

Surrogate outcomes in Clinical Trials

Graham Byrnes

Centre for Genetic Epidemiology

University of Melbourne

The Idea:

- Testing drugs against HIV in early 90's.
- AIDS arises through loss of immune system components, particularly CD4+ cells;
- If a drug reduces loss of CD4+, then it should improve survival;
- If there is no difference in CD4+, then survival will be the same;

- Causal pathway is

$$\text{Drug} \implies \text{CD4+} \implies \text{Survival}$$

- So testing for change in CD4+ associated with the treatment group should be **equivalent** to looking for an association between treatment and survival.
- *Equivalent* in the sense of asymptotically unbiased with regard to outcome;

- Actual relationship is

$$\text{Drug} \xRightarrow{\text{noise}} \text{CD4+} \xRightarrow{\text{noise}} \text{Survival}$$

- So if we use survival as an endpoint we have

$$\text{Drug} \xRightarrow{2 \times \text{noise}} \text{Survival}$$

compared to

$$\text{Drug} \xRightarrow{\text{noise}} \text{CD4+}$$

- So we get more power (and a quicker result).

These type of considerations motivated Prentice to give the following criterion for a surrogate endpoint:

“A response variable for which a test of the null hypothesis of no relationship to the treatment groups under consideration is also a valid test of the corresponding null hypothesis based on the true endpoint” .

T = the true endpoint;

S = the surrogate endpoint;

Z = the treatment group.

Prentice's criterion can then be re-phrased as

$$F(S|Z) = F(S) (\iff F(T|Z) = F(T)). \quad (1)$$

However this needs to be operationalised, which Prentice did with the following four criteria. These are meant to be taken as establishing the validity of a trial based on a surrogate endpoint.

2. $F(S|Z) \neq F(S)$;

3. $F(T|Z) \neq F(T)$;

4. $F(T|S) \neq F(T)$;

5. $F(T|S, Z) = F(T|S)$.

Note that demonstrating 3 would make the surrogate redundant!
Buyse and Molenberghs (1998) have shown that 4 and 5 are equivalent to (1) for binary outcomes.

That still leaves us with a big problem: we can't statistically demonstrate equivalence as required in 5.

We *could* do an equivalence trial

$$H_0 : D(F(T|S, Z), F(T|S)) > \delta,$$

or equivalently do a test with high power given an alternate hypothesis of an effect size $\geq \delta$.

Either would require very large numbers.

We may not really care about the truth of 5. In reality, any intervention will have *some* direct effect on T .

What is important is:

- how large is the direct effect?
- is it in the same, or opposite, direction to the effect mediated by S ?

Freedman's Proportion Explained

Freedman had the idea of converting the troublesome criterion 5 into an estimation problem:

P_E = the proportion of the treatment effect explained by the surrogate.

A more precise definition requires assuming that T , S and Z are related by GLM's.

Suppose that for appropriate link functions g_S , g_T and g_P ,

$$g_S(E(S|Z)) = \mu_S + \alpha Z;$$

$$g_T(E(T|Z)) = \mu_T + \beta Z;$$

$$g_P(E(T|S, Z)) = \mu + \beta_S Z + \gamma S.$$

Freedman then defined

$$P_E = 1 - \frac{\beta_S}{\beta}.$$

Some problems with P_E

1. Flandre and Yacine complain that P_E tends to have a very wide CI;
2. What if there is an interaction?

$$g_P(E(T|S, Z)) = \mu + \beta_S Z + \gamma S + \delta SZ;$$

3. P_E is not really a proportion and can lie outside $[0, 1]$.

In fact Flandre and Yacine go so far as to say that the high variability of P_E indicates that

“the measure simply should not be used in practice”.

In fact, estimating P_E simply exposes the basic silliness of the criterion $F(T|S, Z) = F(T|S)$, which is equivalent to requiring $P_E = 0$. If it's impractical to estimate P_E with precision, then it's impractical to verify the criterion.

With this lack of power in mind, the concern with interactions raised by Buyse *et al* (2000) seems pointless. The fact that P_E may not lie in $[0, 1]$ just indicates that a better name might have been chosen.

There is a much bigger problem: the above approaches require that any given surrogate must be validated jointly with the clinical outcome **and the treatment in question**.

However, if it's necessary to do a trial involving both the treatment Z and the clinical outcome T , what is the use of having a surrogate?

Is it the right question?

The problem is that without a way of experimentally modifying S , it's difficult to establish that a changes in S *causes* a change in T .

However, if we intervene to change S , there is the possibility that the intervention is directly acting on T , with S coming along for the ride.

I would suggest this is not a statistical problem, but a medical one.

A seductive example

Consider patients who have recently suffered an ischaemic stroke.

- The clinical outcome is the Rankin score (very noisy);
- The surrogate is the volume of infarct as measured by MRI.
- It seems intuitively reasonable that infarct volume is causally related to Rankin score (data supports a linear association);
- Is it reasonable to use infarct size as a surrogate marker for the efficacy of thrombolytic drugs in improving stroke outcomes?

This example illustrates a further advantage of surrogates.

- Patients arrive with very different initial infarct volumes;
- Typical drug action is to reduce the growth of the infarct by restoring blood supply;
- Using MRI, it's possible to measure the *change* of infarct volume. This would be expected to be a much more sensitive indicator of biological activity, since we can factor out a large amount of noise (initial infarct volume).

With the advantages, there are new questions:

- The *change* in infarct volume is the most sensitive measure of drug activity;
- *Final volume* is the best predictor of final Rankin score.
- Using the Prentice philosophy, final volume would be the most appropriate surrogate.

The clinical outcome is necessarily an estimate of marginal effect, because there is no reliable way of measuring patient cognitive function in the acute stage.

However, whenever it is possible to measure a conditional effect, clinical trials will do so to gain power.

Another advantage:

The effect of a drug on infarct volume is best seen on the log scale: in fact the simplest reasonable model seems to be something like

$$\log \left(\frac{V_f}{(V_i)^{2/3}} \right) = \alpha + \beta Z + \epsilon_1.$$

However the relation between final infarct volume and Rankin score is roughly linear,

$$R = \beta V_f + \epsilon_2,$$

implying

$$R = \exp(\alpha + \beta Z + \epsilon_1) \times V_i^{2/3} + \epsilon_2.$$

However the trial would almost certainly be analysed by imposing a model

$$R = \gamma + \delta Z + \epsilon_3,$$

with ϵ_3 assumed to be normal.

So using a surrogate allows us to gain power in two ways:

1. we can estimate a conditional effect rather than a marginal effect;
2. we avoid a miss-specified model.

In other words, the surrogate allows us to conduct a fundamentally different type of trial, which we would conventionally consider better.

However this makes it fundamentally impossible to validate the surrogate under any test of *equivalence*.