Verification bias

Jack W O'Sullivan,[1] Amitava Banerjee,[2] Carl Heneghan,[3] Annette Pluddemann[3]

10.1136/bmjebm-2018-110919

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/bmjebm-2018-110919).

[1]Centre for Evidence-Based Medicine, Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK
[2]Farr Institute of Health Informatics, University College London, London, UK
[3]Department of Primary Care Health Sciences, University of Oxford, Oxford, UK

Correspondence to:
Dr Jack W O'Sullivan, Centre for Evidence-Based Medicine, Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, OX1 2JD, UK; jack.osullivan@phc.ox.ac.uk

Check for updates

To cite: O'Sullivan JW, Banerjee A, Heneghan C, *et al*. *BMJ Evidence-Based Medicine* 2018;**23**:54–55.

Abstract

This article is part of the Catalogue of Bias series. We present a description of verification bias, and outline its potential impact on research studies and the preventive steps to minimise its risk. We also present teaching slides in the online supplementary file. Verification bias (sometimes referred to as 'work-up bias') concerns the test(s) used to confirm a diagnosis within a diagnostic accuracy study. Verification bias occurs when only a proportion of the study participants receive confirmation of the diagnosis by the reference standard test, or if some participants receive a different reference standard test.

Background

Diagnostic accuracy studies determine the ability of a new test to rule in (confirm) or rule out (exclude) a disease. To achieve this, investigators subject all study participants to both an index test (the new test) and a reference standard test (usually the test that is considered to be best at diagnosing the target condition and is often referred to as the 'gold standard'). The results of the index test and the reference standard test are then compared, and the number of patients who tested true positive (TP), true negative (TN), false positive (FP) and false negative (FN) is determined. The sensitivity and specificity of the index test can then be calculated (sensitivity=TP/(TP+FN), specificity=TN/(TN+FP)).

Verification bias occurs when only some of the participants who received the index test go on to have the reference standard test, or when some participants receive one reference standard test and others have a different reference standard test. Accurate and consistent confirmation of disease is crucial in diagnostic accuracy studies; if two different reference standard tests are used, varying accuracy of disease confirmation is introduced.

There are two types of verification bias: *partial* verification bias, where only some patients receive the reference standard test, with the other patients not receiving any reference standard test; and *differential* verification bias, where two different reference standard tests are used, typically alternating depending on whether the index test was positive or negative.

Many reference tests are invasive, expensive or carry a procedural risk (eg, angiography, biopsy and surgery), and therefore in many studies verification bias is unavoidable. This paper details examples, impact and preventive steps for verification bias.

Example

A study assessed the accuracy of D-dimer testing for diagnosing deep vein thrombosis (DVT).[1] Patients who had a positive D-dimer result were further assessed with ultrasonography (reference standard test 1), whereas patients who had negative D-dimer results were assessed with routine 3-month clinical follow-up (reference standard test 2). Therefore, patients who had a DVT but a negative D-dimer may not have been diagnosed by routine follow-up (symptoms may have resolved in the interim). This study design thus risks underestimating the number of FNs and thus may overestimate the sensitivity of a new test.

Impact

Verification bias affects the accuracy of an index test in a diagnostic accuracy study. Partial verification bias will frequently underestimate the number of FN patients, and as such will often overestimate the sensitivity. The impact of differential verification bias is less clear-cut. The effect of differential verification bias on the sensitivity and specificity of the index test depends on the diagnostic accuracy of the two reference standard tests, relative to each other.

Further research is required to adequately quantify the effect of verification bias on diagnostic accuracy. A 2006 analysis of 31 meta-analyses of diagnostic accuracy studies stated that 'studies that relied on 2 or more reference standards to verify the results of the index test reported (diagnostic) odds ratios (DOR) that were on average 60% higher than the (diagnostic) odds ratios in studies that used a single reference standard'.[2] The result, however, was not statistically significant. The same study reported that studies that were subject to partial verification bias overestimated diagnostic ORs (DORs) by 10%, although this was also non-significant[2] (DOR is a single estimate of a test's accuracy, taking into account both sensitivity and specificity).[3]

Studies where the reference standard test is an expensive or invasive test are particularly prone to verification bias. For instance, studies assessing the diagnostic accuracy of faecal occult blood test (FOBT) often only use a confirmatory colonoscopy on patients who test positive with FOBT. A meta-analysis comparing the diagnostic accuracy of FOBT for colorectal cancer found that the pooled sensitivity of FOBT without verification bias was significantly lower than those studies with this bias (0.36 vs 0.70). The pooled specificity of the studies without verification bias was also higher (0.96 vs 0.88).[4] The authors concluded that 'the sensitivity of guaiac-based FOBT for colorectal cancer has been overestimated as a result of verification bias. This

test may not be sensitive enough to serve as an effective screening option for colorectal cancer'.[4]

Preventive steps

Ideally, in a diagnostic accuracy study, all patients should receive the same reference test.

However, obtaining a reference test in every patient may not be ethical, practical or cost-effective. When this is not achievable, there are a collection of statistical methods that can be employed to try and account for this bias. Like all statistical adjustments,[5] correction for verification bias attempts to reclassify patients into a group that reflects their actual outcome. For correction of verification bias, statistical approaches attempt to reclassify patients who tested negative into the FN category (to account for the number of FNs missed due to verification bias). Begg and Greenes[6 7] proposed a widely used method to correct for verification bias.[8] Their method uses Bayesian techniques; an empirical probability of verification (receiving the reference standard test) is calculated and then applied to the observed TP, FP, TN and FNs to generate the adjusted estimates (this paper from Cronin and Vickers[8] and its appendix files explain this in more detail).

Nevertheless, statistical adjustment in diagnostic accuracy studies subject to verification bias should be approached with caution. Diagnostic accuracy studies subject to verification bias often have low numbers of FN patients. In these cases, the application of statistical adjustment can substantially, and most likely inappropriately, affect the results.[8] When the total number of FNs is low, reclassification can have dramatic effects on the sensitivity and specificity of a test. For instance, in a study that aimed to determine the accuracy of a human papillomavirus DNA test to diagnose cervical cancer,[9] the reported sensitivity was 100%, but the reclassification of one patient into the FN category would have reduced the sensitivity to 70%.[8]

The obvious solution to avoid verification bias is to use one reference standard test in all patients. When this is not possible, the above statistical adjustment techniques are appropriate in situations where there are an adequate number of FN patients. When the number of FNs is low, randomly sampling a number of TN patients and then confirming disease status with the reference standard test is recommended, although this may unnecessarily risk adverse effects to healthy patients or be expensive.

For teaching purposes, we have provided slides in the online Supplementary file 1.

Discussion

Verification bias is common and can have dramratic effects on the sensitivity and specificity of diagnostic tests.[8] We have detailed what verification bias is, how it can impact real clinical practice and steps to avoid its effect. Researchers should be familiar with this common bias and its consequences. Where verification bias is unavoidable, researchers should always clearly discuss the potential impact of this bias on their results, as well as the potential clinical consequences.

Acknowledgements We would like to acknowledge the support of the McCall MacBain Foundation, which made the Catalogue of Bias possible.

Contributors JWOS, AB, CH and AP conceived the idea. JWOS, AB and AP wrote the initial outline, which JWOS converted into a manuscript. All authors reviewed and approved the manuscript.

Competing interests None declared.

Provenance and peer review Commissioned; internally peer reviewed.

References

1 **Büller HR**, Ten Cate-Hoek AJ, Hoes AW, *et al*. Safely ruling out deep venous thrombosis in primary care. *Ann Intern Med* 2009;**150**:229–36.
2 **Rutjes AW**, Reitsma JB, Di Nisio M, *et al*. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ* 2006;**174**:469–76.
3 **Glas AS**, Lijmer JG, Prins MH, *et al*. The diagnostic odds ratio: a single indicator of test performance. *J Clin Epidemiol* 2003;**56**:1129–35.
4 **Rosman AS**, Korsten MA. Effect of verification bias on the sensitivity of fecal occult blood testing: a meta-analysis. *J Gen Intern Med* 2010;**25**:1211–21.
5 **O'Sullivan J**. Controversies in PSA screening. *Evid Based Med* 2017;**22**:198.
6 **Gray R**, Begg CB, Greenes RA. Construction of receiver operating characteristic curves when disease verification is subject to selection bias. *Med Decis Making* 1984;**4**:151–64.
7 **Begg CB**, Greenes RA. Assessment of Diagnostic Tests When Disease Verification is Subject to Selection Bias Published by : International Biometric Society Stable URL. 2009;**39**:207–15.
8 **Cronin AM**, Vickers AJ. Statistical methods to correct for verification bias in diagnostic studies are inadequate when there are few false negatives: a simulation study. *BMC Med Res Methodol* 2008;**8**:1–9.
9 **Dannecker C**, Siebert U, Thaler CJ, *et al*. Primary cervical cancer screening by self-sampling of human papillomavirus DNA in internal medicine outpatient clinics. *Ann Oncol* 2004;**15**:863–9.
10 **Whiting PF**, Rutjes AW, Westwood ME, *et al*. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011;**155**:529–36.