Published in final edited form as:
Cancer Epidemiol Biomarkers Prev. 2022 August 02; 31(8): 1517–1520. doi:10.1158/1055-9965.EPI-22-0232.

Estimating cancer screening sensitivity and specificity using healthcare utilization data: defining the accuracy assessment interval

Jessica Chubak[1,2], **Andrea N. Burnett-Hartman**[3,4], **William E. Barlow**[5], **Douglas A. Corley**[6], **Jennifer M. Croswell**[7], **Christine Neslund-Dudas**[8], **Anil Vachani**[9], **Michelle I. Silver**[10], **Jasmin A. Tiro**[11,12], **Aruna Kamineni**[1]

[1]Kaiser Permanente Washington Health Research Institute, Seattle, WA

[2]Department of Epidemiology, University of Washington, Seattle, WA

[3]Kaiser Permanente Colorado Institute for Health Research, Aurora, CO

[4]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA

[5]Cancer Research and Biostatistics, Seattle, WA

[6]Kaiser Permanente Northern California, Oakland, CA

[7]National Cancer Institute, Bethesda, MD

[8]Department of Public Health Sciences and Henry Ford Cancer Institute, Henry Ford Health System, Detroit, MI

[9]Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA

[10]Division of Public Health Sciences, Washington University School of Medicine, St. Louis, MO

[11]Department of Population and Data Sciences, University of Texas Southwestern Medical Center, Dallas, TX

[12]Simmons Comprehensive Cancer Center, Dallas, TX

Abstract

The effectiveness and efficiency of cancer screening in real-world settings depend on many factors, including test sensitivity and specificity. Outside of select experimental studies, not everyone receives a gold standard test that can serve as a comparator in estimating screening test accuracy. Thus, many studies of screening test accuracy use the passage of time to infer whether or not cancer was present at the time of the screening test, particularly for patients with a negative screening test. We define the accuracy assessment interval as the period of time after a screening test that is used to estimate the test's accuracy. We describe how the length of this interval may bias sensitivity and specificity estimates. We call for future research to quantify bias and uncertainty in accuracy estimates and to provide guidance on setting accuracy assessment interval lengths for different cancers and screening modalities.

Corresponding author: Jessica Chubak, Kaiser Permanente Washington Health Research Institute, 1730 Minor Avenue, Ste. 1600, Seattle, Washington 98101, Jessica.Chubak@kp.org, Telephone: 206-287-2556.

Keywords

cancer screening; clinical epidemiology; methodology; accuracy; verification bias

Epidemiologists have an important role in evaluating healthcare interventions, such as screening. The effectiveness and efficiency of cancer depends on many factors, including test sensitivity and specificity. Estimating screening test sensitivity and specificity using observational data deserves greater attention in the epidemiologic literature, particularly with respect to a concept that we herein define as the *accuracy assessment interval*.

To estimate sensitivity and specificity, a perfect gold standard test would, ideally, be administered to everyone at the same time as the screening test being evaluated. However, this usually does not occur in practice. With an invasive test (e.g., biopsy) as the gold standard, only persons who screen positive are likely to have screening results verified, which can lead to verification bias (1, 2). Studies of screening test accuracy often use the passage of time to infer whether or not cancer was present at the time of the screening test, particularly for patients with a negative screening test (2–4). This is an example of differential verification bias: some people receive a gold standard test (i.e., biopsy or imaging), while others receive an imperfect referent standard (i.e., presence or absence of a cancer diagnosis during a particular time interval) (5). We define the *accuracy assessment interval* as the period of time after a screening test that is used to estimate its accuracy. In this commentary, we explore how the length of the accuracy assessment interval may contribute to bias in estimates of screening test sensitivity and specificity.

The chosen length of the accuracy assessment interval can affect accuracy estimates, especially sensitivity. For example, Hofvind et al. estimated sensitivity of mammography using 2-year and 1-year follow-up intervals (74.9% vs 82.0%, respectively) (6). There were no differences in specificity. In a study of hemoccult testing for colorectal cancer, Allison et al. reported sensitivities of 50%, 43%, and 25% using 1-year, 2-year, and 4-year follow-up periods, respectively (7). Specificity did not differ meaningfully among the follow-up period (98.8% for 1-year and 2-year follow-up periods and 98.7% for 4-year follow-up period). Others have noted that the “optimal duration of follow-up has not been standardized”(4) and that short follow-up intervals might miss cancers that were truly present at screening while long intervals might include cancers that developed after screening (4, 8). But to our knowledge, this issue has not been addressed in detail. We posit that one should attempt to select an accuracy assessment interval, or intervals, that will help them most accurately estimate the screening test’s true sensitivity and specificity. Ideally, the accuracy assessment interval is long enough such that any cancer present at the time of screening will be diagnosed during the interval, while also short enough that new cancers are unlikely to develop, be detected during the interval, and be falsely classified as having been present at the time of the screening exam. For example, if the accuracy assessment interval for FIT is set to 2 years, we would want the following conditions to be met:

1. If a person does not truly have cancer at the time of a screening FIT, the person will not develop cancer and be diagnosed within 2 years.

2. If a person truly has cancer at the time of a screening FIT, the cancer will be diagnosed within 2 years.

It is unlikely that all these conditions will be met for everyone included in an analysis and there will therefore be error in the estimates of sensitivity and specificity; thus, the question becomes how to minimize these errors. Table 1 shows screening test classification based on observed data (i.e., screening test result and accuracy assessment interval classification) and the truth. Classification according to these three factors allows us to conceptualize screening test results as: correct true positives (cTP), incorrect true positives (iTP), correct false positives (cFP), incorrect false positives (iFP), correct true negatives (cTN), incorrect true negatives (iTN), correct false negatives (cTN), and incorrect false negatives (iFN). The terms correct vs. incorrect describe the agreement between the assessment interval classification and the truth. Positive vs. negative refer to the screening test result. True vs. false describe agreement between the screening test results and the accuracy assessment interval classification. For example, a person with a negative FIT who is not diagnosed with cancer during the accuracy assessment interval (e.g., 2 years) is a *correct* true negative if cancer was truly absent at the time of the negative FIT and an *incorrect* true negative if cancer was truly present at the time of the FIT. A person with a positive FIT who is not diagnosed with cancer during the accuracy assessment interval is a *correct* false positive if cancer was truly absent at the time of the negative FIT and in *incorrect* false positive if cancer was truly present at the time of the FIT.

Observed sensitivity and specificity depend, in part, on the relative frequency of different types of errors (i.e., misclassifying false positives as true positives [cFP→iTP], true positives as false positives [cTP→iFP], false negatives as true negatives [cFN→iTN], and true negatives as false negatives [cTN→iFN]) as given by the equations below.

$$Sensitivity_{observed} = \frac{cTP + iTP}{cTP + iTP + cFN + iFN}$$

$$Specificity_{observed} = \frac{cTN + iTN}{cTN + iTN + cFP + iFP}$$

There are tradeoffs associated with lengthening and shortening the accuracy assessment interval. With a longer accuracy assessment interval, we are more likely to correctly classify a cancer that is present at the time of the screening test as “present.” Some negative screening tests shift from being classified as true negatives to false negatives (which decreases estimated sensitivity and specificity) and some positive screening tests shift from being classified as false positives to true positives (which increases estimated sensitivity and specificity). For example:

- Increasing the length of the accuracy assessment interval risks ***misclassifying* TNs as FNs** (cTN→iFN) and ***misclassifying* FPs as TPs** (cFP→iTP). As a result, we might mistakenly conclude that new cancers developing during the accuracy assessment interval had been present at the time of the screening test.

- Increasing the length of the accuracy assessment interval helps ***correctly* identify FNs** (i.e., iTN→cFN) and ***correctly* identify TPs** (i.e., iFP→cTP). Having a longer accuracy assessment interval helps identify cancers that were truly present at the time of the screening test.

The reverse occurs as the length of the accuracy interval is shortened. There is, thus, an inherent tradeoff between lengthening vs. shortening the accuracy assessment interval. Table 2 and Supplementary Figure 1 show the complexity of potential problems caused by accuracy assessment intervals that are too long or too short.

Table 3 presents a hypothetical example showing how changing the accuracy assessment interval can affect estimates of sensitivity and specificity. In this hypothetical population with a 1% cancer prevalence, true sensitivity and specificity are, respectively, 80.0% and 98.0%. We assume that during a 6-month accuracy assessment interval, 0.02% of cancer-free people develop and are diagnosed with cancer and that 70% who screened positive receive a cancer-confirming follow-up test. We assume that during a 12-month accuracy assessment interval 0.05% of cancer-free people develop and are diagnosed with cancer and that 90% who screened positive received a cancer-confirming follow-up test. In this particular hypothetical example, the estimates of specificity are quite similar (and close to the truth) for both accuracy assessment intervals. Sensitivity is underestimated using both accuracy assessment intervals, but to a greater degree with the shorter accuracy assessment interval. Different assumptions would yield different patterns; thus, the table is intended primarily to show that different accuracy assessment intervals can indeed give rise to different accuracy estimates.

Studies that compute sensitivity based on cancers diagnosed *between* screening rounds (9–14) implicitly use the screening interval as the accuracy assessment interval. This approach is intuitive and reasonable, but it *may* not always be the best choice. Apparent interval cancers (i.e., those that occur after a negative screening test and before the next screening test) likely include both those that were missed at a screening test (false negatives) as well as *de novo* cancers. Thus, although the observed interval cancer rate is an important screening quality measure, it is not a pure measure of test sensitivity and may also have limitations with respect to computing specificity. Thus, the screening interval and accuracy assessment interval need not be the same length. They are distinct concepts that serve different purposes. However, setting the accuracy interval to be the same as the screening interval may have some advantages, including increasing the likelihood that a cancer missed by the first screening test will be diagnosed (i.e., that false negatives are correctly classified as such rather than misclassified as true negatives). Future work should comprehensively (and quantitatively) evaluate the benefits and drawbacks of using the screening interval as the accuracy assessment interval. Factors to consider include adherence (particularly differential adherence) to screening guidelines, disease natural history, and the implications of different screening intervals across screening modalities for a particular cancer.

There is need for guidance in the literature about how to set the length of the accuracy assessment interval. Doing so requires information or assumptions about: 1) how rapidly most new cancers develop; 2) how long it takes new cancers to become symptomatic and/or detectable; 3) if/when people will present for follow-up testing after a positive screen and for diagnostic testing if cancer symptoms are present, 4) the recommended screening interval, and 5) rates of loss to follow-up. Many nuances need consideration, such as how to establish accuracy assessment interval(s) when comparing different screening modalities (e.g., FIT versus screening colonoscopy; Pap test alone versus co-testing with Pap and HPV testing).

We acknowledge that there is unlikely to be a perfect accuracy assessment interval for a particular screening test. For example, three years might be within the time frame needed to detect missed cancers but past the point at which some new cancers develop. Ultimately, setting the length of the accuracy assessment interval is a decision based on weighing these tradeoffs. Future studies (both empirical and simulation based) should investigate how to correct for bias and incorporate uncertainty in estimates due to the inherent challenges in having to artificially set an accuracy assessment interval. Existing research on verification bias and imperfect gold standards (15) may help epidemiologists develop guidance and tools to set accuracy assessment intervals and quantify the resulting bias and uncertainty in sensitivity and specificity estimates.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements:

The authors thank the participating PROSPR II Research Centers. A list of the PROSPR II investigators and contributing research staff is provided at: http://healthcaredelivery.cancer.gov/prospr/. The authors are also grateful to Dr. Rebecca Hubbard for her comments on the manuscript.

This manuscript was written as part of the NCI-funded consortium Population-based Research to Optimize the Screening Process (PROSPR II) consortium. The overall aim of PROSPR II is to conduct multi-site, coordinated, transdisciplinary research to evaluate and improve cervical, colorectal, and lung cancer screening processes. The three PROSPR II Research Centers and their associated sites reflect the diversity of US delivery system organizations. UM1CA222035 (PIs Chubak, Corley, Halm, Kamineni, Schottinger, Skinner), UM1CA229140 (PIs Haas, Kamineni, Tiro), UM1CA221939 (PIs Ritzwoller, Vachani), UM24CA221936 (PIs Li and Zheng).

The views expressed here are those of the authors only and do not represent any official position of the National Cancer Institute or National Institutes of Health.

Conflicts of interest:

Dr. Anil Vachani reports personal fees as a scientific advisor to the Lung Cancer Initiative at Johnson & Johnson, grants from MagArray, Inc., Broncus Medical, and Precyte, Inc., outside the submitted work. From 2019-2020, Dr. Chubak was Principal Investigator of a contract from Amgen, Inc. awarded to the Kaiser Foundation Health Plan of Washington to evaluate the accuracy of using electronic health record data to identify individuals with reduced ejection fraction heart failure. No other authors report potential conflicts of interest.

REFERENCES

1. Pepe MS, Alonzo TA. Comparing disease screening tests when true disease status is ascertained only for screen positives. Biostatistics 2001; 2:249–60. [PubMed: 12933537]
2. O'Sullivan JW, Banerjee A, Heneghan C, Pluddemann A. Verification bias. BMJ Evid Based Med 2018; 23:54–5.
3. Lin JS, Piper MA, Perdue LA, Rutter CM, Webber EM, O'Connor E, et al. Screening for Colorectal Cancer: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force. JAMA 2016; 315:2576–94. [PubMed: 27305422]

4. Rosman AS, Korsten MA. Effect of Verification Bias on the Sensitivity of Fecal Occult Blood Testing: a Meta-Analysis. Journal of General Internal Medicine 2010; 25:1211–21. [PubMed: 20499198]

5. Alonzo TA, Brinton JT, Ringham BM, Glueck DH. Bias in estimating accuracy of a binary screening test with differential disease verification. Stat Med 2011; 30:1852–64. [PubMed: 21495059]

6. Hofvind S, Geller BM, Skelly J, Vacek PM. Sensitivity and specificity of mammographic screening as practised in Vermont and Norway. Br J Radiol 2012; 85:e1226–32. [PubMed: 22993383]

7. Allison JE, Feldman R, Tekawa IS. Hemoccult screening in detecting colorectal neoplasm: sensitivity, specificity, and predictive value. Long-term follow-up in a large group practice setting. Ann Intern Med 1990; 112:328–33. [PubMed: 2407166]

8. Glueck DH, Lamb MM, O’Donnell CI, Ringham BM, Brinton JT, Muller KE, et al. Bias in trials comparing paired continuous tests can cause researchers to choose the wrong screening modality. BMC medical research methodology 2009; 9:4. [PubMed: 19154609]

9. Blom J, Tornberg S. Interval cancers in a guaiac-based colorectal cancer screening programme: Consequences on sensitivity. Journal of medical screening 2017; 24:146–52. [PubMed: 28142309]

10. Bordas P, Jonsson H, Nystrom L, Lenner P. Interval cancer incidence and episode sensitivity in the Norrbotten Mammography Screening Programme, Sweden. Journal of medical screening 2009; 16:39–45. [PubMed: 19349530]

11. Elena PM, Nehmat H, Ermes M, Piera C, Maria Q, Guia M, et al. Quality of mammography screening in the Milan programme: evidence of improved sensitivity based on interval cancer proportional incidence and radiological review. Breast (Edinburgh, Scotland) 2009; 18:208–10.

12. Hakama M, Auvinen A, Day NE, Miller AB. Sensitivity in cancer screening. J Med Screen 2007; 14:174–7. [PubMed: 18078561]

13. Sarkeala T, Hakama M, Saarenmaa I, Hakulinen T, Forsman H, Anttila A. Episode sensitivity in association with process indicators in the Finnish breast cancer screening program. International journal of cancer 2006; 118:174–9. [PubMed: 16003756]

14. Zorzi M, Guzzinati S, Puliti D, Paci E. A simple method to estimate the episode and programme sensitivity of breast cancer screening programmes. J Med Screen 2010; 17:132–8. [PubMed: 20956723]

15. Umemneku Chikere CM, Wilson KJ, Allen AJ, Vale L. Comparative diagnostic accuracy studies with an imperfect reference standard - a comparison of correction methods. BMC medical research methodology 2021; 21:67. [PubMed: 33845775]

Table 1.

Screening accuracy classification based on observed screening results, observed cancer diagnoses during accuracy assessment interval, and true cancer status at screening

Screening test result	Cancer diagnosed during accuracy assessment interval	No cancer diagnosed during accuracy assessment interval
Positive	Cancer is truly present: Correct true positives (cTP) and Cancer is truly absent: Incorrect true positives (iTP) (should be false positives) Misclassifying false positives as true positives: • increases estimated sensitivity by inflating its numerator • increases estimated specificity by deflating its denominator	Cancer is truly absent: Correct False Positive (cFP) and Cancer is truly present: Incorrect False Positive (iFP) (should be true positives) Misclassifying true positives as false positives: • decreases estimated sensitivity by deflating its numerator • decreases estimated specificity by inflating its denominator
Negative	Cancer is truly present: Correct False Negative (cFN) and Cancer is truly absent: Incorrect False Negative (iFN) (should be true negatives) Misclassifying true negatives as false negatives: • decreases estimated sensitivity by inflating its denominator • decreases estimated specificity by deflating its numerator	Cancer is truly absent: Correct True Negative (cTN) and Cancer is truly present: Incorrect True Negative (iTN) (should be false negatives) Misclassifying false negatives as true negatives: • increases estimated sensitivity by deflating its denominator • increases estimated specificity by inflating its numerator

cTP=correct true positive, cTN=correct true negative, cFP=correct false positive, cFN=correct false negative, iTP=incorrect true positive, iTN=incorrect true negative, iFP=incorrect false positive, iFN=incorrect false negative

Table 2.

Impact on estimated sensitivity and specificity from accuracy assessment intervals that are too long or too short

Accuracy assessment interval length	Problem	Possible causes	Impact on estimates of sensitivity and specificity
Too long	Cancers that develop and are detected after screening are incorrectly classified as having been present at screening	Rate of new cancer development is fast relative to accuracy assessment interval length	Misclassifying false positives as true positives (cFP→iTP) increases estimated sensitivity and specificity
			Misclassifying true negatives as false negatives (cTN→iFN) decreases estimated sensitivity and specificity
	Among people with a negative screening test, cancers detected at a *subsequent* screening test are incorrectly classified as having been present at the initial screening test	Recommended screening interval is shorter than accuracy assessment interval	Misclassifying true negatives as false negatives (cTN→iFN) decreases estimated sensitivity and specificity
Too short	Cancers that were truly present at screening are not detected during the accuracy assessment interval	People with positive screening tests do not have sufficiently rapid follow-up diagnostic tests relative to accuracy assessment interval length	Misclassifying true positives as false positives (cTP→iFP) decreases estimated sensitivity and specificity
		Slow progression from asymptomatic cancer to symptom-detected cancer	Misclassifying true positives as false positives (cTP→iFP) decreases estimated sensitivity and specificity
			Misclassifying false negatives as true negatives (cFN→iTN) increases estimated sensitivity and specificity
		Recommended screening interval is longer than accuracy assessment interval, which decreases opportunities for detection	Misclassifying false negatives as true negatives (cFN→iTN) increases estimated sensitivity and specificity

cTP=correct true positive, cTN=correct true negative, cFP=correct false positive, cFN=correct false negative, iTP=incorrect true positive, iTN=incorrect true negative, iFP=incorrect false positive, iFN=incorrect false negative

Table 3.

Hypothetical example of the impact on estimated sensitivity and specificity of using different accuracy assessment interval lengths

Unobserved truth						
	Cancer	No cancer	Total			
Screen +	8	20	28			
Screen −	2	970	972			
Total	10	990	1000			
Sensitivity	80.0%					
Specificity	98.0%					
Observed with 6-month accuracy assessment interval						
	Cancer	No cancer	Total		TP misclassified as FP	2.400
Screen +	5.604	22.396	28		TN misclassified as FN	0.194
Screen −	2.194	969.806	972		FP misclassified as TP	0.004
Total	7.798	992.202	1000		FN misclassified as TN	0
Sensitivity	71.9%					
Specificity	97.7%					
Observed with 12-month accuracy assessment interval						
Total	Cancer	No cancer	Total		TP misclassified as FP	0.800
Screen +	7.210	20.790	28		TN misclassified as FN	0.485
Screen −	2.485	969.515	972		FP misclassified as TP	0.010
Total	9.695	990.305	1000		FN misclassified as TN	0
Sensitivity	74.4%					
Specificity	97.9%					

FP=false positive; FN=false negative; TP=true positive; TN=true negative

[1] We assume that during a 6-month accuracy assessment interval, 0.02% of the no-cancer group develops and is diagnosed with cancer and that 70% of the screen-positive group receives a cancer-confirming follow-up test.

[2] We assume that during a 12-month accuracy assessment interval 0.05% of the no-cancer group develops and is diagnosed with cancer and that 90% of the screen-positive group receives a cancer-confirming follow-up test.