary, and S and T are end points of any type encountered in clinical trial practice. S and T can be of the same type but do not have to be. An example of differing end points is when S is tumor response (binary or categorical) in oncology, with T time-to-progression (time-to-event). When S is time-to-progression or death, and T is time to death, then S and T are different even though both of them are event times. Yet another situation arises when S and T are the same end point but ascertained at different times, such as, for example, when S and T are visual acuity of patients with age-related macular degeneration at 6 months and 12 months, respectively.

The preprentice era

The aforementioned potential benefits (savings in terms of time and patients) are so appealing that surrogates were used long before they were formally studied [5]. At the same time, they were surrounded with controversy [6]. This can be explained in part by a number of unfortunate instances. One of the better known incidents is the US FDA approval of a triple of anti-arrhythmia drugs (encainide, flecainide and moricizine), because they effectively suppressed arrhythmias. As a result of the relationship between arrhythmias and cardiac-related deaths, it was believed that the drugs would also reduce death rate. In other words, arrhythmias were assumed to be a surrogate for death. Unfortunately, the death rate among the active patients treated with encainide and flecainide was twice that among placebo patients [7].

The lesson learned was that a mere association (or correlation) between a potential surrogate and a clinical end point is not sufficient to 'validate' a surrogate. Indeed, the question to be asked is what type of relationship is expected to exist between S and T. It is not simply a matter of replacing correlation with another measure. Rather, one has to reflect on the level at which a correlation exists. The authors will return to this when touching upon the meta-analytic framework. Note that validation is placed between inverted commas, because it comes with the connotation that the process involved is dichotomous: a candidate passes the test of surrogacy or fails to do so. However, it is fair to say that surrogacy comes in shades of grey, from surrogates that are definitely not valid to those that may be valid under some specific circumstances, all the way to those that are universally valid. It is therefore better to use the term 'evaluation,' which refers to the quantification of surrogacy according to one or more criteria. Such a surrogacy measure can then be used as a component in the decision process, rather than being a decision in itself. It is at the same time a more modest and more realistic approach.

Surrogacy as a response to fast-paced events

Evaluating a surrogate takes time and energy and, especially in a multiple-trial setting (see the meta-analytic framework below), large amounts of data are required. When a therapeutic area enters an era of fast-paced evolution, time and/or data may be lacking. One such example is HIV/AIDS [8]. Until the early 1990s, AIDS was an acute disease that almost always led to death over a relatively short time span. Trials could be designed based on time-to-death as a clinical end point, without the need of a surrogate. The first modest therapeutic successes changed this somewhat. At that time, CD4 count came into use as a surrogate end point or, more accurately put, as an alternative end point. We now know that even repeatedly measured CD4, while related to the clinical end point of mortality, was not the best surrogate [9,10].

Indeed, there is a lot of instantaneous variability in the measurement, because the immune system quickly responds to minor attacks, and because CD4 value was, and to some extent still is, hard to measure with high precision.

Matters changed again with the ability to measure viral load (VL). While CD4, at least in theory, reflects important information on the status of the immune system, VL is even more directly related to therapeutic success or absence thereof. One could then use VL alone or VL combined with CD4. Note that both can be considered continuous end points.

A further dramatic change occurred with the advent of highly active antiretroviral therapies, in so far as these therapies routinely lowered VL to undetectable levels. While this evolution is evidently beneficial, it poses a further methodological challenge. Statistically, VL is considered a potentially 'truncated' outcome. An easy and at the same time intuitively appealing solution is to use 'VL below detectable limit' (VL-BDL) as an end point in its own right. The original time-to-event end point has become binary (VL-BDL yes/no), via two continuous intermissions. Technically, VL-BDL is still a surrogate for the original clinical end point, 'time to death'. However, given the tremendous gap between the time where AIDS was an acute disease and its current-day chronic status, it is arguably more sensible to talk about a shift in clinical end point rather than the adoption of a (non)validated surrogate.

Surrogacy's quadrature of the circle: the class

The above example brings out a generic issue. The evaluation of a potential surrogate is not context-free. For example, it is not possible to say whether progression-free survival is a good surrogate for survival in advanced forms of cancer. It might be in advanced colorectal cancer but not in advanced breast cancer and less so in advanced ovarian cancer [3]. This example demonstrated that not only the broad therapeutic area but also the specific disease should be taken into account. Furthermore, an important but difficult question is concerned with the breadth of the drug class over which a surrogate is considered properly evaluated. For example, a surrogate considered acceptable over a class of traditional cytotoxic drugs should perhaps not be used without further evaluation over a class of monoclonal antibodies. Thus, drastic changes in therapeutic behavior, stemming from emerging novel drugs or procedures, would in the worst case require one to return to the drawing board and restart the evaluation process. Evidently, there is no clear-cut, let alone a mathematical, answer as to when the continued use of a previously validated surrogate is still warranted. Judicious deliberation by subject-matter experts is needed.

Prentice's definition

Prentice defined a valid surrogate as one where testing the null hypothesis of no treatment effect on the true end point T is equivalent to testing the null hypothesis of no treatment effect on the surrogate end point S [1]. Appealing though the definition is, it is hard to make operational in practice, for various reasons. First, 'equivalent' is a fuzzy concept in this context, in spite of a precise mathematical connotation. Indeed, the definition cannot mean that the two null hypotheses would be jointly rejected or not rejected in each and every case. Then, the question is just how much discrepancy could be tolerated. Second, while a pair of nonrejections is concordant and hence apparent evidence in favor of surrogacy, it could merely be the result of too small a sample size, reminiscent of the well-known issue with equivalence trials. Prentice was well aware of this issue and offered a set of criteria in the same paper. These form the basis of the single-trial framework, to be discussed next [1].

The single-trial framework

The first three of Prentice's criteria are as follows:

- Criteria 1: The treatment Z has an effect on the true end point T;
- Criteria 2: The treatment Z has an effect on the surrogate end point S;
- Criteria 3: The surrogate end point S has an effect on the true end point T.

At first sight, these criteria are appealing and intuitive. Unlike the definition, they can be tested directly from a set of data, provided that for every patient the triplet (Z, T, S) is available. They can be summarized by stating that each member of the triplet has an impact on the other two.

The fourth criterion, sometimes simply called Prentice's criterion, is:

- Criteria 4: There is no further effect of the treatment Z on the true end point T, after correcting for the surrogate S (Figure 1).

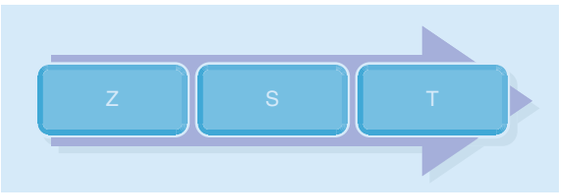

Figure 1. Prentice's criterion.
S: Surrogate end point; T: Clinical (or true) end point; Z: Treatment.

In other words, all the effect of the treatment on the true end point is mediated via the surrogate end point. This would happen if the surrogate were an integral part of the mechanism of the disease or, in other words, of the causal chain leading to disease. An example would be where Z refers to dietary changes, S is the occurrence of colorectal polyps and T is the occurrence of colorectal adenocarcinomas. However, if S is hyperproliferation, but there is a route from treatment to end point via apoptosis as well, then the surrogate is no longer perfect.

The above example indicates that Prentice's criterion 4 is intuitively plausible. However, there are nontrivial issues with the criteria [11]. First, the criteria are neither sufficient nor necessary for the definition, except when all members of the triplet (Z, T, S) are binary. This implies that verifying the criteria does not guarantee that the surrogate fulfills the definition. Second, while criteria 1–3 require a null hypothesis to be rejected, the reverse is true for criterion 4. This raises issues of equivalence testing.

For these reasons, criterion 4 was taken up by Freedman, Graubard, and Schatzkin who proposed a measure to be estimated from the data, the so-called proportion explained (PE; also known as proportion of treatment effect explained) [2]. Its definition is relatively simple. Let β be the treatment effect on the true end point (termed the unadjusted treatment effect) and β_S the treatment effect on the true end point after correcting for the surrogate (termed the adjusted treatment effect).

$$PE = \frac{\beta - \beta_s}{\beta}$$

Equation 1

To fix ideas, both could be estimated from simple linear regressions if S and T were Gaussian. Note that β follows from a model corresponding to criteria 1 and β_S from a model corresponding to criteria 4. Again, this is at first sight an intuitive and worthwhile measure, inspired by the attributable fraction in epidemiology. The intuition is that, if $\beta_S = 0$ (all effect mediated) then PE = 1, whereas if there is no mediation ($\beta = \beta_S$), PE = 0. However, the above is flawed, because β_S is not necessarily zero when there is full mediation, and β and β_S are not necessarily equal when there is no mediation. As a result, the PE is not even confined to the unit interval. This is not merely an academic observation, but it has been shown to happen in a variety of applications [12]. A technical treatment of these issues can be found in the literature [13,14]. In summary, the PE can be very misleading and it has fallen out of favor.

An important feature of a properly defined measure, is that it shifts from the relatively dichotomous hypothesis testing paradigm to estimation. In this spirit, Buyse and Molenberghs proposed two new measures of surrogacy: the adjusted association ρ and the relative effect (RE) [15]. The adjusted association is nothing other than the correlation (for continuous end points at least) between S and T, after correcting for Z. For non-Gaussian end points, appropriate association measures (Kendall's tau, the odds ratio and so forth) have to be used. The relative effect is the ratio $RE = \beta/\alpha$, where α is the treatment effect on the surrogate end point. The idea is that, if RE is sufficiently precisely estimated, it could be used to predict the treatment effect in a new trial, even without measuring the true end point, using the relationship $\beta_0 = RE \times \alpha_0$, where α_0 (β_0) is the treatment effect on the surrogate end point (true end point) in the new trial. The concept of ρ is relatively appealing, and can be estimated with sufficient reliability given that patients are ordinarily replicated within trials. However, the RE is based on a single trial only. It can be estimated because a multiplicative relationship is assumed. This comes down to a regression line through (0,0) and (α,β), a strong and, moreover, unverifiable assumption.

The latter problem stems from there being only one trial to verify the treatment effects. The solution to this is relatively simple, of course: conduct validation in several trials, a topic that will be touched upon next.

The meta-analytic framework

One of the main problems with single-trial validation is that a single trial replicates patients and, hence, provides a basis for inference about patient-related characteristics, such as the correlation between S and T within a patient, but not about characteristics that replicate from trial to trial, rather than from patient to patient. This is a serious problem, because we are interested in the agreement between β and α, the treatment effects on the true and surrogate end points, respectively. While the RE was designed to capture this agreement, it unfortunately does so by making strong and unverifiable assumptions. Even when the multiplicative relationship behind the RE would be correct, there is still a problem. To see this, rewrite the relation as $\beta = RE \times \alpha$ or, to make it slightly more general, $\beta = \mu + RE \times \alpha$ that is, with an intercept included. The question then, is how accurate is this relationship. To study this in proper statistical terms, a final rewrite is necessary: $\beta = \mu + RE \times \alpha + \varepsilon$, where an error term is included. The magnitude of the error, is usually captured by σ^2 where ε follows a normal distribution with zero mean and variance σ^2. This means that, no matter how correct the regression relationship is, σ^2 can be determined only when there is appropriate replication. Here, replication is taken to mean multiple copies of the pair (β,α), which is equivalent to saying

that multiple trials are necessary. Thus, the regression can be finalized as:

$$\beta_i = \mu + RE \times \alpha_i + \varepsilon_i$$

Equation 2

where i = 1,…,N refers to trials. If the regression relationship is correct and σ^2 is small, then, in a new trial, β_0 is very well predictable from α_0. A consequence of these developments is that, next to hypothesis testing and the estimation of surrogacy measures, one now considers prediction. Arguably, prediction, and measures that quantify the quality with which it can be done, should be central to surrogate marker evaluation.

In this spirit, the need for replication at the trial level was noticed by several authors [12,16–18] and developed into a full theory. The above regression for β_i reflects replication at the trial level, on top of the already present replication at the individual level. Thus, the triplet could be denoted by (Z_{ij},T_{ij},S_{ij}), where the additional index ranges over $j = 1,\ldots,n_i$, with n_i the number of patients in trial i. Most of the multitrial paradigms are based on models of the form:

$$S_{ij} = \mu_{Si} + \alpha \times Z_{ij} + \varepsilon_{Sij}$$

Equation 3

$$T_{ij} = \mu_{Ti} + \beta_i \times Z_{ij} + \varepsilon_{Tij}$$

Equation 4

This is a so-called hierarchical model [19]. The treatment effects (β_i,α_i) have the same interpretation as in the above 'RE regression', but the two-level structure, with patients nested within trials, is now properly reflected. Note that the two treatment effects appear in a different equation. It is therefore not immediately clear how these models can be used to predict β_i from α_i. Technically, this is done by assuming a joint distribution for the quadruplet $(\mu_{Ti},\mu_{Si},\beta_i,\alpha_i)$. In this case, a four-variate normal distribution is a convenient choice. From this, the strength of the association between the treatment effects can be quantified, as well as the predictive ability for β_i from α_i and μ_{Si}. The use of the intercept can, in principle, (slightly) improve the predictive ability but this is a minor point. The corresponding measure takes the form of a conventional R^2; we denote it by $R^2(\beta/\alpha)$ and term it the 'trial-level surrogacy' [16]. Likewise, the association between the surrogate and true end point at the individual level is captured by a squared correlation, $R^2(\varepsilon_{Tij}/\varepsilon_{Sij})$, termed 'individual-level surrogacy'. It is reminiscent of the adjusted association ρ, in the sense that ρ is the correlation between both end points at the individual level, adjusting for treatment but not for the hierarchical structure in the data (i.e., trial), whereas $R^2(\varepsilon_{Tij}/\varepsilon_{Sij})$ adjusts for the latter as well. Put differently, if there were no heterogeneity between trials, then $R^2(\varepsilon_{Tij}/\varepsilon_{Sij})$ would equal ρ^2.

The meta-analytic framework: extensions & issues

While the meta-analytic framework is appealing, there are unavoidably a few drawbacks. First, the R^2 measures are very much tied to the (multivariate) normal framework. However, in clinical trial practice, it is very common to record binary outcomes, time-to-event responses, and so forth. While extensions have been considered in such settings, in a sequence of papers and reviewed in [3], the ensuing 'catalogue' of measures can be seen as a complication. For example, when both outcomes are binary, an odds ratio can be used, for time-to-event data, Kendall's τ has been used, and so forth. The silver lining is that this disparate collection of measures is needed for the individual-level surrogacy only. At the trial level, arguably the most important of both in clinical trial practice, R^2 measures can keep being used.

Second, matters complicate further if at least one of the outcomes, and sometimes both, are measured longitudinally. From a clinical standpoint, it may be wise to consider longitudinal surrogates. For example, take PSA in a prostate cancer study. It has been documented that PSA may be a poor surrogate with only a cross-sectional consideration, whereas its evolution over time may be more indicative. For example, if PSA remains relatively stable or increases linearly, it may be an indication of benign prostatic hyperplasia at most; super-linear increases (quadratic, exponential) may be an indication of locoregional or metastatic cancer [19]. Another example is when the brief psychiatric rating scale is studied as a surrogate end point for clinician's global impression in a trial in schizophrenia, with the understanding that both are measured longitudinally. These examples underscore the need to properly deal with longitudinal surrogates for time-to-event end points in the first case and a longitudinal surrogate for a longitudinal true end point on the other. Evidently, many more settings can be considered where at least one of the end points and sometimes both are longitudinal. There are several questions that can be asked from a clinical standpoint, such as whether one should use the entire longitudinal sequence, a few selected measures, the earlier portion, and so forth. Intuitively, the more measurements that are used, the better the prediction will become, but also the more time and financial resources will be needed. Quite a bit of work has been devoted to longitudinal outcomes, including appropriate extensions of the R^2 measures mentioned above [3,20,21].

Third, even in the case of two continuous end points, fitting the models and deriving the R^2 measures is challenging from a computational standpoint. Ideally, one

would like to make use of full linear mixed model technology [19], but this proves to be challenging in many practical situations. Reasons include, but are not limited to, the fact that only a small number of trials are typically available, that the trials vary widely in sample sizes, and so forth. A number of alternative computational strategies have been proposed [22], the most promising a so-called two-stage approach, where the model is fitted to each trial separately in the first stage, and then combined into meta-analytic summaries in the second stage.

In view of these concerns, a number of steps have been taken to unify, streamline and make the existing approaches computationally simpler.

Unification: an information–theoretic approach

A unifying step was made by introducing so-called information–theoretic concepts into surrogate marker validation, building upon earlier work [20,21,23]. In its simplest form for two continuous end points, and omitting some technical details, it comes down to 'comparing' the model fits of the following pair:

$$T_{ij} = \mu_{Ti} + \beta_i \times Z_{ij} + \varepsilon_{Tij}$$

Equation 5

$$T_{ij} = \mu_{Ti} + \gamma_i \times Z_{ij} + \delta \times S_{ij} + \varepsilon_{Tij2}$$

Equation 6

Note that this pair brings us back to models underpinning two out of the four Prentice criteria. Only now, one neither performs the hypothesis-testing characteristic for the criteria, nor is the PE calculated from it. Rather, in this case, the likelihood ratio for comparing both models is used as a basis to derive an information theoretic measure of surrogacy. There are several advantages to this approach. First, because a pair of separate models is fitted, rather than a joint, bivariate model, the computations are simpler. Second, still related to computation, one can fit this pair of models for each trial separately, and then combine the results, via appropriate linear combinations, to span the entire meta-analysis. Third, the resulting information–theoretic validation measure, denoted R_h^2 in [23], has the interpretation of a squared correlation coefficient, whatever the type of the end points, regardless of whether S and T are of the same or rather of a different type. Fourth, R_h^2 reduces to measures derived from a meta-analytic framework perspective for a pair of continuous end points, whether they are measured once or rather longitudinally [24,25].

Prediction: the surrogate-threshold effect

The authors have seen so far that the original definition was morphed over time into a predictive paradigm, via the intermediate stages of hypothesis testing and the estimation of surrogacy measures. Of course, even when one could accurately predict the treatment effect on the true end point from analyzing data from a trial that only records the surrogate, the question remains as to whether this treatment effect would be significant had it been observed directly. In conventional hypothesis testing, uncertainty is taken into account from the fact that a finite sample is taken from a larger, possibly infinite population. However in this case, two other sources of variability come into play. The second source is the uncertainty resulting from the fact that the evaluation of the surrogate is also based on a finite collection of data, from a finite set of trials. The third and last source is the prediction uncertainty; that is, the information that is lost because the surrogate rather than the true end point is observed. Burzykowski and Buyse studied this problem in detail and proposed the so-called surrogate-threshold effect (STE), defined as the minimal effect needed on the surrogate end point to ensure a significant treatment effect on the prediction for the effect on the true end point [26].

Not surprisingly, it follows from carefully scrutinizing the STE that often very large sample sizes and very large effects on the surrogate would be needed to offer a sufficiently promising perspective of a significant effect on the true end point. This implies that we need to reflect very carefully on the use of a surrogate. The authors return to this in the 'Future perspective' section. The STE is finding its way in current-day practice [27].

Alternative paradigms

The meta-analytic framework to evaluate surrogate end points is now well established, especially in its information theory format. The resulting measures are quite intuitive and the underlying computational requirements are not complicated either, whatever the outcome types, and whether or not the surrogate and/or true end point are measured once or repeatedly.

It is important to realize that it is not the only paradigm. A separate strand of research is of a causal inference nature [28,29]. Frangakis and Rubin approach surrogate-end point evaluation from a causal inference standpoint, using so-called principal stratification [28]. They use classical concepts from causal inference, often referred to as Rubin's causal model and to a large extent laid out by Holland [29]. It took two decades after the publication of Prentice's seminal paper until an attempt was made to review, classify, and study similarities and differences between the various paradigms [30]. Joffe and Greene essentially saw two important dimensions [30]. First, some methods are based on a single trial while others use several trials, namely meta-analysis. Second, some approaches are based on association, while others are based on causation. Because the meta-analytic framework described earlier is

based on association and uses multiple trials, on the one hand, and because the causal framework initially used a single trial [28], on the other, the above dimensions got convoluted and it appeared that correlation/meta-analysis had to be a pair, just like causal/single trial.

However, it is useful to disentangle the two dimensions and to keep in mind that proper evaluation of the relationship between the treatment effect on the surrogate and true end points is ideally based on meta-analysis. Joffe and Green argue that the meta-analytic approach is essentially causal in so far as the treatment effects observed in all trials are in fact average causal effects [30]. If a meta-analysis of several trials is not possible, then causal effects must be estimated for individual patients, which requires strong and unverifiable assumptions to be made. Recently, progress has been made regarding the relationship between the association and causal frameworks [31]. In a paper submitted for publication, Alonso *et al.* consider a quadruple [31]:

$$Y_{ij} = (T_{ij}[Z_{ij} = 0], T_{ij}[Z_{ij} = 1], S_{ij}[Z_{ij} = 0], [Z_{ij} = 1])$$

Equation 7

which is observable only if patient 'j' in trial 'i' would be assessed under both control and experimental treatment. Evidently, this is not possible and hence some of the outcomes in the quadruple are 'counterfactual'. Counterfactuals are essential to the causal-inference framework, while the above equation also carries a meta-analytic structure. Alonso van der Elst, Molenberghs, Buyse, and Burzykowsk [31] assume a multivariate normal for Y_{ij}, in order to be able to derive insightful expressions. It is clear that both paradigms root their validation approach in causal

Executive summary

Terminology & concepts

- Surrogate end points used in lieu of true end points in clinical trials need careful definition before progress can be made in their evaluation.

The preprentice era & the surrogacy as a response to fast-paced events

- It is argued that surrogate end points are such a natural concept that they amply appeared in the literature and in practice, before they were even formally defined and before an evaluation paradigm was set up.

The class of surrogates

- The evaluation is not context free. This means that a surrogate may have passed the evaluation test in a specific situation, namely within a certain class of treatment, on a certain population, and so forth. Extrapolation will always be needed to some extent if new drugs are investigated. The practical consequence is that one needs to reflect carefully on whether the leap inherent in using a surrogate is biologicaly, clinicaly and statisticaly warranted. Statistical surrogate-marker evaluation can contribute information to this endeavor, but not replace biological and/or clinical judgment.

Prentice's definition & the single-trial framework

- The definition and every single-trial assessment attempt are surrounded with problems. It is necessary to replicate treatment effects on true and surrogate end points for a satisfactory, data-based assessment. Such is possible only in a meta-analytic setting. Furthermore, sufficiently promising surrogates need to be available. This was a problem in the early years, but currently, thanks in particular to the genetics and 'omics' revolutions, the problem is increasingly to pick the most promising candidates for surrogacy from a large number of candidates.

The meta-analytic framework

- The meta-analytic framework allows to overcome the issues stated in the previous paragraph, provided that sufficient data are available. Also, it comes at a computational cost and leads to an eclectic collection of surrogacy measures. This is why an information–theoretic unification is beneficial.

Prediction: the surrogate-threshold effect

- The surrogate-threshold effect research indicates that using the surrogate on the one hand, and then making precise (significant) statements about the true end point on the other, may be asking too much. It implies that we need to think about how a surrogate could be used. There are several options:
  - A surrogate can be used for a temporary or conditional approval. After such a step, the true end point can still be observed, for further confirmation or, should it be needed, for reversal of the decision.
  - A surrogate, when properly validated from current and historic studies, can be used as the true end point of tomorrow's study. In other words, the surrogate would be considered as a clinically relevant end point in its own right, if it has a sufficiently strong connection with the previously used end point.

Alternative paradigms

- Apart from the meta-analytic framework, there are alternative methods available, predominantly based on causal inference. While these two families have been viewed as diametrically opposed, ongoing work shows that there is a closer-than-anticipated connection between both.

effects of treatment. However, there is an important difference. While the causal inference line of thinking places emphasis on individual causal effects, in a meta-analytic approach the focus is on the expected causal treatment effect. These authors show that, under broad circumstances, when a surrogate is considered acceptable from a meta-analytic perspective at both the trial and individual level, then it would be good as well from a causal-inference angle. These authors also carefully show, in line with comments made earlier, that a surrogate, valid from a single-trial framework perspective using individual causal effects, may not pass the test from a meta-analytic view-point when heterogeneity from one trial to another is large and the causal association is low. Evidently, more work is needed, especially for end points of a different type, but at the same time it is comforting that, when based on multiple trials, the frameworks appear to show a good amount of agreement.

Future perspective

In summary, the evaluation and use of surrogate markers brings new opportunities, but there are a lot of inherent caveats and problems. These will lead to disappointment if one looks upon surrogate marker evaluation as a decision process that will lead, in an automated fashion, to the use or rejection of a surrogate. A more modest and at the same time realistic goal, is to view surrogate marker evaluation as a statistical, quantitative component in the decision process that leads towards the adoption of a candidate surrogate in one of the ways alluded to before.

Financial & competing interests disclosure

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

References

Papers of special note have been highlighted as:

- ■ of interest

1 Prentice RL. Surrogate end points in clinical trials: definitions and operational criteria. *Stat. Med.* 8, 431–440 (1989).

- ■ **Seminal paper that generated the field of surrogate marker evaluation. Prentice's definition and criteria are set out here.**

2 Freedman LS, Graubard BI, Schatzkin A. Statistical validation of intermediate end points for chronic diseases. *Stat. Med.* 11, 167–178 (1992).

- ■ **Proposed the proportion explained. In so doing, the estimation paradigm was brought to surrogate marker evaluation.**

3 Burzykowski T, Molenberghs G, Buyse M. *The Evaluation of Surrogate Endpoints.* Springer, NY, USA (2005).

- ■ **Provides an overview of surrogate marker evaluation from the perspective of the editors and their collaborators, as well as from a host of experts in the field, including Prentice. Apart from methodological development, the focus is also on a variety of areas, such as oncology and HIV.**

4 Biomarkers Definitions Working Group. Biomarkers and surrogate end points: preferred definitions and conceptual framework. *Clin. Pharmacol. Ther.* 69, 89–95 (2001).

5 Ellenberg SS, Hamilton JM. Surrogate end points in clinical trials: cancer. *Stat. Med.* 8, 405–413 (1989).

6 Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? *Ann. Int. Med.* 125, 606–613 (1996).

7 The 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.* 321, 406–412 (1989).

8 Hughes MD. The evaluation of surrogate end points in practice: experience in HIV. In: *The Evaluation of Surrogate Endpoints.* Burzykowski T, Molenberghs G, Buyse M (Eds). Sprinter, NY, USA, 295–322 (2005).

9 Diggle PJ, Liang K-Y, Zeger SL. *Analysis of Longitudinal Data.* Clarendon Press, Oxford, UK (1994).

10 Brookmeyer R, Gail MH. AIDS *Epidemiology: a quantitative approach.* Oxford University Press, NY, USA (1994).

11 Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. Statistical validation of surrogate end points: problems and proposals. *Drug Infor. J.* 34, 447–454 (2000).

12 Albert JM, Ioannidis JPA, Reichelderfer P *et al.* Statistical issues for HIV surrogate end points: point and counterpoint. *Stat. Med.* 17, 2435–2462 (1998).

- ■ **These authors argue for the use of several trials rather than a single one. In so doing, the way is paved for a meta-analytic approach.**

13 Molenberghs G, Buyse M, Geys H, Renard D, Burzykowski T, Alonso A. Statistical challenges in the evaluation of surrogate end points in randomized trials. *Control. Clin. Trials* 23, 607–625 (2002).

14 Molenberghs G, Burzykowski T, Alonso A, Buyse M. A perspective on surrogate end points in controlled clinical trials. *Stat. Meth. Med. Res.* 13, 177–206 (2004).

15 Buyse M, Molenberghs G. The validation of surrogate end points in randomized experiments. *Biometrics* 54, 1014–1029 (1998).

- ■ **Proposed the relative effect and the adjusted association.**

16 Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The validation of surrogate end points in meta-analysis of randomized experiments. *Biostatistics* 1, 49–67 (2000).

- ■ **Introduces the meta-analytic framework to surrogate marker evaluation.**

17 Daniels MJ, Hughes MD. Meta-analysis for the evaluation of potential surrogate markers. *Stat. Med.* 16, 1515–1527 (1998).

- ■ **Introduces the meta-analytic framework to surrogate marker evaluation. These authors employ a Bayesian paradigm, but use summaries rather than individual data from the various trials.**

18 Gail MH, Pfeiffer R, van Houwelingen HC, Carroll RJ. On meta-analytic assessment of surrogate outcomes. *Biostatistics* 1, 231–246 (2000).

- ■ **Introduces the meta-analytic framework to surrogate marker evaluation.**

19 Verbeke G, Molenberghs G. *Linear Mixed Models for Longitudinal Data*. Springer, NY, USA (2000).

20 Alonso A, Geys H, Molenberghs G, Kenward MG. Validation of surrogate markers in multiple randomized clinical trials with repeated measures. *Biometrical J.* 45, 931–945 (2003).

21 Alonso A, Geys H, Molenberghs G, Kenward MG. Validation of surrogate markers in multiple randomized clinical trials with repeated measurements: canonical correlation approach. *Biometrics* 60, 845–853 (2004).

22 Tibaldi FS, Cortiñas Abrahantes J, Molenberghs G *et al.* Simplified hierarchical linear models for the evaluation of surrogate end points. *J. Statist. Comput. Simul.* 73, 643–658 (2003).

23 Alonso A, Molenberghs G. Surrogate marker evaluation from an information theory perspective. *Biometrics* 63, 180–186 (2007).

■ **In this paper, the information theory approach to surrogate marker evaluation is introduced.**

24 Assam P, Tilahun A, Alonso A, Molenberghs G. Information theory based surrogate marker evaluation from several randomized clinical trials with continuous true and binary surrogate end points. *Clin. Trials* 4, 587–597 (2007).

25 Alonso A, Molenberghs G. Evaluating time to cancer recurrence as a surrogate marker for survival from an information theory perspective. *Stat. Meth. Med. Res.* 17, 497–504 (2008).

26 Burzykowski T, Buyse M. Surrogate threshold effect: an alternative measure for meta-analytic surrogate end point validation. *Pharma. Stat.* 5, 173–186 (2006).

27 Mauguen A, Pignon JP, Burdett S *et al.* On behalf of the Surrogate Lung Project Collaborative Group. Surrogate end points for overall survival in chemotherapy and radiotherapy trials in operable and locally advanced lung cancer: a re-analysis of meta-analyses of individual patients' data. *Lancet Oncol.* 14, 619–626 (2013).

28 Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics* 58, 21–29 (2002).

■ **This paper is instrumental in proposing a causal inference approach, based on principal stratification, towards surrogate end point evaluation.**

29 Holland PW. Statistics and causal inference. *J. Am. Statist. Assoc.* 81, 945–960 (1986).

30 Joffe MM, Greene T. Related causal frameworks for surrogate outcomes. *Biometrics* 65, 530–538 (2009).

■ **Describes, classifies and compares the various paradigms available towards surrogate end point evaluation.**

31 Alonso A, van der Elst W, Molenberghs G, Buyse M, Burzykowski T. On the relationship between the causal-inference and meta-analytic paradigms for the validation of surrogate end points. *Tech. Rep.* (2013).

■ **Formally compares the meta-analytic and causal-inference based paradigms. The authors show strong mathematical connections and bring to the forefront assumptions on which each of the frameworks rest.**