Recognising bias in studies of diagnostic tests part 2: interpreting and verifying the index test

Bory Kea,[1] M Kennedy Hall,[2] Ralph Wang[3]

[1]Emergency Medicine, Oregon Health and Science University, Portland, Oregon, USA
[2]Emergency Medicine, University of Washington, Seattle, Washington, USA
[3]Emergency Medicine, University of California, San Francisco, San Francisco, California, USA

Correspondence to
Dr Bory Kea, Emergency Medicine, Oregon Health and Science University, Portland, OR 3181 SW, USA; kea@ohsu.edu

Received 13 January 2019
Revised 27 March 2019
Accepted 2 April 2019
Published Online First 20 June 2019

ABSTRACT

Multiple pitfalls can occur with the conduct and analysis of a study of diagnostic tests, resulting in biased accuracy. Our conceptual model includes three stages: patient selection, interpretation of the index test and disease verification. In part 2, we focus on (1) Interpretation bias (or workup bias): where the classification of an indeterminate index test result can bias the accuracy of a test or how lack of blinding can bias a subjective test result, and (2) Disease verification bias: where the index test result is incorporated into the gold standard or when the gold standard is applied only to a select population as the gold standard is an invasive test. In an example with age-adjusted D-dimer for pulmonary embolism, differential verification bias was a limitation due to the use of two gold standards—CT for a high-risk population and follow-up for symptoms in a low-risk population. However, there are circumstances when certain choices in study design are unavoidable, and result in biased test characteristics. In this case, the informed reader will better judge the quality of a study by recognising the potential biases and limitations by being methodical in their approach to understanding the methods, and in turn, better apply studies of diagnostic tests into their clinical practice.

In part 1 of *Recognising Bias in Studies of Diagnostic Tests*, we discussed how suboptimal patient selection could lead to bias in studies of diagnostic tests[1]. In part 2 of this series, we explain how the interpretation and verification of diagnostic test results can lead to bias and methods to minimise those biases (table 1).

EXAMPLE: ULTRASOUND FOR DEEP VEIN THROMBUS

A healthy 85-year-old patient presents to your ED with a swollen right leg for 2 days. She has no history of travel, cancer, recent immobility, tobacco use, surgery or prior deep vein thrombosis (DVT). She does have a history of heart failure with some lower extremity oedema bilaterally, but the swelling is asymmetrical and painful. Her vital signs are normal, and the patient denies chest pain or shortness of breath. You are concerned about a DVT, and wonder if an ED point-of-care ultrasound (POCUS) can accurately diagnose DVT. You are concerned about the test characteristics of POCUS—will a negative compression test effectively rule out DVT? Can you use a negative test result and confidently send the patient home?

In order to answer this question, you review the methods of two studies of POCUS for lower extremity DVT, paying special attention to how the index test was performed and interpreted. You discover that one study reported excellent sensitivity and specificity, but did not describe blinding of the POCUS operators, while a second study reported more modest results but described blinding in detail. You suspect that the true test characteristics of POCUS may not be as excellent as the first study indicated. In part 2, we examine how test characteristics might change on alteration of how the index test is interpreted and verified.

INTERPRETATION BIAS

All diagnostic tests must be interpreted, and how they are interpreted and/or included in the analysis of a study can alter the performance of the test. Diagnostic test results are not always clearly positive or negative—either due to the limitations of the test or the ability of the interpreter.

Interpretation bias due to indeterminate results

Excluding indeterminate results from an analysis may result in spectrum bias. If patients with indeterminate results are not excluded, investigators must carefully consider and explicitly state a priori whether indeterminate results will be considered positive or negative in the analysis.

A priori decisions on indeterminate results will allow for a clear interpretation of the results, including any sensitivity analyses that are conducted. A familiar example of the problem with indeterminate exams occurs in patients undergoing POCUS, where some studies are technically difficult. In a study of POCUS for DVT by Frazee *et al*, the investigators categorised indeterminate exams as positives a priori.[2] This would have the effect of potentially increasing the number of false positives, thus decreasing the specificity. However, the authors' decision to manage indeterminates in this manner is reasonable because in clinical practice, those with indeterminate POCUS exams would go on to receive a confirmatory study, as would those with positive studies. In addition, sensitivity analyses can be performed to understand how sensitivity and specificity vary depending on how indeterminates are handled. Recognising this bias would require a close reading of the methods to determine exactly how the investigators planned to deal with indeterminate diagnostic studies and whether this would apply to your own setting.

Interpretation bias due to review bias

All clinicians who interpret tests are subject to the influence of prior information, or the available

Linked

► http://dx.doi.org/10.1136/emermed-2019-208446

Check for updates

To cite: Kea B, Hall MK, Wang R. *Emerg Med J* 2019;**36**:501–505.

Table 1 Types of bias introduced through diagnostic testing. Greyed out text denotes the biases discussed in, 'Recognising Bias in Studies of Diagnostic Tests Part 1: Patient Selection' [1]. Bold text denotes the biases discussed in this manuscript, 'Recognising Bias in Studies of Diagnostic Tests Part 2: Interpreting and Verifying the Index Test'

	Type of bias		*Recognising bias*	*Effect on accuracy*
Part 1: Suboptimal patient selection	Partial verification workup or referral bias		Only patients tested with the gold standard are included; patients with positive index test are more likely to get the gold standard.	Falsely increases sensitivity by lowering the rate of false negatives.
	Spectrum bias through case-control design		Inclusion of 'sickest of the sick' or 'wellest of the well'	Falsely increases sensitivity and specificity.
	Spectrum bias through dropping indeterminate subjects		Ask 'did they describe their method for handling indeterminates?'	Falsely increases sensitivity if excluded indeterminates have mild disease. Falsely increases specificity if excluded indeterminates are not diseased.
	Spectrum bias through convenience sampling		Look for screening modality in methods section.	Falsely elevates sensitivity and specificity when sampling excludes difficult, indeterminate or ambiguous patients.
Part 2: Interpreting and verifying the index test	**Interpretation**	**Indeterminate**	**When indeterminate results are considered dichotomously as positive or negative.**	**It can overestimate or underestimate the accuracy of the test depending on how indeterminates are included.**
		Review	**Occurs when the person interpreting the diagnostic test has access to the gold standard test.**	**Usually falsely increases the sensitivity and specificity of the index test.**
	Verification	**Incorporation**	**Occurs when the criteria for a gold standard includes the results of the diagnostic test.**	**Falsely increases the sensitivity and specificity of the index test.**
		Double gold standard (differential verification)	**Occurs when gold standard test is invasive or expensive, and is only performed when index test result is positive.**	**Falsely increases the sensitivity and specificity of the index test.**

context of the index test (test under study) results. In clinical practice, imaging orders often require reason for exam information, which provide radiologists with the context of the image they will interpret. In most cases, this is beneficial to the interpreter in such a way that they can be more observant of the findings, and improve their diagnostic accuracy—such as an acute fracture diagnosis due to the knowledge of new pain at the concordant location compared with the interpretation of an age-indeterminate fracture or artefact as the radiologist is unaware of acute pain and trauma to the area. Thus, the subjective components can influence the interpretation of a diagnostic test.

In studies of diagnostic tests, review bias can occur when the interpreter of an index test is unblinded to whether the patient received the gold standard (verification test) or its results. Consider this imaginary study where a new diagnostic test, CT coronary angiography (CTCA), is being compared with traditional invasive coronary angiograms (ICAs) (gold standard) for coronary artery disease burden. If the adjudicator of the CTCA results has access to the final ICA results, they may potentially alter their interpretation of the index test to agree with the results of the gold standard—increasing agreement between their interpretation and the gold standard results, thus falsely increasing the sensitivity and specificity of the CTCA. Blinding is key to minimising review bias.

DISEASE VERIFICATION

Disease verification allows the investigator to determine if the disease is present or absent in the study participants. In studies of diagnostic tests, participants should receive both the index test and the gold standard test, which verifies the presence of absence of disease. The gold standard test often consists of an expensive or invasive procedure, or expert case review. Bias may occur if the gold standard is applied only to a subset of the cohort or includes the interpretation of the index test as part of the gold standard.[3] However, in certain circumstances, it may be unethical and/or infeasible to perform an invasive test to verify a disease state in all participants.

Disease verification due to differential verification

Despite the similarities in terminology, *partial verification* and *differential verification* cause bias through different mechanisms. *Partial verification bias* applies when participants with a positive index test are more likely to receive the gold standard (eg, positive electrokardiogram (EKG) stress test patients undergo coronary catheterisation) and only those who receive the gold standard are included in the patient population. This enriches the study population with true positives causing a bias towards increased sensitivity. *Differential verification bias*, also known as double gold standard bias, occurs when all patients are verified but more than one gold standard is used—such that two gold standards classify the presence of disease differently.[3] This often occurs when the gold standard test is invasive or expensive and is only performed when the index test result is positive. For example, the positive-test group receives an immediate invasive imaging study to determine if disease is present or not, while the negative-test group receives a 3-month follow-up appointment and is assessed according to symptoms.

Reconsider our initial 85-year-old female patient, and instead of presenting with leg swelling, she presents with shortness of breath, cough and some pleuritic chest pain for more than 8 hours. Because of the shortness of breath and pleuritic chest pain, you are concerned about an acute pulmonary embolism (PE). Of course, you would like to avoid unnecessary CT, especially one that requires contrast in an older patient.

After risk-stratification with the Wells' Criteria for Pulmonary Embolism,[4] you determine that a D-dimer test (rather than immediate CT) is appropriate. The test comes back at 750 μg/L, which is marked 'abnormal result'. You are aware that the D-dimer test has excellent sensitivity but poor specificity, in part because D-dimer is elevated in older patients—not necessarily

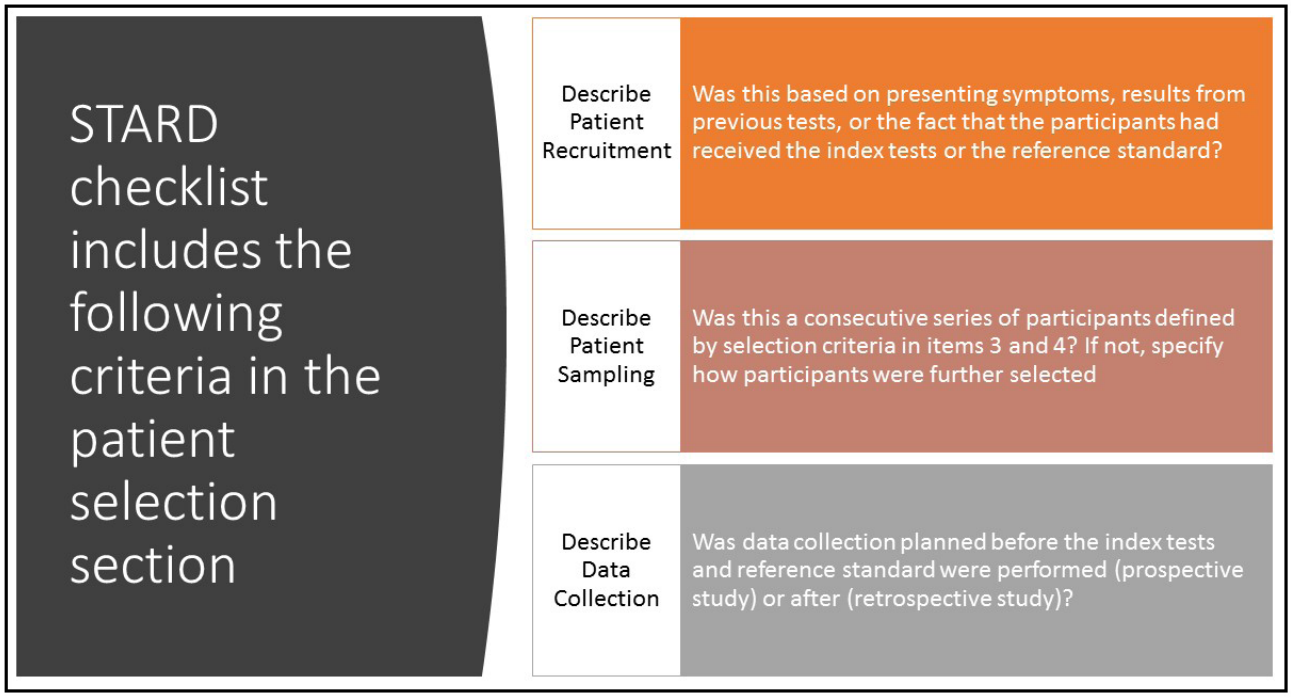

Figure 1 The STARD checklist on patient selection.[10] This checklist is a subsection of the entire STARD checklist. Within the Patient Sampling subsection, item 3 refers to, 'Describe the study population: the inclusion and exclusion criteria and the settings and locations where the data were collected' and item 4 refers to, 'Describe participant recruitment: was this based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?'.

due to PE. You recently read the American College of Emergency Physicians' clinical policies for acute PE,[5] which suggested that clinicians could use an age-adjusted D-dimer result to change the testing threshold required to exclude the diagnosis of acute PE in low-risk to intermediate-risk patients older than 50 years without missing cases of PE. Can this be applied to your patient? The answer to this question depends on the validity of the age-adjusted D-dimer studies.

The age-adjusted D-dimer for PE Study[6] is a classic example of how differential verification bias can affect the test characteristics of the index tests. The investigators recognised the potential for false positivity in older patients, and an age-adjusted D-dimer cut-off could increase the specificity for PE and safely reduce unnecessary CT imaging. The index test was the simplified, revised Geneva Score[4 7 8] or Wells Score and D-dimer test.[4 8] A D-dimer test was performed for those with low/intermediate risk, and if the D-dimer test result was negative (below the threshold), they did not receive a CT. Instead, they received follow-up at 3 months. Those with high-risk scores, or a positive D-dimer test (above the threshold) proceeded to CT pulmonary angiography (CTPA). This strategy is typical of studies validating a diagnostic strategy for PE, as it would be unethical to order CT scans for low-risk participants. As a result of this diagnostic strategy, there were two gold standards. The low-risk/negative D-dimer cohort received observation/follow-up, versus the high-risk cohort received CTPA. CTPA is more likely to identify small subsegmental PEs, which might be missed by follow-up (as they had no clinically significant symptoms) and result in different classification in disease status. In the cohort with follow-up alone (without CTPA), the sensitivity and negative predictive value of the risk scores with negative D-dimer test result will be falsely raised (due to unknown incidence of missed clinically insignificant PEs). The investigators chose to report the incidence of missed PE as the primary outcome, or 1−NPV. In the low-risk group, this is calculated by those who had PE on follow-up. Ultimately, this was a well-conducted study, and although at risk of bias, there are ethical pragmatic considerations that would prevent all participants undergoing an invasive verification test when D-dimer test results are lower than a testing threshold that would require additional imaging. In clinical practice, one should keep in mind that the rate of missed PE in low-risk patients is based on follow-up.

Incorporation bias

In studies in which disease is adjudicated by experts (including chart review), incorporation bias might affect study results. This occurs when the index test results are included in the adjudication process. Incorporation bias falsely results in elevated sensitivity and specificity.

In a recent study where high-sensitivity troponin T (hsTnT) was the index test,[9] the authors describe the gold standard determination of myocardial infarction as: 'An independent clinical events committee (CEC), made up of 2 cardiologists and one emergency physician, adjudicated the acute myocardial infarction (AMI) diagnosis for each patient per the Third Universal Definition of AMI criteria. The CEC had access to all clinical data (including the local troponin assay results) but was blinded to hsTnT… results and the local diagnosis.' If the CEC (gold standard: expert panel) was not blinded to the hsTnT results, circular reasoning would have resulted whereby

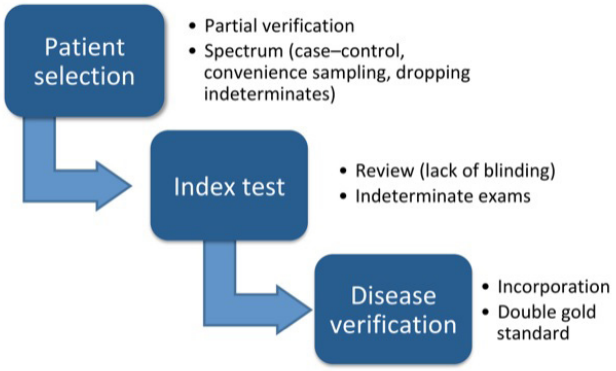

Figure 2 Recognising biases of studies of diagnostic tests. The different stages of a study of a diagnostic test are prone to certain types of biases.

the expert panel (gold standard) incorporates the index test into their final diagnosis and leads to an overestimation of the test accuracy. Fortunately, they minimised this bias by not including the results of the index test (blinding) in the criteria to establish the diagnosis of AMI. Incorporation bias should be suspected when the gold standard consists of expert medical record review, and the assessors were not blinded to the index test.

Despite their best intentions to minimise incorporation bias, two challenges arise with their blinding, (1) Unknown rise/fall of hsTnT to fulfil the universal definition of AMI. (2) Use of a different troponin assay for the same samples. The results would consequently be collinear (some element of incorporation bias still involved) and a small proportion of patients may not develop a rise/fall using another assay—underestimating the true sensitivity and specificity of the index test. These issues are challenging, highlighting the need for pragmatism when there is no perfect approach. Sensitivity analyses can potentially help evaluate the impact of any assumptions and make a study more robust.

HOW TO MITIGATE BIAS

When conducting or evaluating the study of a diagnostic test, the methodology should be considered carefully in order to mitigate potential bias. While investigators may encounter challenges unique to individual studies, they should be familiar with basic methodological principles prior to conducting a study. These principles can be classified in the same categories as we have organised the biases—patient selection, interpretation of the index test and disease verification.

Investigators should seek to include participants who are similar to those in clinical practice. Furthermore, the study cohort should represent the entire spectrum of illness, including those with severe or moderate presentations, and mild symptoms. This can be accomplished by prospectively enrolling consecutive participants from diverse study sites.

The index test should be applied to all participants in the study in a blinded fashion. If the interpreter of the test is not blinded to the results of the gold standard test, then their interpretation of the index test could be influenced. The results of the index test, similarly, should be masked from the assessors of the gold standard.

Finally, investigators should seek to apply the same method of disease verification to all participants in an independent, blinded fashion. Biases may arise if the disease verification is influenced by the index test, or if there is not a uniform, consistent application of the gold standard. Based on the clinical circumstances, this may not always be feasible.

The Standards for Reporting of Diagnostic Accuracy Studies (STARD) guidelines[10] are a checklist to improve the reporting of methods in studies of diagnostic tests (figure 1). In this way, they help the consumers of medical literature judge the risk of bias by increasing the transparency of how the study was conducted. Furthermore, by understanding STARD guidelines in the planning of their studies, investigators may develop research plans and conduct studies in such a way so as to mitigate sources of bias.

SUMMARY

Multiple pitfalls can occur with the conduct and analysis of a study of diagnostic tests. Figure 2 illustrates the three stages: patient selection, interpretation of the index test and disease verification. In this second part, we focused on (1) Interpretation bias (or workup bias): where the classification of an indeterminate index test result can bias the accuracy of a test or how a priori information can bias a subjective test result. (2) Disease verification bias: where the index test result is incorporated into the gold standard or when the gold standard is applied to only a select population. In an example with age-adjusted D-dimer for PE, differential bias was a limitation due to the use of a double gold standard; however, there are times when certain biases are an acceptable limitation of a study. Nevertheless, the informed reader can better judge the quality of a study by recognising the potential biases and limitations by being methodical in their approach to understanding the methods, and in turn, better apply studies of diagnostic tests into their clinical practice.

Correction notice This article has been corrected since it was published Online First. The title of the companion article was updated to *Recognising Bias in Studies of Diagnostic Tests Part 1: Patient Selection* in table 1 and in the references.

Contributors BK, MKH and RW generated the content together. BK wrote the review article, KH edited the article and drafted tables, and RW edited the article and drafted figures and table. All authors have read and approved the manuscript.

Funding BK is supported by NHLBI K08 (grant # 1K08HL140105-01).

Competing interests BK is the site investigator for Ortho-Clinical Diagnostics.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Hall K. Recognising Bias in Studies of Diagnostic Tests Part 1: Patient Selection. *Emerg Med J* 2019;36.
2 Frazee BW, Snoey ER, Levitt A. Emergency Department compression ultrasound to diagnose proximal deep vein thrombosis. *J Emerg Med* 2001;20:107–12.
3 Kohn MA, Carpenter CR, Newman TB. Understanding the direction of bias in studies of diagnostic test accuracy. *Acad Emerg Med* 2013;20:1194–206.
4 Wells PS, Anderson DR, Rodger M, *et al*. Excluding pulmonary embolism at the bedside without diagnostic imaging: management of patients with suspected pulmonary embolism presenting to the emergency department by using a simple clinical model and d-dimer. *Ann Intern Med* 2001;135:98–107.
5 Wolf SJ, Hahn SA, Nentwich LM, *et al*. Clinical Policy: Critical Issues in the Evaluation and Management of Adult Patients Presenting to the Emergency Department With Suspected Acute Venous Thromboembolic Disease. *Ann Emerg Med* 2018;71:e59–e109.
6 Righini M, Van Es J, Den Exter PL, *et al*. Age-adjusted D-dimer cutoff levels to rule out pulmonary embolism: the ADJUST-PE study. *JAMA* 2014;311:1117–24.
7 Le Gal G, Righini M, Roy PM, *et al*. Prediction of pulmonary embolism in the emergency department: the revised Geneva score. *Ann Intern Med* 2006;144:165–71.
8 Wells PS, Anderson DR, Rodger M, *et al*. Derivation of a simple clinical model to categorize patients probability of pulmonary embolism: increasing the models utility with the SimpliRED D-dimer. *Thromb Haemost* 2000;83:416–20.

9 Peacock WF, Baumann BM, Bruton D, *et al*. Efficacy of High-Sensitivity Troponin T in Identifying Very-Low-Risk Patients With Possible Acute Coronary Syndrome. *JAMA Cardiol* 2018;3:104–11.

10 Cohen JF, Korevaar DA, Altman DG, *et al*. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open* 2016;6:e012799.