Invited Commentary

Invited Commentary: The Importance of Prevalence in the Effectiveness of a (Bio)marker

Arpita Ghosh and Philip E. Castle*

* Correspondence to Dr. Philip E. Castle, ASCP Institute, 1225 New York Avenue NW, Suite 250, Washington, DC 20005 (e-mail: philip.castle@ascp.org).

Initially submitted August 24, 2010; accepted for publication November 17, 2010.

Metrics such as relative hazards and relative risks do not account for the prevalence of a marker over time and its relation to whether and when an outcome occurs. Uncommon markers that have good predictive values and common markers that are poorly predictive may not be (clinically) useful in predicting disease and other health outcomes. Recent work by Little et al. (Am J Epidemiol. 2011;173(12):1380–1387) highlights the development of a new method that considers both factors in predicting outcomes. Measures that incorporate both marker prevalence and predictive values and therefore are measures of “effectiveness” may be broadly helpful in deciding which markers or exposures are useful in disease screening or should be targeted by health interventions.

In this issue of the Journal, Little et al. (1) present criteria for comparing the effectiveness of time-dependent binary markers in predicting time to an event for all or some parts of a population. As they elegantly illustrate, relative hazards do not account for the prevalence of a marker over time and its relation to the time at which the outcome occurs.

In their example, D60 (the first segment or interval between menstrual bleeds of at least 60 days), D90 (the first segment or interval between menstrual bleeds of at least 90 days), and RR10 (the running range, defined as a difference of more than 42 days between the minimum and maximum segment lengths for 10 consecutive segments) were available as markers of a dichotomous outcome indicating menopausal status, that is, completion of the final menstrual period (FMP). The patterns of relative hazards by age for these 3 binary markers of menopause are very similar. However, by incorporating the age-specific prevalence of the marker as well as its age-specific discriminatory ability, as assessed at least crudely by the hazard ratio, into a measure of “marker effectiveness,” Little et al. (1) were able to choose the best marker for prediction of FMP at a given age better than they would have been able to by relying on the hazard ratio alone.

Because prevalence differs greatly by age, so too does the marker’s effectiveness. Logically, the prevalence of D60 peaks at an earlier age than D90 (the late 40s vs. the early 50s), and therefore, for women under 50 years of age, D60 is much more effective at predicting FMP than is D90, even though the hazard ratio for D90 is somewhat higher than that for D60.

The methodological work presented here adds to the growing body of research on biomarker development. Over the last decade, several groups of investigators have proposed/developed and evaluated methods of assessing the usefulness of a biomarker in ways that incorporate both the strength of association and the prevalence of the marker (i.e., the odds ratio and the prevalence of exposure, respectively, in the familiar 2 × 2 table of “exposure” status crossed with disease status). Previous statistical methods have relied on comparing data in the row margins or column margins in the standard 2 × 2 table. A recent study of genetic markers and family history as predictors of future prostate cancer (Figure 1) (2) illustrates the importance of the distribution of exposure. This example illustrates the generality of the issues we discuss here: Genotypes at 5 genetic markers plus family history can be seen as equivalent to a biomarker. The cost of improved
discrimination between cases and controls, however, is a paucity of subjects (1.4%) with sufficient numbers of these markers to be considered at elevated risk. To approach clinically useful risk stratification, one needs to account for marker prevalence, as described by Pepe et al. (3).

The method presented by Little et al. (1) could be extended or generalized in several useful ways. First, the method could incorporate nonmonotonic markers as well as monotonic (e.g., D60 for FMP) markers. That is, the effectiveness of a marker may depend not only on when it is detected but also on time elapsed since detection. Case in point: The prevalent detection of carcinogenic human papillomavirus (HPV) DNA, a necessary cause of cervical cancer, is strongly linked to the development of cervical precancer and cancer (4). However, the risk of cervical precancer or cancer from HPV infection is not very high until the infection has persisted (failed to clear) for months or years (5, 6). Fluctuating or intermittent patterns of HPV positivity probably carry very little risk of cervical disease (unpublished observations). It seems likely that in many scenarios, the persistence of a biomarker related to health risk follows a similar pattern: As time elapses from detection without clearance or becoming negative, the hazard of the health outcome increases.

Second, as illustrated in the example of the use of D60 to predict FMP, a single marker may not by itself distinguish well between persons who will experience an event sooner and those who will experience it later. The average time to FMP after a positive D60 at age 45 years is only a few years less than it is after a negative D60 at age 45 years. Thus, it may be useful to compare different combinations of markers for their effectiveness in predicting the timing of an event. For example, using D60 or RR10 to increase the sensitivity for detecting any change in the segments might further differentiate persons who will have an early FMP from those who will not, as compared with either marker alone. In addition, it might prove useful to extend this method for markers with multiple categories with varying hazards. As an example of the latter, using HPV to illustrate, the hazard depends greatly on which of the HPV genotypes are detected, with HPV16 representing the most risky genotype (4, 7).

Finally, this marker effectiveness measure might be adapted for direct risk prediction for a rare outcome of interest, like the diagnosis of a tumor (as described above), rather than for timing of an event that will be experienced by nearly everyone, like menopause among women. Both the risk and the timing of the event are clinically important in determining when to intervene. However, as illustrated in Figure 1, a marker that is useful clinically must also predict a sizeable proportion of the outcome. That is, it is not sufficient to have good predictive value if the marker predicts only a small fraction of the clinically important outcomes. For example, although mutations in the breast cancer susceptibility genes (8) are strongly linked to risks of breast and ovarian cancer, there is little utility in screening for them in the general population because of their rarity (unless it becomes extremely cheap to do so).

In a very preliminary way, we now show how the idea proposed by Little et al. (1) applies to evaluation of markers for risk prediction of binary outcomes, as well as in the clinical context. To evaluate a marker for risk prediction, we usually focus on the receiver operating characteristic curve and measures derived from it or on predictive values. However, as Pepe et al. (9) have pointed out, risk distributions in the population and risk distributions that are conditional on disease status are the most informative for deciding on the clinical usefulness of the marker. We emphasize that the risk distribution corresponding to prediction of outcome \( Y \) on the basis of marker

---

**Figure 1.** Log odds ratio for prostate cancer (\( y_1 \)-axis; ■) and percentage of prostate cancer cases positive for each specific number of markers (\( y_2 \)-axis; ●) according to number of genetic risk factors and/or family history of prostate cancer (\( x \)-axis), Sweden, 2001–2003. Based on data adapted from the study by Zheng et al. (2). The odds ratio values are shown above the black bars (1 = referent).
$X$ depends on both 1) the risk model or the risk prediction function $P(Y = 1 | x)$ and 2) the marker distribution $P(x)$.

Recognition of the limitation of focusing only on the first component leads us to account for the biomarker distribution while designing a measure of marker utility. We can construct a measure of marker effectiveness by combining functions of the marker distribution and the risk prediction model. The measure of marker effectiveness proposed by Little et al. (1) can be thought of as the product of a function of the marker distribution, which Little et al. refer to as the prevalence factor, and a function of the risk model, which they refer to as the discriminatory ability of the marker.

The same 2-element measures of marker effectiveness are found in some commonly used measures in risk prediction. The population attributable risk (PAR) (10), a measure of the impact of an exposure on disease, can be seen as the product of a measure of prevalence and a measure of risk, specifically

$$\text{PAR} = \frac{I_c - I_0}{I_c} = \frac{p(I_1 - I_0)}{I_c} = p \left( \frac{I_1 - I_0}{I_c} \right),$$

where $p$ is the proportion exposed, $I_c$ is the crude or overall risk in the population, $I_i$ is the risk in the exposed, and $I_0$ is the risk in the unexposed. Another measure that better captures the effect of a prevention effort in the population is the attributable community risk (10), which can be expressed as the product of the proportion exposed (prevalence factor) and the risk difference (discriminatory ability); note that the attributable community risk is equal to $I_c - I_0 = \text{PAR} \times I_c$. In a recent publication, Park et al. (11) defined the effect size for a single nucleotide polymorphism marker as $2\beta^2 f(1 - f)$, where $f$ is the minor allele frequency. This definition also has the 2 components: discriminatory ability and prevalence factor.

The timing of life events, also interpreted as intermediate events between exposure and disease or death, has significant health implications. Therefore, identifying the best markers for predicting these events may directly or indirectly aid in predicting the risk of health outcomes. In the example discussed by Little et al. (1), women with markers associated with later FMP may in turn be at increased risk of “menopause-related cancers” of the breast, ovary, and endometrium (12).

Effectiveness measures that incorporate both marker prevalence and predictive values can be broadly helpful in deciding which markers are useful in screening or which exposures should be targeted by other health interventions. Effectiveness (absolute population benefit) rather than efficacy (relative benefit in the exposed compared with the unexposed) is critical to evaluation of any public health intervention. Simply put, considering an entire population eligible for an intervention, how many will benefit? And if the work by Little et al. (1) is extended, how soon? An intervention based on the presence of a common marker that is a weak predictor of disease can have a more substantial impact than an intervention based on the presence of a rare marker that is a strong predictor of disease (10). The converse can also be true, depending on the strengths of association and the prevalences. Continuing with the example of HPV, prophylactic vaccination against HPV appears to be equally efficacious at all ages, but the effectiveness (public health benefit or attributable community benefit) diminishes with increasing age, because the attack rate of new HPV infections in unvaccinated women decreases significantly with increasing age (13). Epidemiologists can learn from health economists, who have long recognized the need to use measures of effectiveness rather than efficacy in their decision analysis models. The continued development of measures of effectiveness will be broadly applicable to epidemiology, public health, and medicine.

**ACKNOWLEDGMENTS**

Author affiliations: Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland (Arpita Ghosh); and ASCP Institute, Washington, DC (Philip E. Castle).

Conflict of interest: none declared.

**REFERENCES**


