Partial verification bias and incorporation bias affect diagnostic studies for biomarkers that are part of an existing composite gold standard

Annika Karch[1]	karch.annika@mh-hannover.de
Armin Koch[1]	koch.armin@mh-hannover.de
Antonia Zapf[2]	antonia.zapf@med.uni-goettingen.de
Inga Zerr[3]	ingazerr@med.uni-goettingen.de
André Karch[3,4,5*]	andre.karch@helmholtz-hzi.de

[1] Institute for Biostatistics, Hannover Medical School, Carl-Neuberg-Str. 1, 30625 Hannover, Germany;

[2] Department of Medical Statistics, University Göttingen, Humboldtallee 32, 37073 Göttingen, Germany;

[3] National Reference Centre for TSE, Department for Neurology, University Göttingen, Robert-Koch-Str. 40, 37075 Göttingen, Germany;

[4] Department of Epidemiology, Helmholtz Centre for Infection Research, Inhoffenstr. 7, 38124 Braunschweig, Germany;

[5] German Center for Infection Research, Hannover-Braunschweig site, Germany

*Corresponding Author:

André Karch
Research Group Epidemiological and Statistical Methods (ESME)
Helmholtz Centre for Infection Research
Inhoffenstr. 7
38124 Braunschweig, Germany
Phone: +4953161813113
Fax: +495115324295
andre.karch@helmholtz-hzi.de

Counts:

Characters Title: 123 (20 words)

Words Abstract: 200

Words Paper: 5,128

Number of references: 31

Number of tables and figures: 6

Number of supplements (web-only files): 3 (1 Table, 5 Figures, 1 Table)

Disclosure of interests:

Annika Karch reports no disclosures. Dr Koch reports no disclosures. Dr Zapf reports no disclosures. Dr Zerr reports no disclosures. Dr André Karch reports no disclosures.

Contributions

AnnK and AndK designed and conceptualized the simulation study, AnnK performed the simulations and the statistical analysis of the data, AnnK, AndK, AK, AZ and IZ interpreted the data, AnnK drafted the manuscript, AndK, AK, AZ and IZ revised the manuscript. All authors approved the manuscript.

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

Running title:

Competing biases in diagnostic studies for already established biomarkers

What is new:

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

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 (CT) 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], 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 post-mortem and even then just in a small proportion of suspected patients. Thus, composite gold standards (CGS), 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, CGS are permanently re-evaluated and the diagnostic accuracy of many individual tests is reanalysed.

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 fulfils the combination of (a) rapidly progressive dementia; (b) at least two out of four clinical symptoms; (c) at least one out of three diagnostic tests [9]. One of these three diagnostic tests is 14-3-3 (Figure 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 non-diseased 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 of the CGS, decision on autopsy is rather enforced by inconsistency between 14-3-3 and the lifetime CGS than by positive results of 14-3-3 alone. Up to now, this special form of 'discordant' partial verification bias has not been described, and it is unknown how it might affect sensitivity and specificity estimates. In the majority of other studies reassessing the diagnostic accuracy of 14-3-3, lifetime CGS was used as the diagnostic gold standard, potentially overestimating 14-3-3 accuracy due to incorporation bias.

1.4 Aims of this simulation study

We aimed to assess the quantity of bias introduced by study design in a study setting with post-mortem examination as the diagnostic gold standard (resulting in discordant partial verification bias) and compared it with a study setting with a lifetime CGS (resulting in incorporation bias). Additionally, we investigated whether bias can be reduced when a gold standard based on the best available information ("BEST" gold standard: autopsy where available, CGS otherwise) is used.

2. Methods

In order to quantify the bias in estimating sensitivity and specificity of 14-3-3 introduced by

(a) using only autopsy results as a diagnostic gold standard and ignoring cases for which no autopsy was performed (autopsy study design),

(b) using lifetime CGS including the diagnostic test under evaluation (CGS study design) or

(c) using the best available gold standard information, i.e. autopsy where available and CGS otherwise (BEST study design),

we performed simulation studies based on realistic parameter values from diagnostic studies of 14-3-3 and CJD.

2.1 Design of simulation study

Simulations were performed under prevalence rates of 3% and 10% (similar to diagnostic studies based on surveillance data), and 50% (similar to specifically designed diagnostic studies (Supplementary Table A1)). Datasets with N=5,000 patients (annual number of referrals to the German CJD reference centre) were simulated with 10,000 simulation runs for each scenario. In a sensitivity analysis, we repeated all analyses simulating N=10,000 patients in order to assess the stability of the simulation results. Simulations were designed as within-subject comparisons of the investigated diagnostic tests, i.e. for each patient, 14-3-3 test result, CGS diagnosis and (for a certain subset) autopsy diagnosis were simulated. Data for sensitivity and specificity were simulated separately. Test results of 14-3-3 and CGS were generated using a multivariate normal distribution with shifted means according to the respective pre-set sensitivities and specificities, equal standard deviations and a fixed correlation between 14-3-3 and CGS of 0.8. The generated data were dichotomized at a threshold of 0.

2.2 Choice of parameter values

Assumptions for sensitivities and specificities of 14-3-3 (sensitivity: 0.70–0.90; specificity: 0.50–0.90) and the imperfect CGS (sensitivity: 0.90–0.98; specificity: 0.70–0.90) were based on estimates from previous diagnostic studies (Supplementary Table A1) [8,16,18]. Sensitivity and specificity of autopsy were considered to be 100%. The correlation between 14-3-3 and CGS was chosen to be 0.80 for all simulations, reflecting the dependence of CGS on 14-3-3 as found in the database of the German CJD reference centre. This takes into account that CGS diagnosis is not independent of 14-3-3 test results because 14-3-3 partly determines CGS diagnosis.

According to the a priori hypothesis that autopsy rates increase with uncertainty of diagnosis (defined as discordant results of 14-3-3 and CGS), different scenarios for the probability of a patient being autopsied were simulated (Supplementary Table A1). If both 14-3-3 and CGS were positive, an autopsy probability of 40% was assumed [14]; if both were negative, autopsy probability was set to 20%. In case the 14-3-3 and CGS results were discordant, probabilities were varied from 40% to 60%. These parameter assumptions were again based on data from the German CJD reference centre.

All possible combinations of different simulation properties were simulated, except for combinations where specificity of 14-3-3 was higher than specificity of CGS, as these scenarios were considered unrealistic. This resulted in a total of 378 scenarios. One scenario (prevalence of CJD=10%, sensitivity of 14-3-3=90%, specificity of 14-3-3=70%, sensitivity of CGS=90%, specificity of CGS=90%, autopsy probability for discordant test results=60%) was considered to be the most realistic scenario for diagnostic studies of 14-3-3 and CGS which are typically performed in tertiary care referral centres (defined as scenario S1).

2.3 Analysis of simulation study

Simulations and analyses were performed in R 3.1.2 [20]. Sensitivity and specificity of 14-3-3 were calculated as observed in the study designs against (a) autopsy, (b) CGS and (c) the best available information (BEST) as gold standard. We followed the recommendations of Burton and colleagues and reported three outcomes of our simulation studies: absolute bias, coverage and mean square error [21]. The quantity of bias – defined as the difference between the estimates observed in the simulation and the true parameter value – was determined and averaged over all simulation runs for the respective scenario. Coverage of corresponding 95% Wald confidence intervals for sensitivities and specificities (defined as the proportion of simulation runs in which the confidence interval included the true value) and mean square errors (a measure of the stability of simulation estimates) are reported in Supplementary Table C1. All results are presented as boxplots.

3. Results

3.1. Summarized results over all scenarios

When assessing all simulation scenarios together, sensitivity of 14-3-3 was revealed to be nearly unbiased in the autopsy design, but was highly biased when using CGS as the gold standard (Figure 2A). In the CGS study design, underestimation (max. –26%) as well as overestimation (max. +24%) of sensitivity occurred; however, overestimation was more common and was observed in more than 75% of all scenarios.

Specificity of 14-3-3 was underestimated in the autopsy study design by at least –7% up to –24% (Figure 2B). In contrast, overestimation up to +20% was observed for specificity compared with CGS.

The study design based on the BEST gold standard showed results similar to those obtained in the CGS study design, because results were influenced to a large extent by those patients diagnosed by CGS. Both sensitivity and specificity were therefore overestimated, however to a smaller degree than in the CGS study design. In particular, bias of specificity was reduced from up to 20% in the CGS design to 0–10% in the BEST study design (Figure 2B).

3.2. Results for the most realistic scenario (S1)

A detailed analysis of one special simulation scenario, assumed to be most realistic one, was performed (scenario S1, see methods). In this scenario, sensitivity and specificity of 14-3-3 in the CGS and BEST study design were moderately overestimated by approximately +5% (Figure 3). Bias on sensitivity was slightly higher in the BEST study design than in the CGS one, whereas it was the other way round for specificity. Sensitivity of 14-3-3 in the autopsy design was on average nearly unbiased (Figure 3A); however, a high variance could be observed, which can be attributed to the diminished sample size (only diseased patients with autopsy were taken into account). Specificity of 14-3-3 was underestimated considerably by more than 20% in the autopsy design (Figure 3B).

The overestimation of specificity (and to a smaller degree of sensitivity) in the CGS study design originated from the high number of false-positive results for 14-3-3 and CGS (n=400, Figure 4B). Owing to the correlation between 14-3-3 and CGS (r=0.8) and the low prevalence of CJD, a large number of truly non-diseased patients were classified as diseased by 14-3-3 and by the imperfect CGS. These false-positive pairs biased sensitivity but, more importantly, specificity upwards in the CGS study design as a large proportion of false-positive 14-3-3 test results no longer contributed to the estimation of specificity and moved instead from the medium grey box in Figure 4A to the black box in Figure 4B.

Although bias in the CGS study design was caused by patients being moved between cells in the contingency table, selected verification of patients' true disease status was responsible for bias in the autopsy study design. Here, specificity was underestimated considerably, as only a small proportion of the true negative results of 14-3-3 (21% of 3,150 [20% of 3,100 plus 60% of 50]=650) but a high proportion of the false-positive results of 14-3-3 were verified (54% of 1,350 [40% of 400 plus 60% of 950]=730) (Figure 4C, light grey and medium grey cells). Estimation of sensitivity, on the other hand, was unbiased as it was only based on true CJD cases, and there was no movement of patients between cells in the contingency table (Figure 4C, black cell).

In the BEST study design (Figure 4D), cell counts were predominantly influenced by CGS results and, as such, the results were similar to those in the CGS study design.

3.3. Effect of single simulation parameters

By varying the values of the different simulation parameters, we evaluated the effect of simulation parameters on the observed level of bias in the three study designs. The following results were again obtained over all simulation scenarios.

3.3.1. Bias of sensitivity in the autopsy study design

Sensitivity of 14-3-3 compared with autopsy results was almost unbiased in all simulation scenarios with a small tendency to underestimation, if the true sensitivity of 14-3-3 was lowest and if discordant results had the highest autopsy probability of 60% (Supplementary Figure B1).

3.3.2. Bias of specificity in the autopsy study design

Specificity of 14-3-3 in the autopsy study design was considerably underestimated; the level of underestimation was dependent on the differences in autopsy probabilities and the true specificity of 14-3-3. With increasing difference between autopsy probabilities (worst case: 20% for double-negative patients, 40% for double-positive patients, 60% for patients with discordant test results), the observed bias increased (Figure 5A). Moreover, the amount of bias decreased with increasing true specificities of 14-3-3 (Figure 5B). The effect of all simulation parameters on bias of specificity in the autopsy study design is illustrated in Supplementary Figure B2.

3.3.3. Bias of sensitivity in the CGS study design

The direction of bias for sensitivity in the CGS study design depended on the true specificity of 14-3-3 (Figure 5C). Although a low specificity of 50% led to overestimation of 14-3-3 sensitivity in the CGS study design, a high specificity of 90% led to underestimation. As described in detail for the example scenario S1 above, overestimation could be attributed to a large number of false-positive pairs. Owing to the high correlation between 14-3-3 and CGS and the low prevalence of CJD in most simulation scenarios, a large number of truly non-diseased patients were classified as diseased by 14-3-3 and by the imperfect gold standard CGS. These false-positive results biased sensitivity estimates upwards. This was mainly triggered by the true specificity of 14-3-3 but was also influenced by the true sensitivity of 14-3-3 (Supplementary Figure B3 (A)). In contrast, underestimation of 14-3-3 sensitivity in the CGS study design occurred if the true specificity of 14-3-3 was high (90%). The number of false-positive pairs incorrectly increasing sensitivity was reduced under such simulation properties. Simultaneously, there was an increasing number of non-diseased patients with positive CGS who were (correctly) classified as non-diseased by 14-3-3, leading to lower estimated sensitivity in the CGS design. The effect of all simulation parameters on bias of sensitivity in the CGS study design is illustrated in Supplementary Figure B3.

3.3.4. Bias of specificity in the CGS study design

Specificity of 14-3-3 in the CGS study design was generally overestimated because of incorporation bias; the amount of overestimation depended predominantly on the true specificity of CGS (Figure 5D). A low true specificity of CGS led to larger overestimation of simulated 14-3-3 specificity (around +18%) than a high specificity of CGS (around +8% for true CGS specificity of 90%). This was because 14-3-3 false-positive patients were less likely to be classified as CJD positive by the imperfect gold standard CGS with increasing specificity of CGS. The effect of all simulation parameters on bias of specificity in the CGS study design is illustrated in Supplementary Figure B4.

3.3.5. Bias of sensitivity in the BEST study design

Bias of sensitivity of 14-3-3 in the BEST study design was influenced by the same mechanisms as described for the CGS study design. Overestimation depended on the true specificity (Supplementary Figure B5 (B)) and true sensitivity (Supplementary Figure B5 (A)) of 14-3-3. The effect of all simulation parameters on bias of sensitivity in the BEST study design is illustrated in Supplementary Figure B5.

3.3.6. Bias of specificity in the BEST study design

Similarly to the CGS study design, the amount of bias was mainly influenced by specificity of CGS with larger overestimation for low CGS specificity (Supplementary Figure B6 (D)). The effect of all simulation parameters on bias of specificity in the BEST study design is illustrated in Supplementary Figure B6.

3.4. Overall results with respect to coverage and mean square error

Coverage rates were low in all simulated scenarios due to the high levels of absolute bias and the large sample size leading to narrow confidence intervals (Supplementary Table C1). For the situation with lowest absolute bias (estimation of sensitivity in the autopsy study design), the coverage was still only 74% instead of the targeted 95%. Mean square errors were highest when estimating specificity (0.029) and lowest when estimating sensitivity (0.001) in the autopsy study design, corresponding to the situation with largest and smallest absolute bias. Variation across scenarios was smaller using the BEST study design compared with the CGS study design (e.g. interquartile ranges were much smaller in all panels of Supplementary Figure B5 compared with Supplementary Figure B3).

Analyses of the repeated simulation with 10,000 instead of 5,000 patients revealed virtually unchanged results regarding bias, mean square error and coverage.

4. Discussion and conclusion

We investigated how different forms of bias affect measures of validity in the reassessment of already established biomarkers. We showed that specificity was considerably underestimated when compared with autopsy because of discordant partial verification bias, whereas sensitivity was almost unbiased. In contrast, sensitivity and specificity of 14-3-3 were generally overestimated in a study design that used CGS as a gold standard due to incorporation bias. If the best available information was used as the gold standard, i.e. autopsy where available and CGS otherwise, sensitivity and specificity were still overestimated but to a smaller degree than in the CGS study design (particularly for specificity).

The reassessment of diagnostic tests that are already part of a lifetime CGS differs from previous examples of partial verification bias, as decisions on disease verification do not depend on the test result itself but on the concordance of the test result with the remaining parts of the CGS. A similar form of bias has been described for cervical cancer screening, in which the results of two index tests have an effect on verification probabilities [22]. However, the cervical cancer example differs from the form of bias presented in our study, as verification probabilities in cervical cancer screening are usually highest if both tests are positive and lowest if both are negative. Verification thus depends on the sum of the test results, not on the discordance between the index test and the CGS as in our example. In specific cases, verification in cervical cancer screening studies might only be performed on discordant results while concordant results are never verified [23]. This represents the most extreme form of discordant verification bias and has been described methodologically by Miller before [24]. It differs from the form of bias presented in our study by assuming that concordant test results are always and discordant test results are never verified, while in our study verification probabilities are modified by the level of discordance between 14-3-3 and the imperfect gold standard. In order to reduce bias, studies from the area of cervical cancer screening recommend either to verify all patients in the study or (if this is too expensive or invasive) to verify a random sample of patients with concordant results. This can easily be done in the context of cervical cancer screening as verification is directly influenced by the treating physicians and researchers whereas in the example of 14-3-3, decision on verification is primarily made by legal next of kins in a much more complex setting.

Results for the analysis of sensitivity and specificity against autopsy as the diagnostic gold standard only partially met the expectations from previous studies about partial verification bias. For classic partial verification bias, it had been shown that sensitivity is raised and specificity is lowered [1,3,5]. Our simulation studies suggest no bias for sensitivity against autopsy. That means overestimation of sensitivity as observed in situations with classic partial verification bias does not occur in the special situation of discordant partial verification bias as present in biomarkers for neurodegenerative diseases. Our investigations, however, confirm a large underestimation of specificity, especially for low true specificities of 14-3-3 in the event of highly differing autopsy rates within the study population (negative/positive/discordant: 20/40/60%). However, even in the case of less diverse autopsy probabilities (20/40/40%), underestimation of around –15% could be observed. This means that the previously described form of partial verification bias – mainly introduced by test-positive selection – is present in diagnostic studies for neurodegenerative disorders as well, but is amplified by an additional verification bias caused by test discordance of the components of a composite gold standard. These two different forms of partial verification bias add up their effects and result in an even stronger underestimation of specificity.

This is of particular interest as, in the diagnostic setting of CJD, particular emphasis is placed on specificity and not on sensitivity. The setting is characterized by critically ill patients with an unknown, probably life-threatening disease and by the fact that an effective therapy for CJD is non-existent. As a consequence, diagnostic tools for CJD seek high specificity in order to avoid false-positive patients who could have been treated effectively for their true differential diagnosis if identified in time.

Given the current literature, overestimation of sensitivity and specificity against CGS (which depends on 14-3-3), as observed in our simulation studies, was expected because of incorporation bias [5,6]. Surprisingly, certain scenarios also resulted in underestimation of sensitivity, particularly when the true specificities of 14-3-3 and CGS were at the highest level and the number of false-positive pairs was small. In scenarios considered to be less extreme, only overestimation occurred.

None of the three proposed diagnostic gold standards was able to provide unbiased estimates for both accuracy measures, as illustrated in Table 1. Use of a composite gold standard (with the index test being included) leads to moderate overestimation of sensitivity and specificity. Using autopsy as the gold standard enables a correct estimation of sensitivity, but leads to unacceptable underestimation of specificity. By using a combination of both gold standards (as proposed in the BEST criteria), sensitivity is still moderately overestimated whereas specificity is only marginally biased.

One solution for this problem might be the application of bias correction methods at the analysis stage of diagnostic studies. However, publications dealing with incorporation bias do not investigate any mathematical correction methods to overcome the bias. Instead, they suggest either excluding the index test from the CGS prior to the study [25] or excluding a defined subset of patients from analysis

[26]. However, this is rarely applicable in studies reassessing already established biomarkers because diagnostic processes follow an established algorithm of combining test results obtained in clinical practice and cannot be constricted for research purposes.

In contrast, several methods have been proposed for the correction of biased specificity resulting from partial verification bias in the autopsy study design; this includes adjustment of accuracy measures by the Begg & Greenes method [27] or correction based on multiple imputation as developed by Zhou et al. [28]. Both concepts have been discussed and compared with each other [29–31], with the conclusion that the use of multiple imputation methods should be recommended, especially in situations in which the mechanism of missing reference data is not known and partial verification cannot be avoided. Interestingly, a more recent work by Xue et al. [22] described, against a background of partial verification bias in cervical cancer screening studies, that these correction methods are almost never applied in real life. Methods are either not applicable or might be too complex to be understood and implemented into statistical analysis. The authors propose an easier, intuitive correction method based on the introduction of balancing weights for subgroups with low verification probability (using weights equal to the inverse of the verification fraction).

However, all these correction methods need to be based on clear assumptions about true test validity, which are rarely available in a clinical context. Moreover, the balancing weights method requires that the diagnostic verification depends only on screening test results and not on any other variables (equivalent to a missing at random assumption). This assumption has to be checked carefully. For the example of 14-3-3 in CJD, the missing at random assumption cannot be fulfilled in real-life scenarios, as e.g. younger patients are more likely to be autopsied and at the same time have a higher probability of being CJD positive independently of their 14-3-3 and CGS status. Available bias correction methods therefore have only limited value for diagnostic studies on neurodegenerative diseases.

A major strength of our simulation study is the rigorous implementation of literature knowledge for all assumptions made as well as the wide range of scenarios simulated. A limitation can be seen in the fact that we only performed simulations based on assumptions for the example of 14-3-3 and CJD. In this example, a high correlation between the index test and CGS had to be assumed. Thus, no statement can be given about less influential components of CGS (e.g. with correlation of 0.5).

This is of particular importance as the problems presented in this study are not exclusive to neurodegenerative diseases, but may occur in all scenarios in which an imperfect CGS is available and only rarely verified by a perfect, mostly invasive gold standard (e.g. liver and kidney diseases, in which biopsies are the gold standard but are not performed in every case). However, diagnostic criteria for neurodegenerative diseases are often more complex (as single diagnostic tests are less accurate) than in other diseases and as the gold standard autopsy is much more difficult to obtain than other perfect gold standards. With new diagnostic approaches available for neurodegenerative diseases, biomarkers, which are already part of composite diagnostic criteria, will increasingly be reassessed for their diagnostic accuracy and compared with newly arising tests. Based on the choice of the gold standard, their accuracy will be either over- or underestimated if the potential for incorporation and verification bias is not addressed in the planning or analysis stage. Moreover, they might be replaced by new diagnostic tests that have been deemed superior based on a biased study design.

This might lead to the clinical application of inferior diagnostic tests and, ultimately, to a deterioration in patient care in diseases such as Alzheimer's disease and CJD. In the case of 14-3-3 and CJD, 14-3-3 could have been replaced by tau after a US study from 2012 that closely resembles our example scenario S1 [11]. As, on account of discordant partial verification bias, specificity was estimated to be only 40% despite a true specificity of 70%, a specificity of 50% would have been enough for tau to replace 14-3-3. This might have led to a situation in which an additional 1,000 non-CJD patients with potentially treatable conditions in Germany per year would have been considered as CJD positive (and because of that might not have received treatment for their curable disease).

Knowledge about the presence of different bias forms will lead to more elaborate study planning and – if relevant pre-information is available – the consideration of bias removal methods at the analysis stage.

If no relevant pre-information for the application of bias removal methods is available, we recommend based on our study results the choice of autopsy as the diagnostic gold standard in the evaluation of sensitivity and the choice of the best available gold standard (autopsy and, where not available, CGS) as the diagnostic gold standard for evaluation of specificity in future diagnostic studies reassessing the accuracy of already established biomarkers for neurodegenerative diseases.

5. References

[1] Whiting PF, Rutjes AWS, Westwood ME, Mallett S. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. J Clin Epidemiol 2013;66:1093–104. doi:10.1016/j.jclinepi.2013.05.014.

[2] Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, Van der Meulen JHP, et al. Empirical evidence of Design-Related Bias in Studies of Diagnostic Tests. JAMA J Am Med Assoc 1999;282:1061–6. doi:doi:10.1001/jama.282.11.1061.

[3] Panzer RJ, Suchman AL, Griner PF. Workup Bias in Prediction Research. Med Decis Mak 1987;7:115–9. doi:10.1177/0272989X8700700209.

[4] European Medicine Agency. Guideline on Clinical Evaluation of Diagnostic Agents. 2009.

[5] Kohn M, Carpenter CR, Newman TB. Understanding the direction of bias in studies of diagnostic test accuracy. Acad Emerg Med 2013;20:1194–206. doi:10.1111/acem.12255.

[6] Rutjes AWS, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PMM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ 2006;174:469–76. doi:10.1503/cmaj.050090.

[7] World Health Organization. Global Surveillance, Diagnosis and Therapy of Human Transmissible Spongiform Encephalopathies: Report of WHO Consultation. 1998.

[8] Zerr I, Bodemer M, Gefeller O, Otto M, Poser S, Wiltfang J, et al. Detection of 14-3-3 protein in the cerebrospinal fluid supports the diagnosis of Creutzfeldt-Jakob disease. Ann Neurol 1998;43:32–40. doi:10.1002/ana.410430109.

[9] Zerr I, Kallenberg K, Summers DM, Romero C, Taratuto a., Heinemann U, et al. Updated clinical diagnostic criteria for sporadic Creutzfeldt-Jakob disease. Brain 2009;132:2659–68. doi:10.1093/brain/awp191.

[10] Hsich G, Kenney K, Gibbs CJ, Lee KH, Harrington MG. The 14-3-3 brain protein in cerebrospinal fluid as a marker for transmissible spongiform encephalopathies. N Engl J Med 1996;335:924–30. doi:10.1056/NEJM199609263351303.

[11] Hamlin C, Puoti G, Berri S, Sting E, Harris C, Cohen M, et al. A comparison of tau and 14-3-3 protein in the diagnosis of Creutzfeldt-Jakob disease. Neurology 2012;79:547–52. doi:10.1212/WNL.0b013e318263565f.

[12] Cramm M, Schmitz M, Karch A, Mitrova E, Kuhn F, Schroeder B, et al. Stability and Reproducibility Underscore Utility of RT-QuIC for Diagnosis of Creutzfeldt-Jakob Disease. Mol Neurobiol 2015. doi:10.1007/s12035-015-9133-2.

[13] Mcguire LI, Peden AH, Orrú CD, Jason M, Appleford NE, Mallinson G, et al. RT-QuIC analysis of cerebrospinal fluid in sporadic Creutzfeldt- Jakob disease. Ann Neurol 2012;72:278–85. doi:10.1002/ana.23589.RT-QuIC.

[14] Robert-Koch-Institut. Epidemiologisches Bulletin Nr. 4: Creutzfeldt-Jakob-Erkrankung in den Jahren 2010 bis 2011. 2013.

[15] Muayqil T, Gronseth G, Camicioli R. Evidence-based guideline: Diagnostic accuracy of CSF 14-3-3 protein in sporadic Creutzfeldt-Jakob disease. Neurology 2012;79:1499–506. doi:10.1212/WNL.0b013e31826d5fc3.

[16] Zerr I, Pocchiari M, Collins S, Brandel JP, de Pedro Cuesta J, Knight RS, et al. Analysis of EEG and CSF 14-3-3 proteins as aids to the diagnosis of Creutzfeldt-Jakob disease. Neurology 2000;55:811–5. doi:10.1212/WNL.55.6.811.

[17] Beaudry P, Cohen P, Brandel JP, Delasnerie-Lauprêtre N, Richard S, Launay JM, et al. 14-3-3 Protein, Neuron-Specific Enolase, and S-100 Protein in Cerebrospinal Fluid of Patients with Creutzfeldt-Jakob Disease. Dement Geriatr Cogn Disord 1999;10:40–6.

[18] Collins S, Boyd A, Fletcher A, Gonzales M, McLean C a, Byron K, et al. Creutzfeldt-Jakob disease: diagnostic utility of 14-3-3 protein immunodetection in cerebrospinal fluid. J Clin Neurosci 2000;7:203–8. doi:10.1054/jocn.1999.0193.

[19] Karch A, Zerr I. Letter to: A Comparison of Tau and 14-3-3 Protein in the Diagnosis of Creutzfeldt-Jakob Disease. Neurology 2013;80:2081–2081. doi:10.1212/WNL.0b013e31829926a2.

[20] R Core Team. R: A language and environment for statistical computing. 2014.

[21] Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Stat Med 2006;25:4279–92. doi:10.1002/sim.2673.

[22] Xue X, Kim MY, Castle PE, Strickler HD. A new method to address verification bias in studies of clinical screening tests: Cervical cancer screening assays as an example. J Clin Epidemiol 2014;67:343–53. doi:10.1016/j.jclinepi.2013.09.013.

[23] Willis BH, Barton P, Pearmain P, Bryan S, Hyde C. Cervical screening programmes: can automation help? Evidence from systematic reviews, an economic analysis and a simulation modelling exercise applied to the UK. Health Technol Assess 2005;9.

[24] Miller WC. Bias in discrepant analysis: When two wrongs don’t make a right. J Clin Epidemiol 1998;51:219–31. doi:10.1016/S0895-4356(97)00264-3.

[25] Worster A, Carpenter C. Incorporation bias in studies of diagnostic tests: How to avoid being biased about bias. Can J Emerg Med 2008;10:174–5.

[26] Gupta A, Roehrborn CG. Verification and incorporation biases in studies assessing screening tests: Prostate-specific antigen as an example. Urology 2004;64:106–11. doi:10.1016/j.urology.2004.02.025.

[27] Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 1983;39:207–15.

[28] Harel O, Zhou X-H. Multiple imputation for correcting verification bias. Stat Med 2006;25:3769–86. doi:10.1002/sim.2494.

[29] Hanley JA, Dendukuri N, Begg CB. Multiple imputation for correcting verification bias: Rejoinder to Multiple imputation for correcting verification bias. Stat Med 2007;26:3046–56.

[30] De Groot JAH, Janssen KJM, Zwinderman AH, Moons KGM, Reitsma JB. Multiple imputation to correct for partial verification bias revisited. Stat Med 2008;27:5880–9. doi:10.1002/sim.3410.

[31] De Groot J a H, Janssen KJM, Zwinderman AH, Bossuyt PMM, Reitsma JB, Moons KGM. Correcting for Partial Verification Bias: A Comparison of Methods. Ann Epidemiol 2011;21:139–48. doi:10.1016/j.annepidem.2010.10.004.

Figure legends:

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

Figure 2: Summarized results over all simulation scenarios: Observed bias (displayed on the y-axis) of (A) sensitivity and (B) specificity of 14-3-3 in the three different study designs (autopsy, CGS and BEST) presented by boxplots. Overestimation is indicated by positive values on the y-axis, underestimation by negative ones.

Figure 3: Results for the most realistic scenario (S1): Simulated (A) sensitivities and (B) specificities after 10,000 simulation runs of one example scenario (with prevalence of CJD=10%, sensitivity 14-3-3=90%, specificity 14-3-3=70%, sensitivity CGS=90%, specificity CGS=90%, autopsy probability for discordant test results=60%); reference lines are displayed for the true sensitivity of 14-3-3 of 90% and for the true specificity of 14-3-3 of 70%.

Figure 4: Patient flowchart for the most realistic scenario (S1): The flowchart illustrates the flow of patients through the different study designs if prevalence of CJD is 10% and true sensitivity and specificity of 14-3-3 are 90% and 70%. Patient numbers in each box are rounded means of the frequency distribution seen in 10,000 simulation runs. The resulting diagnostic 2×2 tables for each study design are given in the right column. Numbers contributing to the diagnostic table with positive 14-3-3 and positive gold standard have a black background filling; numbers contributing to positive 14-3-3 and negative gold standard have a dark grey background filling; numbers contributing to negative 14-3-3 and positive gold standard have a medium grey background filling; and numbers contributing to negative 14-3-3 and negative gold standard have a light grey background filling. Sensitivity and specificity estimates are presented for the different study designs.

Figure 5: Effects of selected simulation parameters on bias: Bias in the estimation of sensitivity in the autopsy study design is influenced by the true specificity of 14-3-3 (A) and the autopsy probability for discordant results of 14-3-3 and CGS (B). In the CGS study design, amount and direction of bias in sensitivity are predominantly dependent on the true specificity of 14-3-3 (C), whereas bias in specificity depends on the true specificity of CGS (D). All results are displayed as boxplots with the observed bias on the y-axis. Overestimation is indicated by positive bias values, underestimation by negative ones.

Table 1: Direction and quantity of bias (visualised by direction and thickness of arrows) of sensitivity and specificity of 14-3-3 as estimated in the different study designs. Ticks indicate the study designs associated with the lowest amount of bias.

Diagnostic gold standard	Sensitivity	Specificity
Autopsy	✓	⬇
CGS	↑	↑
BEST	↑	✓