OXFORD

JNCI J Natl Cancer Inst (2015) 107(8): djv153

doi:10.1093/jnci/djv153
First published online June 24, 2015
Brief Communication

BRIEF COMMUNICATION

Sensitivity, Specificity, PPV, and NPV for Predictive Biomarkers

Richard Simon

Affiliation of author: Richard Simon, DSc, National Cancer Institute, Rockville MD.

Correspondence to: Richard Simon, DSc, National Cancer Institute, 9609 Medical Center Drive, Room 5W110, Rockville MD 20850 (e-mail: rsimon@nih.gov).

Abstract

Molecularly targeted cancer drugs are often developed with companion diagnostics that attempt to identify which patients will have better outcome on the new drug than the control regimen. Such predictive biomarkers are playing an increasingly important role in precision oncology. For diagnostic tests, sensitivity, specificity, positive predictive value, and negative predictive are usually used as performance measures. This paper discusses these indices for predictive biomarkers, provides methods for their calculation with survival or response endpoints, and describes assumptions involved in their use.

Diagnostic tests are often evaluated using the performance indices sensitivity, specificity, positive predictive value, and negative predictive value (1). Today, many oncology drugs are approved for marketing based on extending patient survival relative to a control treatment and are developed with predictive biomarkers used as companion diagnostic tests that attempt to identify the patients who benefit from the drug. The difficulty of computing such measures for predictive markers is discussed here and an approach is proposed, which, if interpreted carefully with awareness of the assumptions, may provide insight into the performance of predictive biomarkers.

Consider a randomized clinical trial with an experimental treatment T, a control C, and a time to event endpoint such as survival or disease-free survival. Data from observational studies are not suitable for the calculation of the measures considered here. Assume that there is a binary biomarker B that is positive or negative and that the hazard ratio (HR) of C vs T is Δ_+ for the biomarker-positive patients and Δ_- for the biomarker-negative patients. These hazard ratios, the P values for testing whether they are statistically significantly different from 1 (indicating no treatment effect), and a P value for testing whether they are statistically significantly different from each other (test of interaction) are the evidence usually provided to indicate the clinical relevance of the marker for that treatment. Regulators, payers, and medical effectiveness evaluators have, however, expressed interest in more direct measures of performance of predictive tests.

The positive predictive value (PPV) of the test can be defined as the probability that the survival of a marker-positive patient will be longer if the patient receives treatment T than if the patient receives control treatment C. Because the outcome for any patient can generally only be observed on a single treatment, patients who have longer survival on the test treatment vs control cannot be identified. Sitlani and Heagerty (2) developed such performance measures for short-term endpoints in which patients were observed on both treatments in cross-over studies. Huang et al. (4) developed such measures for binary endpoints under the strong assumption that if treatment benefited any patients, then it did not harm any others.

If the potential outcomes of the same patient under the two treatments are assumed independent, or conditionally independent given the covariates included in the models used to estimate the hazard ratios, then the performance measures can be computed. These performance measures will be subscripted with an "i" to indicate this assumption. One can avoid concern about the assumptions by thinking of PPV_i as the probability that a patient on treatment T has longer survival than another patient with identical covariates who is on the control C.

If the survival distributions have proportional hazards within biomarker strata and the potential survival time of a patient receiving T can be viewed as independent from the potential survival time of that same patient receiving C, then it is shown in the Supplementary Methods (available online) that

Received: October 17, 2014; **Revised:** March 25, 2015; **Accepted:** May 8, 2015

Published by Oxford University Press 2015. This work is written by (a) US Government employee(s) and is in the public domain in the US.

Downloaded from http://jnci.oxfordjournals.org/ at National Institutes of Health Library on July 7, 2015

$$ppv_i = \frac{\Delta_+}{1+\Delta_+} \quad (1)$$

where Δ_+ is the hazard ratio of C vs T (>1) for biomarker-positive patients. The negative predictive value (NPV) is the probability that a biomarker-negative patient will not have longer survival on T rather than C. It is also shown in the Supplementary Methods (available online) that

$$npv_i = \frac{1}{1+\Delta_-} \quad (2)$$

For example, Amado et al. (3) reported hazard ratios for progression-free survival of best supportive care vs panitumumab in second- or later-line therapy of patients with metastatic colorectal cancer. For patients with wild-type KRAS, the hazard ratio was 2.22, favoring panitumumab with a 95% confidence interval (CI) of 1.69 to 2.94. The hazard ratio for the patients with mutated KRAS was 1.01 with a 95% CI of 0.73 to 1.37. For this data, calling wild-type KRAS marker positive, the PPV_i and NPV_i as calculated from (1) and (2) are 0.69 and 0.50, respectively. We note that expression (2) indicates that an NPV of 0.5 results when there is no treatment difference in the marker-negative stratum.

The sensitivity is the probability that the biomarker is positive for patients who benefit from T relative to C. Specificity is the probability that the biomarker is negative for patients who do not benefit from T relative to C. The usual relationships between sensitivity, specificity, PPV, NPV, and prevalence can be written

$$sensitivity_i = 1 / \left\{1 + \frac{1-npv_i}{ppv_i}\frac{\Pr[B-]}{\Pr[B+]}\right\} \quad (3)$$

$$specificity_i = 1 / \left\{1 + \frac{1-ppv_i}{npv_i}\frac{\Pr[B+]}{\Pr[B-]}\right\} \quad (4)$$

where Pr[B+] denotes the prevalence of biomarker-positive patients and Pr[B-]=1-Pr[B+]. For the Amado et al. (3) data, the prevalence of wild-type KRAS was 0.62 and so the sensitivity and specificity as calculated from (3) and (4) are 0.62 and 0.50, respectively. Predicting treatment outcome is generally more difficult than distinguishing diagnostic categories; consequently, the very high performance indices commonly observed for diagnostic markers should not be expected for predictive biomarkers.

Approximate 95% confidence intervals for the four performance measures can be computed as described in the Supplementary Methods (available online) For the Amado et al. data, approximate 95% confidence intervals for PPV_i and NPV_i are (.63 to .75) and (.42 to .57) respectively. The 95% confidence intervals for sensitivity and specificity are (.58 to .71) and (.47 to .62), respectively.

The robustness of the results (1) through (4) to the assumption of independence was explored in two ways (see the Supplementary Methods, available online). For the bivariate exponential (5), although the potential survival times of a patient on T and C are not independent, equations (1) through (4) are correct. Bivariate Weibull survival distributions were also evaluated.

Although the results appeared reasonably robust to the departures from independence considered, the independence assumption can be weakened by estimating the hazard ratio of treatment using proportional hazards models that include prognostic covariates. Separate models for the two biomarker strata should be used coding the control as 1 and the test treatment as 0. If the potential survival times on T and C are considered independent conditionally on the prognostic factors included in the model, then the estimates given by equations (1) through (4) are valid (see the Supplementary Methods, available online). If proportional hazards models containing prognostic covariates are used to estimate Δ_+ and Δ_-, then the performance indices can be interpreted pragmatically rather than in terms of potential outcomes for the same patient, avoiding unverifiable independence assumptions. For example, PPV_i is interpreted as the probability that outcome for a patient on the test treatment will be better than outcome of a randomly selected other patient with the same covariates who receives the control.

For interpreting the performance indices in terms of potential outcomes of the same patient, the assumption of conditional independence can be further weakened by including a "frailty" term in the proportional hazards models. The frailty represents the effect of unobserved prognostic factors (see the Supplementary Methods, available online). Assuming that the frailty model is correct, the derivation of (A1) of the Supplementary Materials (available online) is valid and one can use the exponentiated regression coefficient for treatment obtained from the frailty model to compute the performance measures (1) through (4). The proportional hazard frailty model is easily fit using the coxme package in the R statistical programming system.

For the Amado et al. (3) data, the reported hazard ratios of 2.22 and 1.01 in the KRAS WT and mutated strata were computed based on separate proportional hazard models, including the treatment indicator and the stratification variables ECOG performance score and region of the institution. Age is an additional prognostic factor, and if it is included in those models the hazard ratios for treatment become 2.30 and 1.02, respectively. If a random frailty term is also included in the models, the hazard ratios for treatment in the KRAS WT and mutated strata become 2.95 (95% CI = 2.07 to 4.18) and 1.09 (95% CI = 0.746 to 1.59), respectively. Using these estimates, the performance indices become: PPV_i = 0.75, NPV_i = 0.48, $sensitivity_i$ = 0.70, and $specificity_i$ = 0.54, rather than the original PPV_i = 0.69, NPV_i = 0.50, $sensitivity_i$ = 0.62, and $specificity_i$ = 0.50.

The diagnostic performance measures PPV, NPV, sensitivity, and specificity cannot be directly measured for binary predictive biomarkers in most clinical trials. The measures computed using the formulas (1) through (4) can be interpreted in one of two ways. First, if they are derived from fitting proportional hazards models with prognostic covariates, they can be interpreted pragmatically without unverifiable assumptions; eg, PPV_i can be interpreted as the probability that a marker-positive individual receiving treatment T will have longer survival than that for a randomly chosen individual with the same covariates and marker value who receives C. Second, to be interpreted in terms of the potential outcomes of the same patient, I recommend models that include prognostic covariates and a random frailty in order to weaken the unverifiable assumption of independence of the potential outcomes on the two treatments.

Notes

I acknowledge the valuable contributions of the Associate Editor and the reviewers to improving an earlier version of this manuscript. I would also like to thank the cancer statistical group at Amgen for performing and making available the additional calculations on data from their clinical trial (3) to enable those results to be used for illustrating the methods described here.

BRIEF COMMUNICATION

Downloaded from http://jnci.oxfordjournals.org/ at National Institutes of Health Library on July 7, 2015

References

1. Pepe MS. *The statistical evaluation of medical tests for classification and prediction*. Oxford, UK: Oxford University Press, 2003.
2. Sitlani CM, Haegerty PJ. Analyzing longitudinal data to characterize the accuracy of markers used to select treatment. *Stat in Med*. 2014;33:2881–2896.
3. Amado RG, Wolf M, Peeters M, et al. Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. *J Clin Oncol*. 2008;26(10):1626–1634.
4. Huang Y, Gilbert PB, Janes H. Assessing treatment-selection markers using a potential outcomes framework. *Biometrics*. 2012;68:687–696.
5. Marshall AW, Olkin I. A generalized bivariate exponential distribution. *J Appl Probab*. 1967;4:291–302.

Downloaded from http://jnci.oxfordjournals.org/ at National Institutes of Health Library on July 7, 2015