Journal of Clinical Epidemiology 170 (2024) 111359

Journal of Clinical Epidemiology

COMMENTARY

Interpreting diagnostic accuracy studies based on retrospective routinely collected data

Stephen H. Bradley[a,*], Bethany Shinkins[b], Gary Abel[c], Matthew E.J. Callister[d,e]

[a]NIHR Academic Clinical Lecturer, Leeds Institute of Health Sciences, University of Leeds, Room 10.39, Worsley Building, LS2 9JT, Leeds, UK
[b]Warwick Medical School, University of Warwick, Coventry, UK
[c]Department of Health and Community Sciences, University of Exeter, Exeter, UK
[d]Leeds Teaching Hospitals NHS Trust, Leeds, UK
[e]Leeds Institute of Health Sciences, University of Leeds, Leeds, UK

Accepted 4 April 2024; Published online 10 April 2024

Keywords: Diagnostic accuracy studies; Routinely collected data; Observational data; Test accuracy; Bias; Reference standard

Producing prospective evidence on the diagnostic accuracy of a test can be costly and time-consuming, particularly in settings where the prevalence of the target disease is low, requiring recruitment of large numbers of individuals. If a test is already in use in routine clinical practice, estimating the diagnostic accuracy of tests based on retrospective, routinely collected health-care data is often an appealing alternative, avoiding the logistical challenges and costs associated with large prospective studies along with potential ethical considerations. However, interpreting the evidence from such studies requires understanding of their underlying assumptions and resulting potential for bias.

1. Use of clinical diagnosis as reference standard

The 'reference standard' is the test(s) or information used to establish the presence or absence of the target condition [1]. Ideally, the reference standard used is highly accurate, undertaken in the whole study population, and conducted at the same time as the index test. All diagnostic accuracy studies rely on the reference standard providing as close to the truth in terms of the disease status of each participant as possible.

In routine clinical practice, particularly in primary care, it would be unusual for everybody who is suspected of having a target condition to receive a reference standard. Typically, a range of different tests and pieces of information are used to rule out the target condition or triage individuals to more definitive investigations. Since these tests are only generally accompanied by a more definitive test if they are positive, retrospective analysis of such tests used in routine practice lacks a consistent contemporaneous reference standard.

To overcome this issue, clinical diagnosis within a specified period of time from the index test being conducted is often used as the reference standard instead [2–7]. This approach may be suitable for diseases which without intervention are not expected to resolve without intervention and requires two further assumptions.

1) If undetected by the test, the target condition would subsequently be diagnosed within the specified time period, that is, the disease would progress and manifest with persisting or worsening symptoms, leading to its diagnosis within that time period.
2) That the target condition was present when the test was performed and did not arise de novo within the specified time period.

Violation of the first assumption will lead to overestimation of diagnostic accuracy (see Fig 1, Patient B). This is because a negative index test result will be incorrectly labeled as a true negative if the disease is present but not diagnosed within the specified period. As the disease is actually present, this result should be labeled as a false negative.

Violation of the second assumption will lead to underestimation of diagnostic accuracy (see Fig 1, Patient C). This is because a negative test result will be incorrectly labeled as a false negative because the disease has been diagnosed during the specified time period. As the disease wasn't present at the time of the index test being conducted the correct designation should actually be true negative.

* Corresponding author. University of Leeds, Room 10.39, Worsley Building, LS2 9JT, UK.
E-mail address: medsbra@leeds.ac.uk (S.H. Bradley).

https://doi.org/10.1016/j.jclinepi.2024.111359

What is new?

Key findings

- Studies often use retrospective routinely collected data to produce estimates of diagnostic accuracy, particularly in low prevalence settings.
- It can be difficult to define an appropriate reference standard when using retrospective routinely collected data without having to make some assumptions.

What this adds to what was known?

- We explain these assumptions and how these may lead to biased estimates of diagnostic accuracy.

What is the implication and what should change now?

- Estimation of diagnostic accuracy using retrospective observational data is subject to biases and the limitations of such approaches should be explicitly acknowledged in published studies.
- Given the limitations inherent in assessing test accuracy using routinely collected data, consideration should be given to prospective evaluations.

As calculations of sensitivity and specificity require accurate estimation of both false negatives and true negatives, respectively, violations of the above assumptions are sufficient to produce biased estimates of diagnostic accuracy. Positive test results are not subject to the same issues as long as the subsequent further definitive investigation that will confirm or exclude presence of the disease occurs in a timely manner. Supplementary figure 1 presents corresponding biases that could result from positive test results.

Selection of an appropriate interval requires clinical judgment; the interval needs to be sufficiently long to capture as many cases of the target condition as possible that were present at the time of the test, but not excessively long such that substantial volumes of the target condition could have arisen after the test. Selecting an interval that optimally achieves this balance is challenging. Consider lung cancer, a disease which is understood to advance rapidly and with fatal consequences if untreated. The mean duration in which lung cancer is detectable before it causes symptoms (sojourn time) has been estimated as between three and 6 years [8]. However, screening programs have demonstrated some instances of highly aggressive cancers that arise much more rapidly between screening rounds [9]. Such 'interval' cancers have been found when screening for several different cancers [10–12] despite having longer estimated sojourn times compared to lung cancer [13–15].

Since these rapidly progressive cancers are relatively atypical, overestimation of diagnostic accuracy by failing to account for slow growing cancers is probably the bigger concern [16]. Smaller tumors, which are generally harder to detect, may also take longer to be diagnosed and be more likely to be picked up outside of the specified time period. This will not only lead to an overestimation

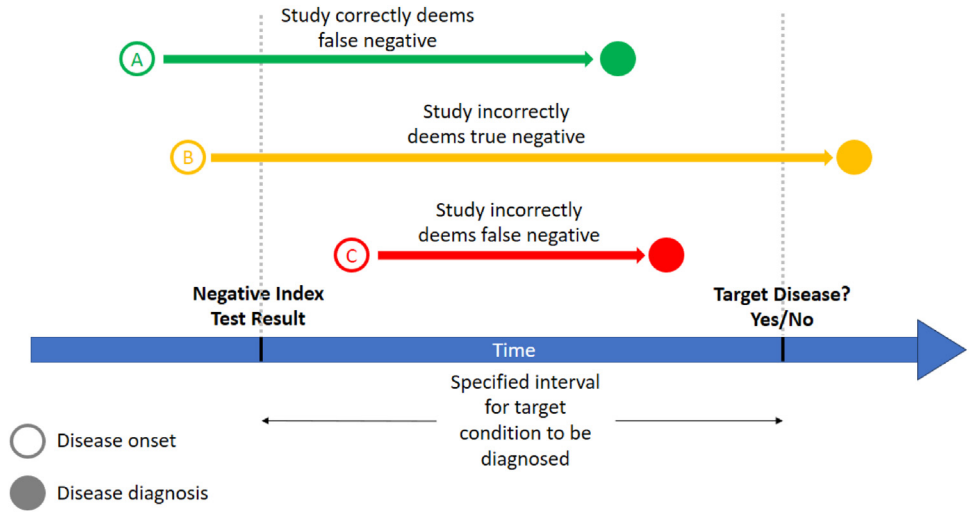

Figure 1. Diagram to show risk of mischaracterization of test performance resulting from reliance on clinical diagnosis within specified interval of the index test (eg, 1 year) as clinical reference standard, for negative test results. Patient A: test does not identify the disease, subsequently diagnosed within specified interval, study correctly labels as false negative. Patient B: test misses the disease, subsequently diagnosed outside the specified interval; this is a false negative, but the study incorrectly labels as true negative thereby overestimating test sensitivity and specificity. Patient C: disease not present at the time of test, subsequently diagnosed within specified interval; this is a true negative, but the study miscategorizes as false negative thereby underestimating sensitivity and specificity. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

of both specificity and sensitivity, but also risks inappropriate inferences about the clinical utility of a test as, arguably, smaller tumors are often those that benefit the most from earlier detection. The opposite can also be true; some smaller, slow progressing, tumors may be less clinically significant as they may not progress to a stage where they cause any problems within the patient's lifetime. Detecting these tumors can lead to overdiagnosis and may result in over treatment. This is why producing evidence on outcomes beyond diagnostic accuracy, particularly in the context of earlier diagnosis, is so important [17]. In the case of lung cancer, data from a screening trial indicates that the true accuracy of chest x-ray may be much lower than has been reported when clinical diagnosis within 1 year is used as a reference standard (Supplementary material).

These considerations are not exclusive to cancer diagnosis and apply to any diseases for which such methodology is used. Such studies also rely on the accuracy of the data source used to ascertain whether the target condition was present or absent. Studies using secondary care records for specific providers [2,6,18] may assume that patients continued to receive care from the same provider and did not change provider due to patient choice or geographic relocation. This can be misleading; for example, studies based on national primary care datasets have been shown to capture varying proportions of cancer diagnoses ranging from 84% of melanoma cases to 97% of lung cancer cases [19]. Where national registers of diagnoses exist, linkage of this data could be used to minimize this issue, although emigration will remain a problem (albeit to a minimal extent).

2. Conclusion

Where tests are already being used in clinical practice, analysis of retrospective routinely collected data can be useful in evaluating the diagnostic accuracy of these tests in settings where prevalence of the target disease is low and prospective studies would be expensive and require large numbers of participants. This type of data is also likely to be increasingly used for postmarket evaluations of test performance to meet new regulatory requirements [20]. Here we explain the limitations of this type of data, particularly the assumptions required to use clinical diagnosis within a specified time period as the reference standard and the potential implications on estimates of diagnostic accuracy. This limitation may be problematic in the context of earlier diagnosis, as it could lead to overestimation of test accuracy by overlooking the very cases of disease which are most amenable to intervention following early detection.

The case for prospective diagnostic accuracy studies remains strong, despite the difficulties in undertaking such studies. Recommendations for conducting prospective diagnostic accuracy studies in contexts where disease prevalence is relatively low have been made and may be useful in thinking through feasible study designs [21]. We recommend that reporting of diagnostic accuracy studies should make explicit reference to the limitations of observational data and sensitivity analyses (ie, varying the length time period for a clinical diagnosis) could be conducted to explore the potential impact that this decision is having on diagnostic accuracy estimates.

CRediT authorship contribution statement

Stephen H. Bradley: Writing − original draft, Conceptualization. **Bethany Shinkins:** Writing − review & editing. **Gary Abel:** Writing − review & editing, Formal analysis. **Matthew E.J. Callister:** Writing − review & editing, Conceptualization.

Data availability

Supplementary material includes discussion of secondary analysis of data from the National Lung Cancer Screening Trial, along with Stata files.

Declaration of competing interest

None.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2024.111359.

References

[1] Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015;351:h5527.
[2] Bradley SH, Bhartia BS, Callister ME, Hamilton WT, Hatton NLF, Kennedy MP, et al. Chest X-ray sensitivity and lung cancer outcomes: a retrospective observational study. Br J Gen Pract 2021; 71(712):e862−8.
[3] Bhartia BSK, Hatton NLF, Aslam R, Bradley SHB, Darby M, Hamilton WT, et al. A prospective cohort evaluation of the sensitivity and specificity of the chest X-ray for the detection of lung cancer in symptomatic adults. Eur J Radiol 2021;144:109953.
[4] Taylor CJ, Ordóñez-Mena JM, Lay-Flurrie SL, Goyder CR, Taylor KS, Jones NR, et al. Natriuretic peptide testing and heart failure diagnosis in primary care: diagnostic accuracy study. Br J Gen Pract 2023;73(726):e1−8.
[5] Funston G, Hamilton W, Abel G, Crosbie EJ, Rous B, Walter FM. The diagnostic performance of CA125 for the detection of ovarian and non-ovarian cancer in primary care: a population-based cohort study. PLoS Med 2020;17(10):e1003295.
[6] Nicholson BD, James T, Paddon M, Justice S, Oke JL, East JE, et al. Faecal immunochemical testing for adults with symptoms of colorectal cancer attending English primary care: a retrospective cohort

study of 14 487 consecutive test requests. Aliment Pharmacol Ther 2020;52(6):1031−41.
[7] Miller A, Nightingale AL, Sammon CJ, Mahtani KR, Holt TA, Mc Hugh NJ, et al. Estimating the diagnostic accuracy of rheumatoid factor in UK primary care: a study using the Clinical Practice Research Datalink. Rheumatology 2015;54:1882−9.
[8] Ten Haaf K, van Rosmalen J, de Koning HJ. Lung cancer detectability by test, histology, stage, and gender: estimates from the NLST and the PLCO trials. Cancer Epidemiol Biomarkers Prev 2015;24(1):154−61.
[9] Horeweg N, Scholten ET, de Jong PA, van der Aalst CM, Weenink C, Lammers J-WJ, et al. Detection of lung cancer through low-dose CT screening (NELSON): a prespecified analysis of screening test performance and interval cancers. Lancet Oncol 2014;15(12):1342−50.
[10] Singh S, Singh PP, Murad MH, Singh H, Samadder JN. Prevalence, risk factors, and outcomes of interval colorectal cancers: a systematic review and meta-analysis. Am J Gastroenterol 2014;109(9):1375−89.
[11] Rebolj M, Cuschieri K, Mathews CS, Pesola F, Denton K, Kitchener H. Extension of cervical screening intervals with primary human papillomavirus testing: observational study of English screening pilot data. BMJ 2022;377:e068776.
[12] Niraula S, Biswanger N, Hu P, Lambert P, Decker K. Incidence, characteristics, and outcomes of interval breast cancers compared with screening-detected breast cancers. JAMA Netw Open 2020;3(9):e2018179.
[13] Brenner H, Altenhofen L, Katalinic A, Lansdorp-Vogelaar I, Hoffmeister M. Sojourn time of preclinical colorectal cancer by sex and age: estimates from the German national screening colonoscopy database. Am J Epidemiol 2011;174:1140−6.
[14] Pashayan N, Duffy SW, Pharoah P, Greenberg D, Donovan J, Martin RM, et al. Mean sojourn time, overdiagnosis, and reduction in advanced stage prostate cancer due to screening with PSA: implications of sojourn time on screening. Br J Cancer 2009;100:1198−204.
[15] Broder MS, Ailawadhi S, Beltran H, Blakely LJ, Budd GT, Carr L, et al. Estimates of stage-specific preclinical sojourn time across 21 cancer types. J Clin Oncol 2021;39:e18584.
[16] Bossuyt PMM, Reitsma JB, Linnet K, Moons KGM. Beyond diagnostic accuracy: the clinical utility of diagnostic tests. Clin Chem 2012;58:1636−43.
[17] di Ruffano LF, Hyde CJ, McCaffery KJ, Bossuyt PMM, Deeks JJ. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. BMJ 2012;344:e686.
[18] Bailey SER, Abel GA, Atkins A, Byford R, Davies S-J, Mays J, et al. Diagnostic performance of a faecal immunochemical test for patients with low-risk symptoms of colorectal cancer in primary care: an evaluation in the South West of England. Br J Cancer 2021;124:1231−6.
[19] Strongman H, Williams R, Bhaskaran K. What are the implications of using individual and combined sources of routinely collected data to identify and characterise incident site-specific cancers? a concordance and validation study using linked English electronic health records data. BMJ 2020;10(8):e037719.
[20] MDCG 2021-24 Guidance on classification of medical devices. European Union; 2021. Available at: https://health.ec.europa.eu/system/files/2021-10/mdcg_2021-24_en_0.pdf. Accessed February 19, 2024.
[21] Holtman GA, Berger MY, Burger H, Deeks JJ, Donner-Banzhoff N, Fanshawe TR, et al. Development of practical recommendations for diagnostic accuracy studies in low-prevalence situations. J Clin Epidemiol 2019;114:38−48.