Clinical Chemistry 62:5
737–742 (2016)

Early-Phase Studies of Biomarkers: What Target Sensitivity and Specificity Values Might Confer Clinical Utility?

Margaret S. Pepe,[1*] Holly Janes,[2] Christopher I. Li,[3] Patrick M. Bossuyt,[4] Ziding Feng,[5] and Jørgen Hilden[6]

BACKGROUND: Many cancer biomarker research studies seek to develop markers that can accurately detect or predict future onset of disease. To design and evaluate these studies, one must specify the levels of accuracy sought. However, justified target levels are rarely available.

METHODS: We describe a way to calculate target levels of sensitivity and specificity for a biomarker intended to be applied in a defined clinical context. The calculation requires knowledge of the prevalence or incidence of cases in the clinical population and the ratio of benefit associated with the clinical consequences of a positive biomarker test in cases (true positive) to cost associated with a positive biomarker test in controls (false positive). Guidance is offered on soliciting the cost/benefit ratio. The calculations are based on the longstanding decision theory concept of providing a net benefit on average in the population, and they rely on some assumptions about uniformity of costs and benefits to those tested.

RESULTS: Calculations are illustrated with 3 applications: predicting colon cancer recurrence in stage 1 patients; predicting interval breast cancer (between mammography screenings); and screening for ovarian cancer.

CONCLUSIONS: It is feasible to specify target levels of biomarker performance that enable evaluation of the potential clinical impact of biomarkers in early-phase studies. Nevertheless, biomarkers meeting the criteria should still be tested rigorously in studies that measure the actual impact on patient outcomes of using the biomarker to make clinical decisions.

Biomarker research has evolved substantially over the past 2 decades. Consortia including the Early Detection Research Network *(1)* have been formed to provide support and coordination to the multidisciplinary team science involved. In addition, rigorous standards for the design of biomarker studies have been formulated *(2)*. A key notion is that the design should be built around the intended clinical use for that biomarker. For example, biological samples should be collected from the target population and in the relevant clinical context.

Another key notion is that the biomarker's performance should be measured against performance that could confer clinical utility in practice. However, defining such performance standards—i.e., values of sensitivity and specificity that could confer clinical utility—is often a woefully neglected component of research design. Yet without having target levels for sensitivity and specificity, it is difficult make much progress: for example, we cannot make conclusions about the success or failure of biomarkers evaluated in studies and neither can we calculate necessary sample sizes *(2, 3)* for study designs. In this article, we apply well-established concepts from decision theory to derive target performance values for biomarkers in early-phase research.

A typical phase II or III biomarker study *(4)* measures biomarkers in samples that were prospectively collected and stored for later use in research. Clinical usefulness cannot be fully determined from these retrospective studies. Subsequent phase IV/V studies in which biomarker tests are incorporated into clinical practice would be needed to really assess clinical utility vis-à-vis the impact on patient outcomes of medical decisions on the basis of the biomarker. Nevertheless, for the purposes of research in earlier phases, we must make some educated guesses about levels of sensitivity and specificity that are likely to confer clinical utility when biomarkers are studied in later phases.

This article presents methods for back-of-envelope calculations to assist investigators in setting target sensi-

[1] Biostatistics and Biomathematics Program, Public Health Sciences Division, [2] Vaccine and Infectious Disease Division, Public Health Sciences Division, and [3] Translational Research Program, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA; [4] Clinical Epidemiology, University of Amsterdam, Amsterdam, Netherlands; [5] Biostatistics, University of Texas MD Anderson Cancer Center, Houston, TX; [6] Department of Biostatistics, Institute of Medical Genetics, University of Copenhagen, Copenhagen, Denmark.

* Address correspondence to this author at: Biostatistics and Biomathematics, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, M2-B500, Seattle, WA 98105. Fax 206-667-7004; e-mail: mspepe@u.washington.edu.

Received November 18, 2015; accepted January 14, 2016.
Previously published online at DOI: 10.1373/clinchem.2015.252163

tivity and specificity values for early-phase biomarker research. The target performance values are important for determining from a completed study whether a marker performs at or above the target level and for designing a marker study to evaluate whether a marker meets this standard. Sample size calculations for biomarker discovery studies require such specification, as described recently *(3)*, and so do more traditional sample size calculations for biomarker validation studies *(2)*. We also describe the assumptions on which these calculations rely so that investigators can be cognizant of their limitations and remain tentative in making conclusions about the value of biomarkers until they are truly tested in late-phase clinical impact studies. The logic of our approach is based on defining an action threshold for medical procedures and treatments offered to those testing positive. This action threshold is derived from considering the potential benefits and harms of the medical procedures and treatment. This approach is grounded in the well-established decision analytic framework for medicine that dates back at least to the 1970s and has been brought to the fore again recently by proponents of decision curves *(5)* for evaluating biomarkers and risk prediction models.

Methods

THE CALCULATION

When an individual tests positive with a biomarker, the result is a true positive if that person has the target disease or will have the specified clinical outcome, and the result is a false positive otherwise. To simplify, we refer to the former as cases and the latter as controls. Within the target population, all noncases are controls.

We write TPR[7] for the true-positive rate or sensitivity (the proportion of cases testing positive) and FPR for the false-positive rate or 1 − specificity (the proportion of controls testing positive). Let ρ be the prevalence of cases in the relevant clinical population (if the biomarker is for screening or diagnostic purposes) or the incidence of cases over a specified time period (if the biomarker is for prognostic or risk prediction purposes). Use of the biomarker leads to a proportion $\rho \times TPR$ cases testing positive and $(1 - \rho) \times FPR$ controls testing positive.

Cases testing positive derive benefit on average from consequent workup or treatment, whereas controls suffer harm from unnecessary procedures or treatment. Although we do not need to explicitly quantify the net benefit B associated with a case that tests positive or the net harm or cost C associated with a control testing positive, we use symbols B and C to motivate an important formula. The total benefit derived from clinical use of the biomarker is derived from positive tests in cases and is quantified by $\rho \times TPR \times B$. Ignoring for now the costs or inconveniences associated with biomarker testing, the cost (nonmonetary as well as monetary) associated with clinical use of the biomarker is derived from positive tests in controls and is quantified by $(1 - \rho) \times FPR \times C$. Overall then, the biomarker is beneficial for use in the population if the benefits outweigh the costs:

$$\rho \times TPR \times B > (1 - \rho) \times FPR \times C \qquad (1)$$

In other words, the minimal requirement for the biomarker to be useful is

$$\frac{\text{TPR}}{\text{FPR}} > \frac{1 - \rho}{\rho} \times r \qquad (2)$$

where r is the cost/benefit ratio (C/B). The ratio TPR/FPR is called the positive diagnostic likelihood ratio *(6)*. Note that if one also takes into account the costs associated with biomarker testing, including nonmonetary costs of obtaining samples and performing assays, the minimal required TPR/FPR to achieve benefit in the population is even higher than $\frac{1 - \rho}{\rho} \times r$.

In some settings, investigators specify a target for sensitivity; for example, high sensitivity may be warranted if the disease is serious. Such a target for sensitivity is easily achieved by choosing an appropriate biomarker positivity criterion. In practice, however, investigators often neglect to specify a corresponding target for specificity or set an arbitrary target for specificity. The above formula shows that the corresponding target for specificity should depend on the prevalence ρ and the cost/benefit ratio r and provides an explicit expression for it. Moreover, the expression provides a set of sensitivities and corresponding specificities that indicate potential for clinical usefulness. In other words, it shows that one does not need to set a single value for target sensitivity. Analogously, in current practice, investigators sometimes fix specificity at a high level, such as with disease screening biomarkers, but set no target or an arbitrary target for corresponding sensitivity. The above formula now provides a rational approach to choosing a target value for sensitivity when specificity is fixed by investigators.

INTUITIVE MEASURES OF THE COST/BENEFIT RATIO

Specifying the costs and benefits associated with workup or treatment can be a very difficult exercise in practice. In many settings, the exact costs and benefits are unknown or unknowable, for example, if the intended workup or treatment is new or has not been defined. Yet, we see that one must make some educated guesses about r to calculate targets for biomarker performance. Educated guesses

[7] Nonstandard abbreviations: TPR, true-positive rate; FPR, false-positive rate; TVS, transvaginal ultrasound.

are appropriate in early-phase research. Here we describe some intuitive ways of articulating the cost/benefit ratio that may allow researchers to feel more comfortable with proposing possible values for it.

From a public health perspective, in the context of screening and diagnosis where the consequence of a positive biomarker test is workup, one can articulate r in terms of the maximum number of controls (N_{max}) we are willing to work up to reap the benefits associated with working up a single case. To see this, observe that the cost associated with working up N_{max} controls is $N_{max} \times C$ and, according to our definition of N_{max}, this is just equal to the benefit associated with working up 1 case, namely B. In other words, $N_{max} \times C = B$, implying that the cost benefit ratio $r = 1/N_{max}$. Note that in the context of screening, B is the expected benefit of working up a case, which is not necessarily the same as the benefit of detecting a case unless the workup is perfect at detecting disease.

As an illustration, consider a biomarker that will be used to select women for mammography. Mammography constitutes the workup here. Experience with mammography in women 50–70 years old indicates that for every case of breast cancer that is screened with mammography, 240 women without breast cancer are screened. Therefore, according to the public health system, the cost/benefit ratio associated with screening mammography must be no more than 1/240: because the health system routinely offers mammography screening to women 50–70 years old, it must consider that screening 240 women without cancer is worth the benefit of screening 1 breast cancer case. Thus N_{max} is ≥ 240, and consequently r is $\leq 1/240$. In summary, we write the following for screening/diagnosis:

$$\frac{1}{r} = \text{maximum number of controls worth working up per case worked up} = N_{max}. \quad (3)$$

Observe that without explicitly stating what are the costs associated with unnecessary workups or what are the net benefits associated with working up a case, these entities are implicitly contained in specification of N_{max}. By using intuition about the potential costs and benefits of workup, researchers may be able to supply likely values or ranges of values for r in terms of N_{max}.

In the context of risk prediction, cancer prevention interventions might be recommended on the basis of a positive biomarker test. In the context of prognosis, therapeutic intervention might be recommended in those testing positive. Analogous considerations imply that the benefit/cost ratio, $1/r$, is the number of people worth unnecessarily subjecting to the downstream intervention to reap the benefits of intervention for 1 case. This is the number of controls testing positive (people treated because of the positive test but who would have had a good outcome in the absence of treatment) relative to the cases testing positive (persons treated who would have had a bad outcome in the absence of treatment). Here we write the following for risk prediction/prognosis:

$$\frac{1}{r} = \text{maximum number of interventions to controls worth intervention to 1 case.} \quad (4)$$

Another method for articulating the cost/benefit ratio is to consider the minimum level of risk at which workup or intervention is considered warranted, denoted by R. This is not a new concept, and it was well described by Pauker and Kassirer *(7)* in 1975. To illustrate, a woman might consider that she should have a mammogram if her chance of having breast cancer is at least 5/1000, but not if her risk is less. In essence, the cost/benefit ratio she implicitly associates with mammography is $0.005/(1 - 0.005)$. The mathematical argument for this equivalence is that at the threshold risk R, her expected benefit, $R \times B$, equals her expected cost, $(1 - R) \times C$, mathematically written as $R \times B = (1 - R) \times C$. This implies that $R/(1 - R) = C/B = r$.

If one can supply an action threshold for risk—that is, a value below which workup or therapeutic intervention would not be performed and above which workup or intervention would be performed—one is implicitly specifying the cost/benefit ratio r. Although specifying a value or range of possible values for the action threshold is not an easy exercise, we have found that researchers are often willing to take on that task much more easily than explicitly specifying costs and benefits separately. We write the following for screening/diagnosis:

$$r = \frac{(\text{threshold risk for workup})}{(1 - \text{threshold risk for workup})}. \quad (5)$$

In the context of risk prediction and prognosis, the technical result is analogous, and we write the following:

$$r = \frac{(\text{threshold risk for intervention})}{(1 - \text{threshold risk for intervention})}. \quad (6)$$

Results

We now sketch 3 clinical scenarios in which a biomarker might be sought and perform corresponding back-of-envelope calculations for target levels of sensitivity and specificity.

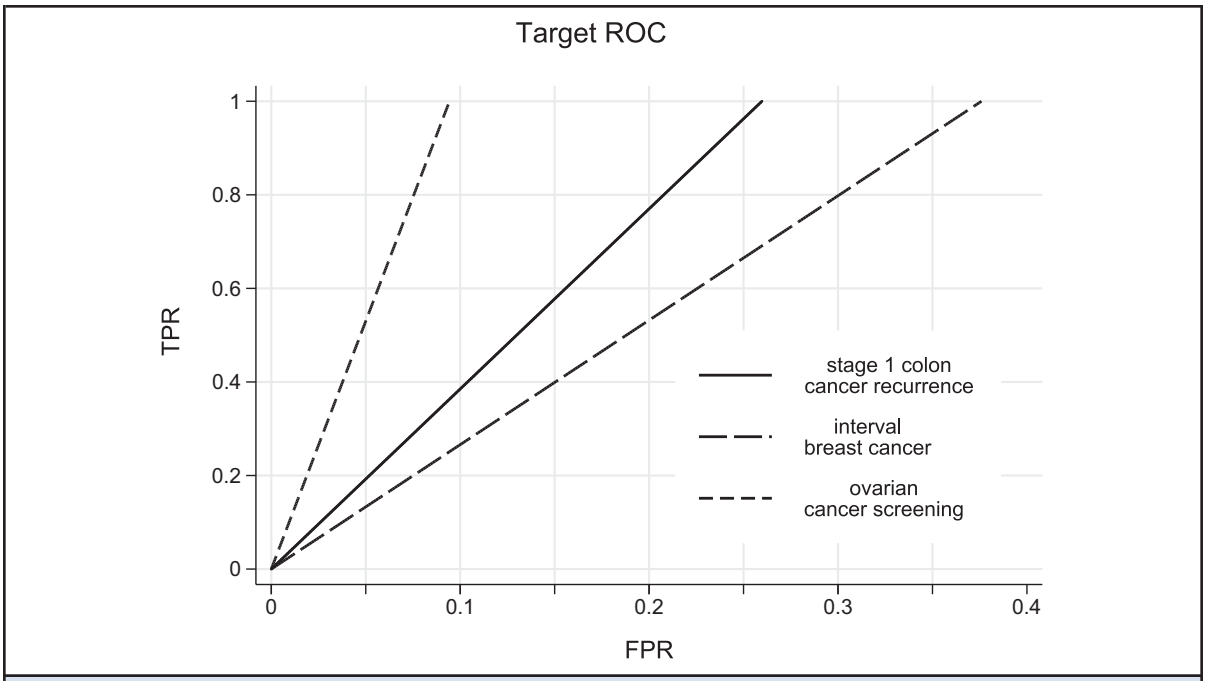

Fig. 1. Minimal requirements for performance of a biomarker test in 3 clinical settings.
A biomarker with (FPR, TPR) anywhere above the target ROC line meets criteria for clinical usefulness under assumptions described in the text.

BIOMARKERS FOR COLON CANCER RECURRENCE

Consider development of a biomarker for risk of recurrence within the first year after surgery for stage 1 colon cancer patients. Stage 1 patients are not routinely offered chemotherapy, and the goal is to use the biomarker to identify stage 1 patients at increased risk of recurrence who might be offered chemotherapy. The overall 1-year recurrence rate for stage 1 patients is 10% *(8)*, $\rho = 10\%$. How high would a stage 1 patient's risk need to be to warrant chemotherapy? Guidance may be derived from the fact that stage 3 colon cancer patients are routinely offered chemotherapy and that as a group their 1-year risk of recurrence is 30% *(9)*. Therefore the minimal risk threshold must be ≤30%. If investigators chose 30% as the risk threshold for recommending adjuvant chemotherapy, then $r = 0.3/(1 - 0.3) = 0.43$.

We now have all we need to calculate target values for biomarker sensitivity (TPR) and specificity (1 − FPR). Any combination of (FPR, TPR) satisfying

$$\frac{TPR}{FPR} \geq \left(\frac{(1-\rho)}{\rho}\right) \times r = \left(\frac{0.9}{0.1}\right) \times 0.43 = 3.85 \quad (7)$$

will do. For example, a biomarker with FPR = 0.10 needs corresponding TPR ≥ 0.385. A marker with FPR = 0.20 would require TPR > 0.77.

Fig. 1 shows minimally acceptable TPR values corresponding to various values of FPR on a target ROC line. Because a marker's TPR cannot exceed 1, markers with FPR >1/3.85 = 0.26 are not viable for this clinical application since they cannot meet the requirement that TPR/FPR > 3.85. For biomarkers measured on a continuous scale, we can set the FPR at higher or lower values by choosing less or more stringent criteria for biomarker positivity. Practical limitations on resources may further limit the acceptable range of FPR. A marker that has even a single (FPR, TPR) above the target ROC could have clinical utility.

BIOMARKERS FOR INTERVAL BREAST CANCER

It is recommended that women 50–74 years old have annual or biennial mammography to screen for breast cancer *(10, 11)*, but interval cancers occur between mammograms. Consider development of a biomarker to recommend additional mammograms at 8 and 16 months after a negative screening mammogram, with the goal of detecting cancers occurring during the 24 months after screening. During this interval, the expected incidence of cancer is 1.36–1.68 per 1000 women (say, ρ = 0.15%) *(12, 13)*. Suppose a guideline panel has evaluated the pros and cons of additional mammograms and suggested that the health care system should support 500 additional mammograms if it ensures that 1 woman with

interval cancer undergoes intermediate mammography: 250 women each undergoing 2. This implies $N_{max} = 250$ and so $r = 1/250$.

Biomarker performance targets can now be calculated as any satisfying

$$\frac{TPR}{FPR} \geq \left(\frac{(1 - 0.0015)}{0.00015}\right) \times \left(\frac{1}{250}\right) = 2.66. \quad (8)$$

For example, if the biomarker FPR is set at 0.05, allowing 5% of controls to undergo intermediate mammograms, the biomarker TPR for interval cancers must exceed $2.66 \times 0.05 = 0.13$. See minimally acceptable TPR values corresponding to alternative FPR values in Fig. 1.

BIOMARKERS FOR OVARIAN CANCER SCREENING

Ovarian cancer is typically diagnosed in advanced stages *(14)*. The annual incidence for 50 to 64-year-olds is about 25 in 100 000 *(15)*. Suppose we seek a biomarker for annual screening and after a positive biomarker result, transvaginal ultrasound (TVS) will be performed. If the TVS is also positive, the woman will be referred to surgery for definitive diagnosis *(16)*. Suppose the health care system requires that at a minimum, 1 ovarian cancer must be discovered for every 10 diagnostic ovarian surgeries performed, i.e., the risk threshold for surgical workup is 1/10. (Again, we do not dwell here on a formal reasoning for this choice, which would necessarily involve considerations of both economic and ethical values.) The criterion of finding 1 cancer for every 10 surgeries implies that the cost/benefit ratio for workup after the combined biomarker/TVS screen is 1/9. The TPR/FPR of the biomarker/TVS screen should therefore be at least $(1 - 0.00025)/[0.00025 \times (1/9)] = 444$:

Minimal performance of biomarker/TVS screen:

$$\frac{\text{TPR}}{\text{FPR}} \geq 444. \quad (9)$$

If the biomarker result and the TVS result are statistically independent (another assumption), the TPR for the combined test is the product of the TPR for TVS, which is 0.755 *(14)* and the TPR for the biomarker, written TPR_b. Correspondingly, and again assuming statistical independence, the FPR for the combined test is the FPR for TVS, 0.018 *(14)*, times the biomarker FPR, written FPR_b. The above criterion is then rewritten as $0.755 \times TPR_b/0.018 \times FPR_b > 444$, implying

$$\text{Minimal performance of biomarker: } \frac{\text{TPR}_b}{\text{FPR}_b} \geq 10.6 \quad (10)$$

If a biomarker detects 80% of ovarian cancers ($TPR_b = 0.80$), the criterion implies that for clinical usefulness it must have a corresponding $FPR_b \leq 0.075$. If its sensitivity is less—say, 50%—the corresponding specificity must be more, ≥0.953 (FPR = 0.047), to confer benefit. Even if the biomarker detects all ovarian cancer (TPR = 1.00), its false-positive rate must be <0.094 for viable clinical use (Fig. 1). Skates et al. *(16)* have argued for early detection markers that are highly specific. With specificity set at 98%, for instance, we see that a marker that is positive in ≥21% of ovarian cancers might be useful clinically.

Discussion

The formulas for performance targets presented here are very simple because they rely on some assumptions. Foremost among these assumptions are the following:

(*a*) the expected net benefit of workup (or intervention) is the same for all cases regardless of their biomarker values and

(*b*) the expected net cost of workup (or intervention) is the same for all controls regardless of their biomarker values.

One can easily envision scenarios in which these assumptions do not readily apply. Suppose age is prognostic, or that a prognostic biomarker is related to age. If younger cases are more likely to benefit from the intervention and suffer less toxicity, assumptions (*a*) and (*b*) do not hold. To partially address this issue, one could specify different cost/benefit ratios for younger and older patient groups, thereby arriving at different biomarker performance targets for the 2 groups.

Our calculations address only scenarios in which

(*c*) intended use for the biomarker is to select some individuals for workup or treatment who would otherwise, under standard of care, not undergo this workup or treatment.

For settings in which the default is that all individuals receive workup or treatment, a biomarker test might be sought to identify a subgroup of individuals to forgo this workup or treatment or receive a different regimen. Calculations pertaining to this setting are described in the Supplemental Materials, which accompany the online version of this article at http://www.clinchem.org/content/vol62/issue5.

Another assumption implicit in our calculations is the following:

(*d*) testing negative with the biomarker has the same effect on patient outcome as not having been tested at all with the biomarker.

A subtle issue relating to (*d*) is that reassurance provided by a negative biomarker test might lead individuals to change health care or lifestyle practices. If that is of concern, it should be factored into considerations about net costs and benefits associated with test results. Finally, we assume the following:

(*e*) harms and side effects associated with the biomarker test are negligible.

In addition to patient outcomes, we ignore costs, including nonmonetary costs such as discomfort, inconvenience, and stress associated with obtaining biological samples. In the online Supplemental Materials, we extend the formulas for performance targets to settings where assumptions (*c*)–(*e*) do not hold.

The assumptions underscore the potential utilities and limitations for the calculations we propose, namely to assist in setting targets for biomarker sensitivity and specificity in early-phase biomarker research. Similar assumptions are inherent in the use of decision curves for biomarker evaluation *(5)*, a methodology based on notions of benefit, which has become popular recently *(17)* in light of shortcomings noted with the area under the ROC curve and the net reclassification index *(18–20)*.

We do not suggest that clinical benefit can be truly evaluated in early-phase research, and we stress that one must be careful to interpret our calculations and those of decision curve analysis in terms of potential benefit rather than actual benefit. However, the advantage of this potential benefit framework is that one can design and evaluate discovery and validation studies within the context of sensitivity and specificity targets that align with the intended clinical application.

Author Contributions: *All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.*

Authors' Disclosures or Potential Conflicts of Interest: *Upon manuscript submission, all authors completed the author disclosure form. Disclosures and/or potential conflicts of interest:*

Employment or Leadership: None declared.
Consultant or Advisory Role: None declared.
Stock Ownership: None declared.
Honoraria: None declared.
Research Funding: Grants funded by the NIH were awarded to M.S. Pepe, R01 GM054438 and U24 CA086368; H. Janes, R01 CA152089; C.I. Li, R01 CA152637.
Expert Testimony: None declared.
Patents: None declared.

Role of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, or preparation or approval of manuscript.

References

1. Feng Z, Kagan J, Pepe M, Thornquist M, Rinaudo J, Dahlgren J, et al. Early detection research network specimen reference sets: paving the way for rapid evaluation of potential biomarkers. Clin Chem 2013;59:68–74.
2. Pepe MS, Feng Z, Janes H, Bossuyt PM, Potter JD. Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. J Natl Cancer Inst 2008;100:1432–8.
3. Pepe MS, Li CI, Feng Z. Improving the quality of biomarker discovery research: the right samples and enough of them. Cancer Epidemiol Biomarkers Prev 2015;24:944–50.
4. Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, et al. Phases of biomarker development for early detection of cancer. J Natl Cancer Inst 2001;93:1054–61.
5. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26:565–74.
6. Pepe MS. The statistical evaluation of medical tests for classification and prediction. Oxford: Oxford University Press; 2003.
7. Pauker SG, Kassirer JP. Therapeutic decision making: a cost-benefit analysis. N Engl J Med 1975;293:229–34.
8. American Cancer Society. Colorectal cancer. http://www.cancer.org/cancer/colonandrectumcancer/detailedguide/colorectal-cancer-survival-rates (Accessed August 2015).
9. Moertel CG, Fleming TR, Macdonald JS, Haller DG, Laurie JA, Tangen CM, et al. Fluorouracil plus levamisole as effective adjuvant therapy after resection of stage III colon carcinoma: a final report. Ann Intern Med 1995;122:321–6.
10. U.S. Preventive Services Task Force. Screening for breast cancer: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med 2009;151:715–26.
11. American Cancer Society. Breast Cancer. http://www.cancer.org/Cancer/BreastCancer/DetailedGuide/index (Accessed August 2015).
12. Bennett RL, Sellars SJ, Moss SM. Interval cancers in the NHS breast cancer screening programme in England, Wales and Northern Ireland. Br J Cancer 2011;104:571–7.
13. Bucchi L, Ravaioli A, Foca F, Colamartini A, Falcini F, Naldoni C. Incidence of interval breast cancers after 650,000 negative mammographies in 13 Italian health districts. J Med Screen 2008;15:30–5.
14. Menon U, Gentry-Maharaj A, Hallett R, et al. Sensitivity and specificity of multimodal and ultrasound screening for ovarian cancer, and stage distribution of detected cancers: results of the prevalence screen of the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS). Lancet Oncol 2009;10:327–40.
15. Howlader N, Noone AM, Krapcho M, Garshell J, Miller D, Altekruse SF, et al., editors. SEER Cancer Statistics Review, 1975–2011. Bethesda (MD): National Cancer Institute. http://seer.cancer.gov/archive/csr/1975_2011/ (Accessed March 2016).
16. Skates SJ, Gillette MA, LaBaer J, Carr SA, Anderson NL, Liebler DC, et al. Statistical design for biospecimen cohort size in proteomics-based biomarker discovery and verification studies. J Proteome Res 2013;12:5383–94.
17. Fitzgerald M, Saville BR, Lewis RJ. Decision curve analysis. JAMA 2015;313:409–10.
18. Pepe MS, Janes HE. Gauging the performance of SNPs, biomarkers and clinical factors for predicting risk of breast cancer. J Natl Cancer Inst 2008;100:978–9.
19. Pepe MS, Janes H, Li CI. Net risk reclassification p-values: valid or misleading? J Natl Cancer Inst 2014;107:355.
20. Vickers AJ, Pepe MS. Does the net reclassification improvement help us evaluate models and markers? Ann Intern Med 2014;60:136–7.