



Prediagnostic evaluation of multicancer detection tests: design and analysis considerations

Stuart G. Baker , ScD,^{1,*} Ruth Etzioni , PhD²

¹Division of Cancer Prevention, National Cancer Institute, Bethesda, MD, USA

²Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

*Correspondence to: Stuart G. Baker, ScD, Division of Cancer Prevention, National Cancer Institute, 9609 Medical Center Dr, Bethesda, MD 20892, USA (e-mail: sb16i@nih.gov).

Abstract

There is growing interest in multicancer detection tests, which identify molecular signals in the blood that indicate a potential preclinical cancer. A key stage in evaluating these tests is a prediagnostic performance study, in which investigators store specimens from asymptomatic individuals and later test stored specimens from patients with cancer and a random sample of controls to determine predictive performance. Performance metrics include rates of cancer-specific true-positive and false-positive findings and a cancer-specific positive predictive value, with the latter compared with a decision-analytic threshold. The sample size trade-off method, which trades imprecise targeting of the true-positive rate for precise targeting of a zero-false-positive rate can substantially reduce sample size while increasing the lower bound of the positive predictive value. For a 1-year follow-up, with ovarian cancer as the rarest cancer considered, the sample size trade-off method yields a sample size of 163 000 compared with a sample size of 720 000, based on standard calculations. These design and analysis recommendations should be considered in planning a specimen repository and in the prediagnostic evaluation of multicancer detection tests.

A multicancer detection test identifies cancer-specific signals in the blood from a variety of preclinical cancers. Technologies under development for multicancer detection testing include assays based on abnormal DNA methylation (1), circulating proteins and variations in cell-free DNA (2,3), fragmentation patterns of cell-free DNA (4), and protein markers in exosomes (5).

Our premise is that the multicancer detection test yields a continuous risk score, which is an estimated probability of developing cancer based on multiple predictors, including markers from the blood specimens (eg, thousands of DNA fragments) and possibly clinical risk factors. In logistic regression, the risk score comes from parameter estimates. In machine learning, the risk score arises from the algorithm, such as the terminal decision tree nodes in a random forest (6) or the last-layer sigmoid activation function in deep learning (7).

The investigator selects a cut point on this risk score such that a risk score larger than this cut point indicates a positive multicancer detection test result. If the multicancer detection test is positive and identifies a tissue of origin, the next step is a diagnostic workup to ascertain whether there is preclinical cancer at the tissue of origin. If the diagnostic workup indicates preclinical cancer, the patient receives appropriate early intervention.

There are 2 types of multicancer detection tests. A 1-stage multicancer detection test involves a separate test for each cancer type, such as with protein markers from exosomes (5). In this case, the tissue of origin is directly known. A 2-stage multicancer detection test first determines whether the test result is positive for any cancer, and then, if possible, assigns a tissue of origin (1-4).

The goal of cancer screening with multicancer detection tests is to detect cancers earlier to decrease cancer mortality rates.

The extent to which screening for cancer with multicancer detection tests might affect cancer mortality, however, remains to be determined. Also, as with any cancer screening test, a multicancer detection test has the potential for harm. One harm is a false-positive finding, which is a screening test that indicates preclinical cancer not confirmed by diagnostic workup (8). Another potential harm is overdiagnosis—that is, the detection of preclinical cancer on workup that would not have developed into symptomatic cancer in the absence of cancer screening.

Pipeline for evaluating multicancer detection tests

Because multicancer detection testing involves potential benefits and harms, it requires evaluation before clinical use. We envisioned multicancer detection test evaluation in 3 successive studies related to the cancer screening biomarker pipeline (9-12): a diagnostic performance study, a prediagnostic performance study, and a cancer screening randomized trial.

A diagnostic performance study evaluates how well the multicancer detection test classified specimens as samples from diagnosed cancer cases vs noncancer controls. The adjective *diagnostic* refers to cancer diagnosis at the time of the blood draw. Many people with diagnosed cancers in this study were likely diagnosed because of symptoms, which is relevant to the interpretation of results. Other names for a diagnostic performance study are phase II study (9) and preliminary performance study (10). Most studies evaluating multicancer detection tests have been diagnostic performance studies (13).

A prediagnostic performance study evaluates how well the multicancer detection test predicts cancer in asymptomatic individuals (9-12). The adjective *prediagnostic* refers to cancer diagnosis after the time of the blood draw. Other names for a prediagnostic performance study are phase III study (9) and retrospective performance study (10). In a prediagnostic performance study, investigators collect specimens from a blood draw; store specimens during the follow-up period (typically, 1 year); and, at the end of the follow-up period, test stored specimens from all cases and a random sample of controls.

In the prediagnostic performance study, *cases* refers individuals who received a blood draw and developed cancer during the follow-up period, and *controls* refers to individuals who received a blood draw and did not develop cancer during the follow-up period. Participants receiving a blood draw are considered a random sample from a target population who would receive a multicancer detection test, if offered. For each case and randomly selected control, the investigators perform a multicancer detection test on the stored specimen.

When evaluating multicancer detection tests based on a prediagnostic performance study, the determination of tissue of origin must be based only on the specimen. For example, a prediagnostic performance study would not be appropriate if tissue-of-origin determination required an imaging scan. The American Cancer Society Cancer Prevention Study-3 (14), which used 294 000 stored specimens, and the Taizhou Longitudinal Study (15), which used 5 000 stored serum specimens, exemplify multicancer detection prediagnostic performance studies.

A cancer screening randomized trial is a major leap from the previous studies in the pipeline. In a screening trial for a multicancer detection test, investigators randomly assign asymptomatic participants to the multicancer detection test or to no multicancer detection test with early intervention, if indicated. Conventional cancer screening occurs during the trial. The primary endpoint is cancer mortality. The National Cancer Institute is developing a pilot cancer screening trial for multicancer detection test evaluation (16).

Because it targets asymptomatic individuals, the prediagnostic performance study is more relevant than the diagnostic performance study for deciding whether to evaluate the combination of the multicancer detection test with early intervention in a cancer screening trial. For illustration, suppose that a multicancer detection test is based only on marker X, which arises after the diagnosis of a symptomatic cancer. This multicancer detection test may perform well in the diagnostic performance study, which involves symptomatic cancers. It would perform poorly, however, in a prediagnostic performance study, where participants are asymptomatic and therefore lack marker X. As a real-world example, a diagnostic performance study found that carcinoembryonic antigen almost perfectly classified participants into diagnosed colorectal cancer (CRC) cases or non-cancer controls (17), while a later prediagnostic performance study found that carcinoembryonic antigen was a poor predictor of the development of CRC in asymptomatic individuals (18). A multicancer detection test that poorly predicts cancer development in a prediagnostic performance study is a poor bet for reducing cancer mortality in a cancer screening trial. Moreover, excellent performance is needed in the prediagnostic performance study to recommend a cancer screening trial because of possible harms from early intervention in the cancer screening trial.

Prediagnostic performance studies

We focused on prediagnostic performance studies, which have received relatively little attention in multicancer detection test evaluation. There are 2 types of prediagnostic performance studies, depending on the data used to develop the prediction model: validation-only studies and development-validation studies.

A validation-only prediagnostic performance study, such as the American Cancer Society Cancer Prevention Study-3 (14), uses 1) the specimens from the diagnostic performance study as a training sample for model development and 2) the specimens from the prediagnostic performance study as a validation sample for performance evaluation. In other words, model development involves classifying specimens into a *current* diagnosis of cancer or no cancer, and performance evaluation involves classifying specimens into a *future* diagnosis of cancer during follow-up or no future diagnosis of cancer.

A development-validation prediagnostic performance study, such as the Taizhou Longitudinal Study (15), splits the set of prediagnostic specimens into a training sample for model development and a validation sample for performance evaluation. Ideally, the split would involve a nonrandom sample to investigate generalizability. Unlike the validation-only study, both model development and performance evaluation involve classifying specimens into a *future* diagnosis of cancer during follow-up or no future diagnosis of cancer. Thus, model development in a development-validation study is more likely to yield good validation performance than model development in a validation-only study. To increase the sample size for model development, it may be possible to use bootstrapping for performance evaluation (19). With a development-validation prediagnostic performance study, the diagnostic performance study can be viewed as proof of principle.

Metrics for prediagnostic performance

A prediagnostic performance study can evaluate multicancer detection tests for specific cancers so that investigators can refine the set of cancers the multicancer detection test is targeting. Evaluating the multicancer detection test for all cancers combined is less useful because overall performance is dominated by some cancers and is not relevant to treatment decisions, which involve specific cancers. Therefore, we consider study design considerations for specific cancers in a multicancer detection test. With some approximations, we evaluate multicancer detection test performance using the same metrics as would be used to evaluate individual cancer tests.

For computing sample size, we use the cancer-specific true-positive and false-positive rates. The true-positive rate (TPR) for cancer j is the fraction of cases of cancer j with a positive multicancer detection test for cancer j . The false-positive rate (FPR) for cancer j is the fraction of the randomly selected controls with a positive multicancer detection test for cancer j . Varying cut points of the risk score to compute pairs of false-positive rate and true-positive rate for cancer j yields a receiver operating characteristic curve for cancer j . As we will discuss, we suggest focusing the analysis on the point on the receiver operating characteristic curve with false-positive rate=0. One can always set false-positive rate=0 by making the cut point of the risk score for a positive multicancer detection test result larger than the largest risk score among the controls.

For interpreting results, we use the cancer-specific positive predictive value (PPV). The PPV for cancer j is the estimated

probability of diagnosing cancer j during follow-up among individuals positive for cancer j on the multicancer detection test. We compute the PPV for cancer j as $PPV_j = TPR_j \times p_j / \{TPR_j \times p_j + FPR_j \times (1 - P)\}$, where p_j is the estimated probability of diagnosing cancer j during follow-up and P is the estimated probability of diagnosing any cancer during follow-up. The PPV formula is exact for a 1-stage multicancer detection test and approximate for a 2-stage multicancer detection test. The exact formula for the 2-stage multicancer detection test includes another term in the denominator, $\sum_{i \neq j} FTOOR_{ij} \times p_i$, where $FTOOR_{ij}$ is the false tissue-of-origin rate for cancer j among individuals with cancer i . The false tissue-of-origin rate was small in a diagnostic performance study (1), so it is likely small in a prediagnostic performance study. Also, because $FTOOR_{ij}$ is multiplied by p_i , the contribution to the denominator is much smaller than the contribution from the term $FPR_j \times (1 - P)$.

An important question is how large a PPV is large enough to recommend moving to the next step of the multicancer detection test evaluation pipeline. To formally address this question based on clinical considerations, we define the PPV threshold for cancer j as the value of the PPV for cancer j that will yield a positive expected utility of prediction for cancer j . If the PPV for cancer j is larger than the PPV threshold for cancer j , we recommend further evaluation of the multicancer detection test for cancer j .

Ignoring the small contribution of the false tissue-of-origin rate in a 2-stage multicancer detection study, the expected that the utility of prediction for cancer j is $U_j = PPV_j \times q_j \times B_j - (1 - PPV_j) \times q_j \times C_j$, where B_j is the anticipated benefit of a true-positive prediction for cancer j , C_j is the anticipated cost of a false-positive prediction for cancer j , and q_j is the probability of a positive test for cancer j . Setting $U_j > 0$ implies a $PPV_j > PPV$ threshold for cancer $j = 1/(B_j / C_j + 1)$. The PPV threshold must be larger than P for a sensible result. One can interpret (B_j / C_j) as the number of false positives one would trade for a true positive. For example, for ovarian cancer screening, where surgery following a true-positive test can lead to major complications and there is no clearly established benefit from early intervention (20), some investigators have discussed B_j / C_j equal to 10 (21). We recommend a sensitivity analysis based on a plausible range of PPV thresholds. For example, for $B_j / C_j = 20, 10,$ and 5 , the PPV threshold for cancer j is $0.05, 0.09,$ and 0.17 , respectively.

Minimum follow-up time

In planning the sample size for a serum repository, a key consideration is the minimum follow-up time for prediagnostic evaluation. The choice of minimum follow-up time for the sample size calculation balances the following considerations:

- Because multicancer detection technology is developing rapidly, a shorter minimum follow-up time makes it more likely that the multicancer detection technology would be relevant at the end of the study.
- Multicancer detection tests likely perform worse with longer follow-up because 1) some cancers arise from preclinical cancers initiated after specimen collection, diluting the cancer cases with cases that could not have been prevented by screening, and 2) predictions farther in the future are less likely to be reliable. For example, in a study of cancer antigen 19 to predict pancreatic cancer, true-positive rates were $0.80, 0.60, 0.39,$ and 0.28 at a false-positive rate of 0.01 for $t = 0.25, 0.5,$ and $1, 2$ years after the blood draw (22). Less dramatically, for the prediction of cancer (CRC, esophageal, liver, lung, or stomach) in China, the Taizhou Longitudinal Study

yielded true-positive rates of $1.00, 0.904, 0.947,$ and 0.839 at a false-positive rate of 0.053 for 0 to $1, 1$ to $2, 2$ to $3,$ and 3 to 4 years after a blood draw (13).

- The shorter the minimum follow-up time, the more likely the multicancer detection test will identify rapidly developing cancers with short preclinical durations.
- Too short a time between the blood draw and cancer diagnosis may yield insufficient lead time for early intervention to be effective, which is an important consideration for further evaluation in a cancer screening trial.
- A shorter minimum follow-up time requires a larger sample size to obtain the required number of cases.

Although these considerations are difficult to balance because many aspects are unknowable, we suggest a minimum follow-up time of 1 year for the sample size calculation, so the sample size is likely feasible, investigators are likely to identify rapidly developing cancers, and the technology is not likely to be outdated at completion. Another reason to consider a 1-year minimum follow-up time is when the clinical application of multicancer detection screening would occur in 1-year intervals.

Sample size

The sample size for the prediagnostic performance study is the number of asymptomatic individuals from whom blood is drawn and specimens are collected. To illustrate the sample size calculation, we consider a validation-only design to evaluate a multicancer detection test involving 9 cancers (ovarian, stomach, pancreatic, liver, bladder, CRC, lung, prostate, breast), with a minimum follow-up time of 1 year. We compute the sample size for the rarest cancer under consideration—here, ovarian cancer—because that sample size will suffice for more commonly occurring cancers. The [Supplementary Material](#) (available online) provides the mathematical details of the sample size calculation. We compute the sample size based on 1) TPR_{TAR} , the target true-positive rate; 2) TPR_{LOW} , the lower bound of the 95% confidence interval for the estimated true-positive rate centered at TPR_{TAR} , (adjusted for chance results when investigating multiple cancers); 3) FPR_{TAR} , the target false-positive rate; and 4) FPR_{UPP} , the upper bound of the 95% confidence interval for the estimated false-positive rate centered at FPR_{TAR} . These target values yield a target value for the lower bound of the estimated PPV for cancer j , $PPV_{LOW(j)} = TPR_{LOW} \times p_j / \{TPR_{LOW} \times p_j + FPR_{UPP} \times (1 - P)\}$. Based on US population data for individuals age 50 to 64 years (23), we use $p_j = 12/100\,000$ for the 1-year incidence rate of ovarian cancer and $P = 528/100\,000$ for the 1-year incidence rate of the 9 cancers considered.

Scenario 0 specifies a standard set of target values for false- and true-positive rates (11) that many investigators would likely find reasonable—namely, $TPR_{TAR} = 0.80,$ $TPR_{LOW} = 0.70,$ $FPR_{TAR} = 0.01,$ and $FPR_{UPP} = 0.03$. As shown in [Table 1](#), a sample with 70 cases and 300 controls could satisfy these targets with at least 59 positives among the cases and at most 3 positives among the controls. Obtaining the required 70 ovarian cancer cases in 1 year with 95% probability requires 720 000 individuals for specimen collection. Also, $PPV_{LOW} = 0.003$, which is too small to provide useful information.

To substantially reduce sample size while increasing PPV_{LOW} , we applied the sample size trade-off method (24,25), which trades imprecise targeting of TPR_{TAR} for precise targeting of $FPR_{TAR} = 0$. Imprecise targeting of TPR_{TAR} means that TPR_{LOW} is much smaller than TPR_{TAR} , which decreases sample size because

Table 1. Sample size and positive predictive value lower bound based on ovarian cancer as the rarest cancer considered in a validation-only prediagnostic performance study with 1-year follow-up

Targeted quantity	Standard sample size trade-off targets		
	Scenario 0	Scenario 1	Scenario 2
TPR_{TAR}	0.80	0.50	0.80
TPR_{LOW}	0.70	0.20	0.50
FPR_{TAR}	0.01	0.0000	0.0000
FPR_{UPP}	0.03	0.0005	0.0005
$PPV_{LOW(j)}$ for ovarian cancer	0.003	0.05	0.11
No. of cases	70	12	12
No. of test-positive cases	59	7	10
No. of controls	300	7500	7500
No. of test-positive controls	3	0	0
Sample size	720 000	163 000	163 000

FPR = false-positive rate; PPV = positive predictive value; TPR = true-positive rate.

fewer cases are needed to obtain a wider confidence interval, and the sample size for specimen collection depends on the number of cases, not on the number of controls. Precise targeting of FPR_{TAR} means that FPR_{LOW} is close to FPR_{TAR} , which increases PPV_{LOW} because the PPV is largely determined by the false-positive rate when the disease is rare.

We consider 2 scenarios under the sample size trade-off method. Scenario 1 specifies $TPR_{TAR} = 0.50$, $TPR_{LOW} = 0.20$, $FPR_{TAR} = 0$, and $FPR_{UPP} = 0.005$. Scenario 2 specifies $TPR_{TAR} = 0.80$, $TPR_{LOW} = 0.50$, $FPR_{TAR} = 0$, and $FPR_{UPP} = 0.005$. As shown in Table 1, a sample with 12 ovarian cancer cases and 7500 controls could satisfy these targets with at least 7 positives among the cases for scenario 1, at least 10 positives among the cases for scenario 2, and 0 positives among the controls. Obtaining the required 12 ovarian cancer cases in 1 year with 95% probability requires only 163 000 individuals for specimen collection—an impressive 77% reduction in the sample size relative to scenario 0. For scenario 1, $PPV_{LOW(j)} = 0.05$, corresponding to $B_j / C_j \leq 19$, and for scenario 2, $PPV_{LOW(j)} = 0.11$, corresponding to $B_j / C_j \leq 8$. These target benefit-cost ratios are reasonable for multicancer detection tests for ovarian cancer, and, if achieved, would suggest moving to the next step for evaluation.

Discussion

In this commentary, we explore the feasibility and sample size requirements for prediagnostic performance studies in the setting of multicancer detection testing. The sample size calculation was based on a validation-only design with a minimum follow-up of 1 year. A good case can be made for a 6-month minimum follow-up to detect more rapidly developing cancers, which would double the sample size. Alternatively, one might halve the sample size by specifying a minimum follow-up of 2 years, anticipating 6 ovarian cancers in year 1 and 6 in year 2. A sample of only 6 ovarian cancers in year 1 is not informative, however, if the goal is to draw conclusions at year 1. To investigate performance at longer times before diagnosis, investigators can follow patients longer than the minimum follow-up time used in the sample size calculation.

The sample size trade-off method is based on the target precision for the estimated true-positive rate when the estimated false-positive rate equals 0, which corresponds to 1 threshold on the receiver operating characteristic curve. A possible concern

with the sample size trade-off method is that estimates of the true-positive rate are imprecise. We believe, however, that this imprecision in estimating true-positive rates is a small price to pay for a greatly reduced sample size. More importantly, PPV, which is more relevant to patients than is the true-positive rate, is estimated precisely because of the precise estimation of the false-positive rate, which is achieved by using many controls.

Another approach to reducing sample size in a study for the early detection of cancer is to focus only on individuals with a high risk of developing cancer. Because multicancer detection tests involve many cancers, the risk factors would have to apply across the cancers tested. The most predictive risk factor applicable to all cancers is age. Implementing the study for older ages would decrease sample size but with less generalizability (26). Because generalizability is important, we recommend applying the prediagnostic performance study design to asymptomatic individuals in an age group corresponding to that targeted by the multicancer detection test.

In summary, serum repositories are likely to be a valuable resource for prediagnostic evaluation of multicancer detection tests, but the field needs guidance regarding how they should be designed and sized. Focusing the analysis on the lower bound of the PPV yields useful results and allows for a feasible sample size. Focusing the sample size on a minimum 1-year follow-up allows for a timely evaluation and good performance for identifying rapidly developing cancers.

Data availability

The data used to compute the incidence of ovarian cancer in the sample size calculations are publicly available from Surveillance, Epidemiology, and End Results, as listed in the references.

Author contributions

Stuart G. Baker, ScD (Conceptualization; Formal analysis; Methodology; Writing—original draft), Ruth Etzioni, PhD (Writing—review & editing).

Funding

Financial support for this study for Ruth Etzioni was provided entirely by grant No. R35 CA274442 from the National Cancer Institute.

Conflicts of interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and publication of this article.

Acknowledgements

Opinions expressed by the authors are their own, and this material should not be interpreted as representing the official viewpoint of the US Department of Health and Human Services, the National Institutes of Health, the National Cancer Institute, or the Division of Cancer Prevention. The funder had no role in the design of the study; the collection, analysis, or interpretation of the data; or the writing of the manuscript and decision to submit it for publication.

References

- Liu MC, Oxnard GR, Klein EA, Swanton C, Seiden MV; CCGA Consortium. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Ann Oncol.* 2020;31(6):745-759.
- Cohen JD, Li L, Wang Y, et al. Detection and localization of surgically resectable cancers with a multianalyte blood test. *Science.* 2018;359(6378):926-930.
- Lennon AM, Buchanan AH, Kinde I, et al. Feasibility of blood testing combined with PET-CT to screen for cancer and guide intervention. *Science.* 2020;369(6499):eabb9601.
- Cristiano S, Leal A, Phallen J, et al. Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature.* 2019;570(7761):385-389.
- Li S, Yi M, Dong B, Tan X, Luo S, Wu K. The role of exosomes in liquid biopsy for cancer diagnosis and prognosis prediction. *Int J Cancer.* 2021;148(11):2640-2651.
- Malley JD, Kruppa J, Dasgupta A, Malley KG, Ziegler A. Probability machines: consistent probability estimation using nonparametric learning machines. *Methods Inf Med.* 2012;51(1):74-81.
- Zhan T. DL 101: Basic introduction to deep learning with its application in biomedical related fields. *Stat Med.* 2022;41(26):5365-5378.
- Baker SG, Kramer BS. Estimating the cumulative risk of a false positive under a regimen involving various types of cancer screening tests. *J Med Screen.* 2008;15(1):18-22.
- Pepe MS, Etzioni R, Feng ZD, et al. Phases of biomarker development for early detection of cancer. *J Natl Cancer Inst.* 2001;93(14):1054-1061.
- Baker SG, Kramer BS, McIntosh M, Patterson BH, Shyr Y, Skates S. Evaluating markers for the early detection of cancer: Overview of study designs and methods of analysis. *Clin Trials.* 2006;3(1):43-56.
- Baker SG. Improving the biomarker pipeline to develop and evaluate cancer screening tests. *J Natl Cancer Inst.* 2009;101(16):1116-1119.
- Baker SG, Kramer BS, Srivastava S. Markers for early detection of cancer: statistical issues for nested case-control studies. *BMC Med Res Methodol.* 2002;2:4.
- LeeVan E, Pinsky P. Predictive performance of cell-free nucleic acid-based multi-cancer early detection tests: a systematic review. *Clin Chem.* 2024;70(1):90-101.
- Patel A, Dur CAC, Alexander G, et al. Methylated DNA biomarkers and incident cancer in the American Cancer Society (ACS) Cancer Prevention Study-3 (CPS-3) cohort. *J Clin Oncol.* 2023;41(suppl 16):3004-3004.
- Chen X, Gole J, Gore A, et al. Non-invasive early detection of cancer four years before conventional diagnosis using a blood test. *Nat Commun.* 2020;11(1):3475.
- Minasian LM, Pinsky P, Katki HA, et al. Study design considerations for trials to evaluate multicancer early detection assays for clinical utility. *J Natl Cancer Inst.* 2023;115(3):250-257.
- Thomson DM, Krupcey J, Freedman SO, Gold P. The radioimmunoassay of circulating carcinoembryonic antigen of the human digestive system. *Proc Natl Acad Sci USA.* 1969;64(1):161-167.
- Thomas DS, Fourkala EO, Apostolidou S, et al. Evaluation of serum CEA, CYFRA21-1 and CA125 for the early detection of colorectal cancer using longitudinal preclinical samples. *Br J Cancer.* 2015;113(2):268-274.
- Collins GS, Dhiman P, Ma J, Schlüssel MM, et al. Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ.* 2024;384:e074819.
- Henderson JT, Webber EM, Sawaya GF. Screening for ovarian cancer: Updated evidence report and systematic review for the US Preventive Services Task Force. *J Am Med Assoc.* 2018;319(6):595-606.
- Skates SJ, Gillette MA, LaBaer J, et al. Statistical design for biospecimen cohort size in proteomics-based biomarker discovery and verification studies. *J Proteome Res.* 2013;12(12):5383-5394.
- Fahrman JF, Schmidt CM, Mao X, et al. Lead-time trajectory of Ca19-9 as an anchor marker for pancreatic cancer early detection. *Gastroenterology.* 2021;160(4):1373-1383.e6.
- Surveillance, Epidemiology, and End Results (SEER) Program. SEERStat Database: SEER Incidence Rates by Age at Diagnosis 2014-2018, National Cancer Institute, DCCPS, Surveillance Research Program, released December 2020. <https://seer.cancer.gov/>. Accessed January 23, 2024
- Baker SG. Cancer screening markers: A simple strategy to substantially reduce the sample size for validation. *Med Dec Making.* 2019;39(2):130-136.
- Baker SG, Kramer BS. Simple methods for evaluating 4 types of biomarkers: surrogate endpoint, prognostic, predictive, and cancer screening. *Biomark Insights.* 2020;15:1-8.
- Baker SG, Kramer BS, Corle D. The fallacy of enrolling only high-risk subjects in cancer prevention trials: can we afford a "free lunch."? *BMC Med Res Methodol.* 2004;4:24.