JAMA Guide to Statistics and Methods

When Can Intermediate Outcomes Be Used as Surrogate Outcomes?

David L. DeMets, PhD; Bruce M. Psaty, MD, PhD; Thomas R. Fleming, PhD

Randomized clinical trials have a long history of success in many medical arenas. Many trials that change clinical practice use clinical outcomes that are direct measures of how a patient feels, functions, or survives. The substantial resources required by trials using such end points are powerful incentive to pursue designs that reduce the numbers of patients required, the length of follow-up, and the trial costs.[1]

What Are Intermediate Outcomes and Why Are They Used?

Patient-centered outcomes—direct measures of how a patient feels, functions, or survives—often reflect the effects of multiple factors, reducing the expected treatment effect and thus increasing the required trial size. To reduce the trial resources, a frequent approach has been to use a biomarker or another replacement end point that is an intermediate outcome thought to capture the causal pathway through which the disease process affects the patient-centered outcomes. Intermediate outcomes, which may be physiological measures, laboratory test results, imaging results, or other such measures, are appealing because trials that use these outcomes are shorter, smaller, and statistically more powerful than those that evaluate patient-centered outcomes.

### What Are the Limitations of Intermediate End Points?

Despite the appeal of using replacement end points, 2 fundamental requirements must be met[2] to ensure that the replacement end point is a "valid surrogate," ie, the effect of the intervention on the replacement end point reliably predicts its effect on the patient-centered outcome. The first requirement is that the replacement end point and the patient-centered outcome are strongly correlated. The second is that effects of an intervention on the replacement end point should fully capture its net effect on the patient-centered outcome. There are several reasons this second requirement is often not met; thus, "a correlate does not a surrogate make."[3]

How Have Intermediate Outcomes Been Used?

An article published in *JAMA Oncology*[4] by Ritchie and colleagues illustrates these issues in the immuno-oncology setting, where checkpoint inhibitors have frequently been evaluated for the treatment of advanced solid cancers. The objective response rate (ORR), determined from decreases in tumor size, is often thought to be a valid surrogate for overall survival simply because, on a patient-specific level, treatment responders live longer than non-responders. The article evaluated the primary end points in 24 randomized controlled phase 2 trials of checkpoint inhibitors. Across these trials of checkpoint inhibitors, the correlation between the observed ORR odds ratio and the overall survival hazard ratio was modest (0.57 [95% CI, 0.23-0.89]). Ritchie et al[4] suggest avoiding ORR as a primary end point in phase 2 trials of checkpoint inhibitors.

Similarly, the findings reported by Ritchie et al[4] suggest that the effects of checkpoint inhibitors on progression-free survival (PFS),

Figure. Effects of an Intervention on a Replacement End Point
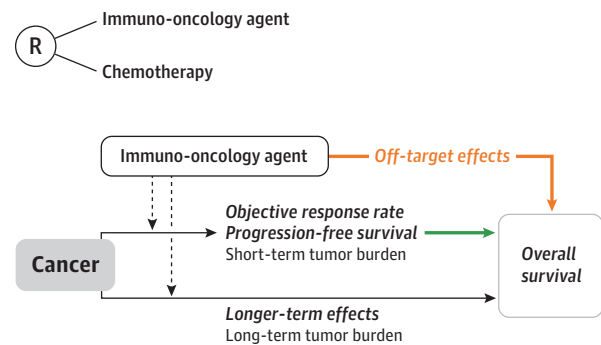
Effects on a replacement end point (R), such as objective response rate (ORR) or progression-free survival (PFS), often do not reliably predict intervention effects on a direct measure of how a patient feels, functions, or survives, such as overall survival, even though the 2 end points are correlated; insights into why that occurs are provided by recognizing the importance of multiple causal pathways of the disease process and multiple mechanisms of action of the treatment intervention.

another replacement end point for overall survival, were only modestly correlated with their effects on overall survival (0.42 [95% CI, 0.04-0.81]). It appears that PFS is insensitive to the longer-term effects of checkpoint inhibitors on overall survival.

How Should Intermediate Outcomes Be Used for Checkpoint Inhibitor Studies?

To understand why intermediate outcomes such as ORR and PFS may be correlates for patient-centered end points such as overall survival and yet might not be valid surrogates, consider the case of a randomized trial that compares an immuno-oncology agent, such as a checkpoint inhibitor, with chemotherapy. The **Figure** shows the multiple disease-process causal pathways and treatment-intervention mechanisms of action that can influence whether a replacement end point is a valid surrogate. First, the replacement end point might not lie in a pathophysiological pathway through which the disease process causally induces effects on the patient-centered end point (ie, the green arrow does not exist). Second, even if the replacement end point is in the pathway, there could be treatment effects on other causal pathways such as influencing tumor burden over longer periods that are inadequately captured by the replacement end point. This mismatch is likely when using short-term outcomes such as ORR and PFS as replacement end points in trials of checkpoint inhibitors that often have important longer-term effects. Third, even if the intervention has the intended mechanism of action (dashed black arrows), its effects on the patient-centered end point might be affected by other mechanisms (orange arrow) that are not captured by the replacement end point.

Even if intermediate end points such as ORR and PFS are not established as valid surrogate end points, they can be useful as supportive end points in phase 3 trials or as primary end points in proof-of-concept trials insofar as they provide substantive evidence about biological effects. Hence, insights about mechanisms of action of interventions are important in formulating those end points. Importantly, Ritchie et al[4] recognized that checkpoint inhibitors and chemotherapy have fundamentally different mechanisms of action. Because these therapies influence tumor biology differently, what is known about the reliability of ORR and PFS as surrogate end points for overall survival in the chemotherapy setting cannot be assumed to hold in the setting of checkpoint inhibitors.

Because it is important to have sensitivity to the longer-term effects of checkpoint inhibitors on tumor burden, duration of response (DOR) is at least as important as ORR. A preferred biomarker might be each patient's "time in response." This end point includes all patients in the analysis, in which nonresponders are included with an outcome of "zero" duration of response. This approach enables an intention-to-treat analysis with increased sensitivity. For example, a doubling in ORR and a doubling in DOR translates to a 4-fold increase in "time in response." Sensitivity could be further improved by using "time in disease control" or "long-term average change in disease burden" biomarkers that would capture causal effects on both disease stability and durability of tumor shrinkage. While such measures would still be relatively insensitive to unintended effects of interventions, this approach would reduce the risk of false-negative conclusions compared with traditional biomarkers such as ORR and PFS.

For proper validation of potential surrogate end points, there must be an in-depth understanding of the multiple causal pathways of the disease process and of the intended as well as unintended mechanisms of action of the treatment intervention.

Because such insights are inherently imperfect, there is also a need for meta-analyses of completed trials to assess the relationship of the net effects of interventions on the potential surrogate end points and on the patient-centered outcomes. Informative illustrations of that process are provided by the validation of "death or cancer recurrence" in the adjuvant colon setting for 5-fluorouracil–based regimens.[5] Another illustration is the validation of systolic and diastolic blood pressure as a surrogate for patient-centered outcomes for antihypertensive drug trials.[6] Intermediate outcome validation is, however, specific for drugs or classes of drugs because the validity of a surrogate might not properly extrapolate across different drug classes (or even to another drug in the same class), especially if important unintended effects are drug- or class-specific. Caution is also required before extrapolating surrogates from adults to children.[7]

If assessments of efficacy are based on replacement end points that are not properly validated as surrogates for direct measures of how a patient feels, functions, or survives, patients may be exposed to interventions that have unfavorable benefit-to-risk profiles. A classic example is the use of class IC antiarrhythmic agents to suppress arrhythmias after myocardial infarction. While beneficial effects on the arrhythmia intermediate outcome led to off-label use of these agents in hundreds of thousands of patients per year, the placebo-controlled Cardiac Arrhythmia Suppression Trial[8] revealed that these drugs tripled the death rate. There are many other examples in which biomarkers, when used as replacement end points, have yielded misleading results about efficacy.[6] While use of replacement end points provides more rapid assessments of experimental interventions, timeliness should not be achieved at the expense of a misleading risk-benefit profile. Registrational or pivotal trials should use direct measures of how a patient feels, functions, or survives whenever replacement end points have not been properly validated.

REFERENCES

1. Wittes J, Lakatos E, Probstfield J. Surrogate endpoints in clinical trials: cardiovascular diseases. *Stat Med*. 1989;8(4):415-425. doi:10.1002/sim.4780080405

2. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. *Stat Med*. 1989; 8(4):431-440. doi:10.1002/sim.4780080407

3. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? *Ann Intern Med*. 1996;125(7):605-613. doi:10.7326/0003-4819-125-7-199610010-00011

4. Ritchie G, Gasper H, Man J, et al. Defining the most appropriate primary end point in phase 2 trials of immune checkpoint inhibitors for advanced solid cancers: a systematic review and meta-analysis. *JAMA Oncol*. 2018;4(4):522-528. doi:10.1001/jamaoncol.2017.5236

5. Fleming TR. Surrogate endpoints and FDA's accelerated approval process. *Health Aff (Millwood)*. 2005;24(1):67-78. doi:10.1377/hlthaff.24.1.67

6. Fleming TR, Powers JH. Biomarkers and surrogate endpoints in clinical trials. *Stat Med*. 2012;31(25):2973-2984. doi:10.1002/sim.5403

7. Barst RJ, Ivy DD, Gaitan G, et al. A randomized, double-blind, placebo-controlled, dose-ranging study of oral sildenafil citrate in treatment-naive children with pulmonary arterial hypertension. *Circulation*. 2012;125(2):324-334. doi:10.1161/CIRCULATIONAHA.110.016667

8. 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(6):406-412. doi:10.1056/NEJM198908103210629