Journal of Clinical Epidemiology 78 (2016) 73–82

Journal of Clinical Epidemiology

ORIGINAL ARTICLES

Partial verification bias and incorporation bias affected accuracy estimates of diagnostic studies for biomarkers that were part of an existing composite gold standard

Annika Karch[a], Armin Koch[a], Antonia Zapf[b], Inga Zerr[c], André Karch[c,d,e,*]

[a]*Institute for Biostatistics, Hannover Medical School, Carl-Neuberg-Str. 1, 30625 Hannover, Germany*
[b]*Department of Medical Statistics, University Göttingen, Humboldtallee 32, 37073 Göttingen, Germany*
[c]*National Reference Centre for TSE, Department for Neurology, University Göttingen, Robert-Koch-Str. 40, 37075 Göttingen, Germany*
[d]*Department of Epidemiology, Helmholtz Centre for Infection Research, Inhoffenstr. 7, 38124 Braunschweig, Germany*
[e]*German Center for Infection Research, Hannover-Braunschweig Site, Feodor-Lynen-Str. 7, 30625 Hannover, Germany*

Accepted 30 March 2016; Published online 21 April 2016

Abstract

Objective: To investigate how choice of gold standard biases estimates of sensitivity and specificity in studies reassessing the diagnostic accuracy of biomarkers that are already part of a lifetime composite gold standard (CGS).

Study Design and Setting: We performed a simulation study based on the real-life example of the biomarker ''protein 14-3-3'' used for diagnosing Creutzfeldt–Jakob disease. Three different types of gold standard were compared: perfect gold standard ''autopsy'' (available in a small fraction only; prone to partial verification bias), lifetime CGS (including the biomarker under investigation; prone to incorporation bias), and ''best available'' gold standard (autopsy if available, otherwise CGS).

Results: Sensitivity was unbiased when comparing 14-3-3 with autopsy but overestimated when using CGS or ''best available'' gold standard. Specificity of 14-3-3 was underestimated in scenarios comparing 14-3-3 with autopsy (up to 24%). In contrast, overestimation (up to 20%) was observed for specificity compared with CGS; this could be reduced to 0–10% when using the ''best available'' gold standard.

Conclusion: Choice of gold standard affects considerably estimates of diagnostic accuracy. Using the ''best available'' gold standard (autopsy where available, otherwise CGS) leads to valid estimates of specificity, whereas sensitivity is estimated best when tested against autopsy alone.

Keywords: Diagnostic validity; Incorporation bias; Partial verification bias; Creutzfeldt–Jakob disease; 14-3-3; Autopsy

1. Introduction

1.1. Bias in diagnostic studies

Diagnostic studies are prone to various types of bias, which can affect accuracy estimates if not taken into account during the design and analysis stage [1]. The choice of gold standard is therefore of particular importance, as perfect gold standards often do not exist or at least are not available for the entire study population. If only a subset of the patients' diagnoses can be verified by the gold standard and only these patients are included in the analysis, so-called partial verification bias might occur whenever verification is dependent on the result of the diagnostic test under evaluation [1,2]. In this case, specificity of the test is underestimated as people with a negative test and a negative gold standard (true negatives) are less likely to be verified than those with a positive test and a negative gold standard (false positives) [1–3]. Partial verification bias typically occurs when the gold standard is invasive (e.g., biopsy or autopsy) or harmful (e.g., computed tomography scans), and the test under evaluation has already been implemented in clinical practice. As a potential solution, alternative diagnostic gold standards, which are less valid than the ''perfect'' gold standard but available for all patients, can be used. These gold standards are often composed of several individual tests. Use of composite gold standards can, however, also lead to biased estimates of test accuracy if, contrary to the recommendation of the guideline on the clinical evaluation of diagnostic agents [4],

Conflicts of interest: Ann.K. reports no disclosures. Arm.K. reports no disclosures. A.Z. reports no disclosures. I.Z. reports no disclosures. And.K. reports no disclosures.

* Corresponding author. Tel.: +4953161813113; fax: +495115324295.
E-mail address: andre.karch@helmholtz-hzi.de (A. Karch).

http://dx.doi.org/10.1016/j.jclinepi.2016.03.022
0895-4356/

What is new?

Key findings

- We showed that, in studies reassessing the diagnostic accuracy of already established biomarkers, use of the ''best available'' gold standard (autopsy where available, otherwise CGS) leads to valid estimates of specificity, whereas sensitivity is estimated best when tested against autopsy alone.

What this adds to what was known?

- We assessed for the first time how choice of gold standard biases estimates of diagnostic accuracy in the reassessment of biomarkers that are already part of a lifetime composite gold standard (CGS) and identified a new type of partial verification bias (which we call ''discordant partial verification bias'').

What is the implication and what should change now?

- Future studies need to take our results into account and should follow our recommendations for study design to prevent overestimation as well as underestimation of diagnostic accuracy.

the test under evaluation is already incorporated in the gold standard (so-called incorporation bias); the true sensitivity and specificity of the diagnostic test will then be overestimated [5,6].

1.2. Diagnostic studies on neurodegenerative diseases

Diagnostic studies for neurodegenerative diseases are a typical example of a situation in which choice of gold standard matters. Neuropathological examination by autopsy is the perfect gold standard for many neurodegenerative diseases (e.g., Alzheimer's disease or Creutzfeldt–Jakob disease [CJD]) but is only available postmortem and even then just in a small proportion of suspected patients. Thus, composite gold standards (CGSs), which are based on several different criteria, have been developed and established in recent decades, allowing a diagnosis during lifetime without acquisition of brain material. With new diagnostic tests available and changes in the spectrum of differential diagnoses over time, CGSs are permanently re-evaluated and the diagnostic accuracy of many individual tests is reanalyzed.

1.3. Biases in diagnostic studies on 14-3-3 and Creutzfeldt–Jakob disease

One example of this is the cerebrospinal fluid (CSF) biomarker 14-3-3, which has been part of the CGS for sporadic CJD since 1998 [7,8]. The CGS for CJD consists of three conditions and classifies a patient as CJD positive if the patient fulfills the combination of (1) rapidly progressive dementia; (2) at least two of four clinical symptoms; (3) at least one of three diagnostic tests [9]. One of these three diagnostic tests is 14-3-3 (Fig. 1). When introduced in the CGS, diagnostic studies had indicated a very good test accuracy for 14-3-3 (sensitivity = 95%; specificity = 90–100%) [8,10].

However, new CSF biomarkers such as total tau or RT-QuIC have been proposed in the meantime as better diagnostic tests than 14-3-3 [11–13]. Moreover, differential diagnoses of CJD have changed in the last 15 years as the number of CSF test referrals has increased considerably (e.g., from 200 to 6,000 per year in the German National Reference Center for Prion diseases) [14]. Concerns have arisen that this might have led to a decrease in 14-3-3 accuracy [11]. Diagnostic studies reassessing the accuracy showed heterogeneous results [15], especially for the specificity of 14-3-3, which varied from 40% to 95% [8,11,16–18]. However, these studies differed from each other with respect to the gold standard used. The lowest specificity (40%) was reported in a US study from 2012, in which 14-3-3 was directly compared with a competitor, total tau [11]. This study was suspected to suffer from partial verification bias, as only autopsy-proven patients were included in the analyses although all patients with a clinical suspicion were tested for 14-3-3 and total tau [19]. As only 14-3-3 but not tau results were reported to the patients' physicians and families, decision on autopsy was directly dependent on 14-3-3, but not on tau. The exclusion of 14-3-3-negative patients who were correctly classified as nondiseased biased specificity down [1,3,5]. The example of 14-3-3 and CJD is, however, not classical for partial verification bias and differs from cases reported in the literature, as 14-3-3 is embedded in a battery of clinical and diagnostic criteria. As 14-3-3 is the best single diagnostic test and part

I. clinical signs	1. myoclonus 2. cerebellar or visual 3. pyramidal or extrapyramidal 4. akinetic mutism	
II. diagnostic tests	1. MRI	High signal abnormalities in caudate nucleus and putamen or at least in two cortical regions (temporal-parietal-occipital) either in DWI or FLAIR
	2. CSF	Detection of protein14-3-3
	3. EEG	Periodic slow wave complexes
Imperfect lifetime composite gold standard (CGS)	**rapid progressive dementia + two out of I + at least one out of II**	
Perfect autopsy gold standard	**neuropathological confirmation (detection of prion protein scrapie by immunocytochemistry/western blot)**	

Fig. 1. Available gold standards for sporadic Creutzfeldt–Jakob disease: Definition of perfect autopsy gold standard (dark gray) and imperfect lifetime gold standard (light gray), which is a composite of three factors (white) [7,9]. CSF, cerebrospinal fluid; EEG, electroencephalography; MRI, magnetic resonance imaging.