Adherence to STARD 2015 Reporting Recommendations in Pathology

By

Garth Fraga
B.A., Swarthmore College, 1989
M.D., University of Chicago, 1994

Submitted to the graduate degree program in Clinical Research and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Master of Science.

Committee Chair: Sue Min Lai, MS, MBA, PhD

Fred Plapp, MD, PhD

Rachel Vukas, MLS

Date Defended: 23 April 2020

The thesis committee for Garth Fraga certifies that this is the
approved version of the following thesis:

Adherence to STARD 2015 Reporting Recommendations in Pathology

Chair: Sue Min Lai MS, MBA, PhD

Graduate Director: Won S. Choi, MPH, PhD

Date Approved: 23 April 2020

Abstract

Diagnostic accuracy studies compare the performance of an index test against a reference standard test. The Standards for Reporting of Diagnostic Accuracy Studies (STARD) is a checklist of items that should be reported in scientific studies to improve completeness and transparency. The primary objective of this project was to calculate the frequency of STARD items reported in diagnostic accuracy studies from the 2017-2018 pathology scientific literature. Two raters independently scored 171 articles for compliance in reporting 34 STARD items. There was excellent inter-rater reliability (Cohen Kappa coefficient = .8773). The mean number of STARD recommended items reported was 15.44 ± 3.59 with a range of 4-28 out of maximum possible score of 34. Excluding not-applicable items such as test-induced adverse events, overall adherence to STARD reporting recommendations was 50%. There was substantial variation in individual item reporting, with > 75% reporting of 8/34 items and < 25% reporting of 11/34 items. Less than 10% of the articles included pre-specified hypotheses, rationale for choice of the reference standard, subgroup analyses for confounding, sample size calculations, subject flow diagrams, time intervals and/or interventions between index and reference tests, adverse events caused by testing, study registration numbers, or links to full study protocols. Significantly more items were reported in articles from journals that encouraged STARD usage in their author guidelines (16.15 vs. 14.84, P = .0165). The frequency of STARD item reporting was independent of journal impact factor, article citation count, ICMJE reporting standards endorsement, anatomic/clinical pathology disciplines, and pathology subspecialty. These findings demonstrate variable compliance in the recent pathology scientific literature with STARD 2015 reporting recommendations. Mandating authors submit completed STARD checklists in the manuscript submission process might improve compliance.

Acknowledgements

I am grateful to Drs. Won Choi and Sue Min Lai for supporting and inspiring me throughout my postgraduate training in clinical research. This project would not have been possible without the dedication and effort of my co-investigator, Dr. Keenan Hogan

This work is not supported by external funding. There are no conflicts of interest to report.

Table of Contents

Chapter 1: Introduction 1

Chapter 2: Methods 4

2.1 Case Selection 5

2.2 Data Extraction 7

2.3 Statistical Analysis 8

Chapter 3: Results 9

3.1 Attributes of the Articles in the Study 10

3.2 Compliance with STARD Reporting Items by Articles in the Study 11

3.3 Statistical Associations with STARD 2015 Compliance 15

Chapter 4: Discussion 18

Limitations 28

Considerations for Future Research 29

References 30

Appendix A: PubMed Search String 34

Appendix B. Data Collection Sheet 35

Appendix C: Journal Attributes 41

List of Figures

Figure 1. Flow chart of selected articles 6

List of Tables

Table 1. Characteristics of the Selected Articles (N = 171) 10
Table 2. STARD 2015 Item Reporting in the Selected Articles (N = 171) 12
Table 3. STARD 2015 Reporting Adherence after Excluding Non-Applicable Cases 14
Table 4. STARD 2015 Reporting Associations by Subgroup 16
Table 5. STARD 2015 Item Reporting in the Selected Journals 17
Table 6. STARD 2015 Adherence by Medical Specialty 20

Chapter 1: Introduction

The volume of published research has grown substantially in the past two decades from 291,806 manuscripts in 1995 to 1,064,266 in 2015 [1]. There is increasing recognition that most published research findings cannot be reproduced [2]. Attempts at experimental replication have failure rates between 75% and 90%. This is important because research funds are limited and should be allocated as efficiently as possible. Most of the US$100 billion expended annually on research was considered wasted in a recent analysis [3]. Irreproducibility is a leading cause of wasted research efforts.

Causes of irreproducibility include small sample sizes, lack of blinding, power analyses, and pre-specified hypotheses, suppression of negative studies, and poor quality data. Authors are expected to acknowledge these weaknesses in their manuscripts. To improve transparency in scientific reporting, consensus guidelines for reporting scientific studies have been designed. The best known is the Consolidated Standards of Reporting Trials (CONSORT) statement. CONSORT includes 25 items that should be reported in clinical trials to ensure transparency [4]. The EQUATOR (Enhancing the QUAlity and Transparency Of health Research) network has developed multiple checklists for reporting other research study designs from case reports to systematic reviews [5].

Diagnostic accuracy studies compare the performance of an index test against a reference standard test. The diagnostic accuracy of a test is not intrinsic to the assay. The choice of methodology, interpretation, setting, and subjects can influence measures of diagnostic accuracy [6]. Articles on diagnostic accuracy tests must include sufficient detail in order for readers to evaluate reported data for potential bias and generalizability to their own clinical setting. The Standards for Reporting of Diagnostic Accuracy Studies (STARD) 2015 statement is a list of items recommended by EQUATOR to enhance transparency and completeness in the scientific reporting of diagnostic accuracy studies [7]. There is 54-69% overall adherence to STARD reporting recommendations in the radiology, ophthalmology, and urology scientific literature [8-11]. Pathology is a diagnostic specialty, but information on the reliability of diagnostic accuracy studies in the pathology literature is limited. A survey of 66 articles from a Korean clinical pathology journal (Annals of Laboratory Medicine) found 37% overall adherence to STARD 2015 reporting recommendations [12]. It is unknown if this finding can be generalized to other pathology journals. Hence there is a compelling need for a systematic study of adherence to STARD 2015 reporting recommendations in the pathology literature.

The primary objective of this project is to calculate the frequency of STARD item reporting in diagnostic accuracy studies from the recent pathology scientific literature. The secondary objective is to calculate the frequency of STARD item reporting from the same data set by anatomic/clinical pathology disciplines and by pathology subspecialty subgroups.

This study will test the null hypotheses of no difference in the frequency of STARD item reporting between STARD endorsing and non-endorsing journals, International Committee of Medical Journal Editors (ICMJE) reporting endorsing and non-endorsing journals, high and low impact factor journals, anatomic and clinical pathology disciplines, frequently and infrequently cited articles, and between pathology subspecialties.

Chapter 2: Methods

This study measures adherence to STARD 2015 reporting items according to a preregistered protocol (researchregistry5286) in 171 diagnostic accuracy studies extracted from the pathology scientific literature between August 2017 and December 2018. Our Institutional Review Board exempted this study from formal ethics review because it was performed upon publically available library materials and did not include any protected health information.

2.1 Case Selection

A PubMed literature search of the top 40 pathology journals ranked by Thomson Reuters' Citation Impact Factor was performed for articles which included terms of diagnostic accuracy such as sensitivity, specificity, predictive value, or area under the curve. Assuming 50% overall adherence to STARD reporting recommendations, a sample size of 171 articles was required to estimate that proportion with good precision (95% confidence interval ± 7.5%). The time span of the search period was adjusted backward from December 2018 to obtain the desired sample size. The full search string is listed in Appendix A. Two reviewers (GF and KH) independently examined the titles and abstracts for study qualification. Those which both reviewers selected as qualified were retrieved for full text review. Inclusion criteria were English language journal of human pathology, comparison studies of an index test against a reference standard, and reporting of a measure of diagnostic accuracy (sensitivity, specificity, positive predictive value, negative predictive value, false positivity, false negativity, area under the curve, and likelihood ratio). Exclusion criteria were prognostic/predictive biomarker studies, missing measurements of diagnostic accuracy, non-human subjects, reviews, and commentaries. The flow chart of case selection is demonstrated in Figure 1.

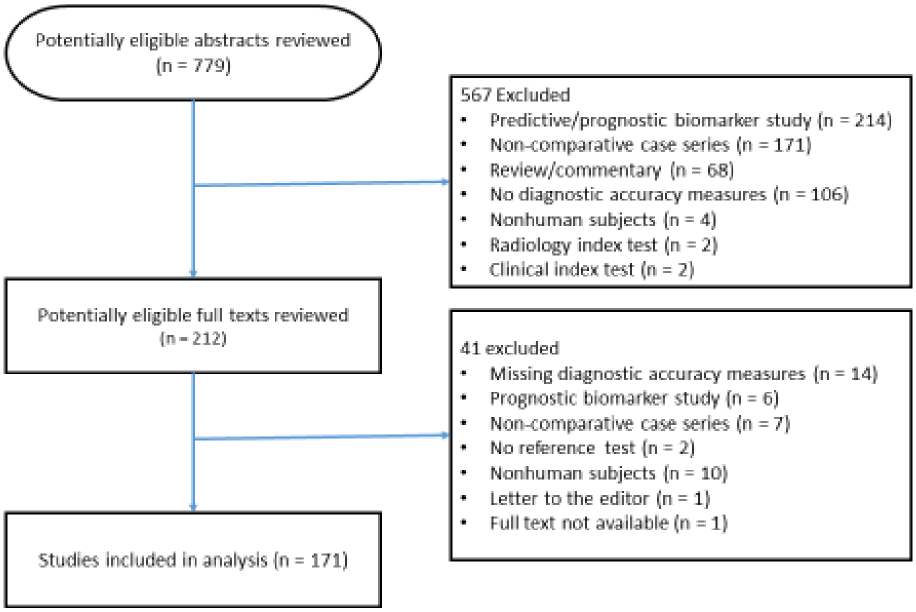

Potentially eligible abstracts reviewed (n = 779)
567 Excluded
Predictive/prognostic biomarker study (n = 214)
Non-comparative case series (n = 171)
Review/commentary (n = 68)
No diagnostic accuracy measures (n = 106)
Nonhuman subjects (n = 4)
Radiology index test (n = 2)
Clinical index test (n = 2)
Potentially eligible full texts reviewed (n = 212)
41 excluded
Missing diagnostic accuracy measures (n = 14)
Prognostic biomarker study (n = 6)
Non-comparative case series (n = 7)
No reference test (n = 2)
Nonhuman subjects (n = 10)
Letter to the editor (n = 1)
Full text not available (n = 1)
Studies included in analysis (n = 171)

Figure 1. Flow chart of selected articles

2.2 Data Extraction

Data collection sheets were designed to capture the thirty items in the STARD 2015 checklist. For items with subcomponents, each subcomponent was scored as an individual item (e.g. item 10a and item 10b in STARD 2015 on index test and reference test procedure details were each tested as individual items). Criteria for defining an item as reported, not reported, or not applicable were refined after two rounds of pilot testing small series of diagnostic accuracy studies from the pathology literature in 2016. Two reviewers (GF, KH) independently scored each article in the final study population for compliance with the final list of 34 recommended reporting items derived from STARD 2015. Study data were collected and managed with REDCap electronic data capture tools hosted at the University of Kansas Medical Center. REDcap (Research Electronic Data Capture) is a secure, web-based software platform designed to support data capture for research studies [13]. Discordant classifications were recorded and resolved in consensus meetings.

The final data collection sheet for each article in the survey included the reporting status of 34 items derived from STARD 2015, whether inter-rater classification on each item was concordant or discordant, journal title, journal impact factor, if the journal endorsed EQUATOR reporting guidelines such as STARD, if the journal endorsed the ICMJE recommendations for reporting scholarly work, and the frequency of citations by other authors. STARD and ICMJE journal endorsement status was abstracted from the journal instructions to authors and the ICMJE website [14]. Study citation frequencies and 2018 journal impact factors were obtained from Web of Science database between January and March, 2020. The final data collection sheet used in this study is included in Appendix B.

2.3 Statistical Analysis

Statistical analysis was performed with SAS 9.4 (SAS Institute, Cary, NC). Significant results were prespecified as $p < .05$. Inter-rater agreement was evaluated with Cohen's Kappa coefficient with results classified as suggested by Fleiss (0-0.39, poor; 0.40-0.75, fair to good; and 0.75-1.00, excellent) [14]. The mean number of STARD items reported in the survey articles was calculated. Shapiro-Wilks test was performed to confirm the sample was drawn from a normal probability distribution. Overall adherence to STARD reporting was expressed as a percentage by dividing the number of reported STARD items by the total number of applicable items (e.g. overall adherence was adjusted to exclude items classified as not applicable by the reviewers). Descriptive statistics were recorded as means with standard deviations for normally distributed data and as median values with interquartile ranges for skewed data.

Articles were subdivided into two or more groups by the following categorical variables: journal STARD reporting endorsement, journal ICMJE reporting endorsement, anatomic/clinical pathology study material, and pathology subspecialty. Articles were also divided into two groups by median split for the numerical variables impact factor and citation count. Two-tailed independent samples *t* test was used to detect significant differences in the number of STARD items reported between STARD endorsers/non-endorsers, ICMJE endorsing/non-endorsing journals, high/low impact factor journals, frequently/infrequently cited articles, and anatomic/clinical pathology disciplines. One-way analysis of variance (ANOVA) was used to detect significant differences in the mean number of STARD items reported in different pathology subspecialties.

Chapter 3: Results

3.1 Attributes of the Articles in the Study

171 articles were retrieved from 20 pathology journals. The frequency by journal and journal attributes are listed in Appendix C. The median journal impact factor was 2.43 (interquartile range/IQR 1.96, 4.43). The median number of article citations was 3 (IQR 1, 6). The characteristics of the articles are listed in Table 1.

Table 1. Characteristics of the Selected Articles (N = 171)

Article characteristics	**n**	**%**
Journal endorses STARD	79	46.20
Journal endorses ICMJE	86	50.29
Pathology discipline		
Anatomic pathology	120	70.18
Clinical pathology	51	29.82
Pathology subspecialty		
Surgical pathology	52	30.41
Cytopathology	43	25.15
Molecular pathology	23	13.45
Clinical chemistry	16	9.36
Hematopathology	12	7.02
Medical microbiology	11	6.43
Dermatopathology	5	2.92
Oral pathology	5	2.92
Pediatric pathology	2	1.17
Clinical informatics	2	1.17

Abbreviations: STARD, Standards for Reporting of Diagnostic Accuracy Studies; ICMJE, International Committee of Medical Journal Editors

3.2 Compliance with STARD Reporting Items by Articles in the Study

The mean number of STARD items reported was 15.44 ± 3.59 with a range of 4-28 out of a maximum possible score of 34. There was excellent agreement between the two reviewers' independent ratings of STARD item reporting with Cohen's Kappa coefficient = .8773 (95% CI .8659, .8888). There was < 90% concordance in the scoring of seven STARD items: index test cut-offs/categories (88%), reference test cut-offs/categories (83%), blinding of test readers (84%), distribution of severity of disease in subjects with the targeted condition (89%), time interval between reference and index tests (77%), cross-tabulation of index test results by reference standard (88%), and reporting sources of funding (87%). Item reporting frequencies and inter-rater concordance proportions are listed in Table 2.

Table 2. STARD 2015 Item Reporting in the Selected Articles (N = 171)

STARD 2015 item number	**Item reported, n (%)**	**Inter-rater concordance, %**
Title and/or abstract		
1. Includes a measure of diagnostic accuracy	130 (76.02)	94.74
2. Structured abstract	105 (61.40)	100.00
Introduction		
3. Scientific and clinical background of index test	171 (100)	100.00
4. Lists study hypotheses and objectives*	10 (5.85)	98.83
Methods		
5. Prospective or retrospective study design	74 (43.27)	94.15
6. Subject eligibility criteria	90 (52.63)	91.23
7. Basis for identification of potential subjects	92 (53.80)	91.23
8. Setting, location, and dates where subjects identified	104 (60.82)	97.08
9. Sequence of subject selection (e.g. random)	52 (30.41)	93.57
10a. Sufficient description of index test for replication	139 (81.29)	92.98
10b. Sufficient description of reference for replication	157 (92.35)	90.64
11. Rationale for choosing reference standard*	5 (2.92)	92.40
12a. Definition of index test cut-offs or result categories	143 (83.63)	87.72
12b. Definition of reference test cut-offs/categories	157 (91.81)	83.04
13. Blinding of test readers (index and reference test)*	38 (22.22)	84.21
14. Description of statistical methods	121 (70.76)	97.08
15. Description of handling of indeterminate results	57 (33.33)	90.06
16. Description of handling of missing results*	28 (16.37)	92.98
17. Analyses for subgroup variability/confounding*	15 (8.77)	94.74
18. Intended sample size and how it was calculated*	9 (5.26)	98.83
Results		
19. Flow diagram of subjects*	12 (7.02)	98.25
20. Baseline demographics and characteristics of subjects	92 (53.80)	94.74
21a. Distribution of severity of disease in positive subjects	82 (47.95)	89.47
21b. Alternate diagnoses in negative subjects	106 (61.98)	83.04
22. Time interval/interventions between index/ref. tests*	12 (7.02)	77.19
23. Cross tabulation of test results by reference standard	97 (56.73)	87.72
24. Estimates of precision (95% confidence intervals)	54 (31.58)	94.15
25. Adverse events caused by index or reference test*	3 (1.75)	94.15
Discussion		
26. Study limitations including bias and generalizability	67 (39.18)	91.81
27. Implications for practice, including clinical role	169 (98.83)	98.25
28. Study registration number*	4 (2.34)	99.42
29. Location of full study protocol*	4 (2.34)	98.83
30a. Disclosure of sources of funding	94 (54.97)	87.13
30b. Disclosure of conflicts of interest	148 (86.55)	97.08

Abbreviations: STARD, Standards for Reporting of Diagnostic Accuracy Studies; *Items with less than 25% reporting. Items reported denotes the number of articles that transparently reported the checklist item. Inter-rater concordance describes the proportion of articles in which raters independently agreed on the status of the checklist item.

Some STARD checklist items within articles were classified as not applicable by raters. For example, STARD item 25, adverse events due to testing, is not relevant in studies of in vitro tests. These items were excluded from the analysis to obtain the overall percent adherence to STARD 2015 reporting recommendations. The percent adherence to each recommended STARD reporting item after excluding non-applicable studies is demonstrated in Table 3. After exclusion of non-applicable STARD items, overall adherence to STARD reporting recommendations was 2,641 reported items/5,319 applicable items (50%).

Table 3. STARD 2015 Reporting Adherence after Excluding Non-Applicable Cases

STARD 2015 item number	Adherence to recommendation, %	Excluded as non-applicable, n
Title and/or abstract		
1. Includes a measure of diagnostic accuracy	76.02	0
2. Structured abstract	61.40	0
Introduction		
3. Scientific and clinical background of index test	100.00	0
4. Lists study hypotheses and objectives*	5.85	0
Methods		
5. Prospective or retrospective study design	43.27	0
6. Subject eligibility criteria	52.63	0
7. Basis for identification of potential subjects	53.80	0
8. Setting, location, and dates where subjects identified	60.82	0
9. Sequence of subject selection (e.g. random)	30.41	0
10a. Sufficient description of index test for replication	81.29	0
10b. Sufficient description of reference for replication	92.35	0
11. Rationale for choosing reference standard	33.33	156
12a. Definition of index test cut-offs or result categories	83.63	0
12b. Definition of reference test cut-offs/categories	93.45	3
13. Blinding of test readers (index and reference test)*	22.22	1
14. Description of statistical methods	70.76	0
15. Description of handling of indeterminate results	33.33	0
16. Description of handling of missing results*	16.37	0
17. Analyses for subgroup variability/confounding*	8.77	0
18. Intended sample size and how it was calculated*	5.26	0
Results		
19. Flow diagram of subjects*	7.02	0
20. Baseline demographics and characteristics of subjects	53.80	0
21a. Distribution of severity of disease in positive subjects	47.95	0
21b. Alternate diagnoses in negative subjects	70.20	20
1. Time interval/interventions between index/ref. tests*	20.34	112
2. Cross tabulation of test results by reference standard	56.73	0
3. Estimates of precision (95% confidence intervals)	31.58	0
4. Adverse events caused by index or reference test	100.00	168
Discussion		
5. Study limitations including bias and generalizability	39.18	0
6. Implications for practice, including clinical role	98.83	0
7. Study registration number*	2.34	0
8. Location of full study protocol*	2.34	0
30a. Disclosure of sources of funding	69.12	35
30b. Disclosure of conflicts of interest	86.55	0

Abbreviations: STARD, Standards for Reporting of Diagnostic Accuracy Studies; *Items with less than 25% adherence to recommendation after excluding non-applicable studies, reported as proportion. Excluded as non-applicable denotes articles in which checklist item did not apply, recorded as frequency.

3.3 Statistical Associations with STARD 2015 Compliance

Significantly more STARD items were reported in articles from journals that encouraged STARD usage in their author guidelines (16.15 vs. 14.84, P = .0165). After excluding non-applicable items, articles from journals which endorsed STARD demonstrated 52% overall adherence with STARD reporting recommendations vs. 48% overall adherence in non-endorsing journals. The frequency of STARD item reporting was independent of journal impact factor, article citation count, ICMJE reporting standards endorsement, anatomic/clinical pathology disciplines, and pathology subspecialty. The frequency of STARD item reporting was lowest in pediatric pathology (11.50) and highest in oral pathology (17.40), but there were only two and five articles in each of these subspecialties respectively, limiting comparability. These results are demonstrated in Table 4. STARD item reporting by journal is displayed in Table 5. The highest mean number of items reported was in the Journal of Oral Pathology and Medicine (17.40 ± 3.05) and the lowest mean number of items reported was in APMIS (11.67 ± 6.81). The difference in number of STARD items reported between the different journals in this survey was not significant (P=0.4768).

Table 4. STARD 2015 Reporting Associations by Subgroup

Subgroup	n (%)	STARD Items Reported, n ± s.d.	P-value
STARD endorsement status			.0165*
STARD endorser	79 (46.20)	16.15 ± 3.57	
Nonendorser	92 (53.80)	14.84 ± 3.51	
Journal impact factor			.3355*
Higher impact (≥ 2.43)	87 (50.88)	15.18 ± 3.39	
Lower impact (< 2.43)	84 (49.12)	15.71 ± 3.79	
Study citations			.9244*
Higher citations (≥ 3)	95 (55.56)	15.42 ± 3.79	
Lower citations (< 3)	76 (44.44)	15.47 ± 3.35	
ICMJE reporting recommendations			.6782*
ICMJE Endorser	86 (50.29)	15.59 ± 3.91	
Nonendorser	85 (49.71)	15.33 ± 3.26	
Pathology discipline			.7927*
Anatomic pathology	120 (70.18)	15.49 ± 3.23	
Clinical pathology	51 (29.82)	15.33 ± 4.35	
Pathology subspecialty			.0854ǂ
Clinical chemistry	16 (9.36)	16.63 ± 3.88	
Clinical informatics	2 (1.17)	17.00 ± 5.66	
Cytopathology	43 (25.15)	16.21 ± 3.17	
Dermatopathology	5 (2.92)	14.20 ± 2.86	
Hematopathology	12 (7.02)	14.83 ± 3.97	
Medical microbiology	11 (6.43)	16.18 ± 5.08	
Molecular pathology	23 (13.45)	13.70 ± 3.32	
Oral pathology	5 (2.92)	17.40 ± 3.05	
Pediatric pathology	2 (1.17)	11.50 ± 0.71	
Surgical pathology	52 (30.41)	15.23 ± 3.35	

Abbreviations: STARD, Standards for Reporting Diagnostic Accuracy Studies; ICMJE, International Committee of Medical Journal Editors; s.d., standard deviation; *t-test; ǂone-way analysis of variance test

Table 5. STARD 2015 Item Reporting in the Selected Journals

Journal	Articles Included, n (%)	2018 Impact Factor	STARD Items Reported, n ± s.d.	P Value
				0.4768*
Am J Clin Pathol	12 (7.02)	1.962	14.42 ± 4.17	
Am J Surg Pathol	15 (8.77)	6.155	13.67 ± 3.18	
Arch Pathol Lab Med	7 (4.09)	4.151	15.43 ± 3.41	
APMIS	3 (1.75)	2.225	11.67 ± 6.81	
Am J Pathol	1 (0.58)	3.762	15.00	
AIMM	3 (1.75)	1.863	14.67 ± 5.13	
Cancer Cytopathol	12 (7.02)	4.425	16.92 ± 2.27	
Diagn Cytopathol	26 (15.20)	1.402	15.58 ± 3.36	
Diagn Pathol	2 (1.17)	2.528	15.00 ± 2.83	
Dis Markers	11 (6.43)	2.761	15.56 ± 3.45	
Exp Mol Pathol	2 (1.17)	2.350	13.50 ± .71	
Histopathol	5 (2.92)	3.294	14.40 ± 3.36	
Hum Pathol	9 (5.26)	2.740	15.11 ± 2.47	
J Clin Pathol	28 (16.37)	2.346	16.89 ± 3.47	
J Cutan Pathol	5 (2.92)	1.524	15.20 ± 4.21	
J Mol Diagn	7 (4.09)	4.426	15.86 ± 6.12	
J Oral Path Med	5 (2.92)	2.030	17.40 ± 3.05	
Mod Pathol	10 (5.85)	6.365	14.50 ± 4.30	
Pathol Onc Res	3 (1.75)	2.433	14.33 ± 1.53	
Virchow Arch	5 (2.92)	2.868	16.40 ± 1.67	

*Analysis of Variance test

Chapter 4: Discussion

This survey of the recent pathology literature demonstrates variable adherence to STARD 2015 reporting recommendations for diagnostic accuracy studies. The mean number of STARD items reported among 171 recent pathology diagnostic accuracy studies was 15.44 + 3.59 with a range of 4-28 out of a maximum possible score of 34. There was broad variation in the completeness of reporting of different items. More than 90% of articles in this study included the scientific background of the index test, a description of the reference standard, reference test cut-offs/categories, and implications of the study for clinical practice. Conversely, less than 10% of the articles in this survey included pre-specified hypotheses, subgroup analyses for variability, sample size calculations, subject flow diagrams, study registration numbers, or links to full study protocols. Excluding not-applicable items such as adverse events, overall adherence to STARD reporting recommendations was 50%. This is less than in radiology (55 – 69%), ophthalmology (54%), and urology (56%). A recent survey of diagnostic tests in laboratory medicine also found lower STARD adherence in pathology (37%), although that study was limited to articles extracted from a single clinical pathology journal from outside North America. Lower STARD 2015 adherence in the pathology scientific literature may be due to the relatively lower prevalence of reporting guideline endorsement and enforcement in pathology [16]. Overall adherence to STARD 2015 recommendations in pathology compared to other medical specialties is demonstrated in Table 6.

Table 6. STARD 2015 Adherence by Medical Specialty

Field (Reference)	Sample Size (N)	Overall Adherence to STARD 2015 (%)	Items with < 25% Reporting
Radiology (8)	151	69	➢ Subgroup analysis for confounding ➢ Sample size determination* ➢ Alternate diagnoses ➢ Adverse events* ➢ Registration number* ➢ Full study protocol*
Radiology (9)	142	55	➢ Sample size determination* ➢ Adverse events* ➢ Registration number* ➢ Full study protocol*
Ophthalmology (10)	106	54	➢ Intended clinical use ➢ Rationale for reference standard ➢ Adverse events* ➢ Sample size determination* ➢ Flow chart ➢ Registration number* ➢ Full study protocol* ➢ Sources of funding
Urology (11)	61	56	➢ Handling of indeterminate results ➢ Handling of missing data ➢ Sample size determination* ➢ Flow chart ➢ Adverse events* ➢ Registration number* ➢ Full study protocol*
Laboratory Medicine (12)	66	37	➢ Study hypotheses ➢ Subject identification ➢ Blinded test readers ➢ Handling of indeterminate data ➢ Handling of missing data ➢ Subgroup analyses for confounding ➢ Sample size determination* ➢ Flow chart ➢ Diagnoses in non-target condition ➢ Time interval index-reference test ➢ Adverse events* ➢ Registration number* ➢ Full study protocol*
Pathology (current study)	171	50	➢ Study hypotheses ➢ Rationale for reference test ➢ Blinded test readers ➢ Handling of missing data ➢ Subgroup analyses for confounding ➢ Sample size determination* ➢ Flow chart ➢ Time intervals between tests ➢ Adverse events* ➢ Registration number* ➢ Link to full study protocol*

*Items with < 25% reporting in all reported studies

There were significantly more STARD items reported in articles from journals that encouraged STARD usage in their author guidelines (16.15 vs. 14.84, P = .0165). The frequency of STARD item reporting was independent of journal impact factor, article citation count, ICMJE reporting standards endorsement, anatomic/clinical pathology disciplines, and pathology subspecialty. These results are consistent with a study of adherence to biomarker reporting guidelines in the pathology literature, which found higher adherence in journals which endorsed EQUATOR reporting guidelines such as STARD [16]. These findings suggest scientific journals can improve reporting compliance by encouraging authors to follow STARD guidelines. A survey of 167 journals found endorsement of reporting guidelines in the instructions to authors increased between 2010 and 2015 [17]. However, most endorsing journals simply refer to the STARD statement instead of explicitly communicating their expectations of authors [18]. This is supported by our results. While STARD compliance was higher in endorsing journals, the overall STARD adherence difference was marginal (52% vs. 48%). Stronger enforcement of guideline compliance by requiring a completed STARD checklist in the manuscript submission process might improve overall adherence.

Studies of adherence to STARD reporting recommendations have uniformly found a lack of reporting of sample size calculations, a priori registration, links to full study protocols, and reporting of test-associated adverse events (see asterisked items in Table 4). Publically registering the study prior to data collection demonstrates that hypotheses are pre-specified and tested according to an a priori protocol. Failure to preregister studies, perform power analyses for sample size determination, and pre-specify testable hypotheses renders data susceptible to *p-hacking*. P-hacking describes the practice of selectively reporting post-hoc analytical decisions by retrospectively fitting data to identify significant results [19]. P-hacking is essentially a technique for identifying false-positive results. Articles that report false-positive results are a root cause of the current crisis in research reproduction. 167/171 (98%) of the articles in this survey were not registered prior to data collection. This suggests research into pathology diagnostic accuracy tests is predominantly exploratory and vulnerable to p-hacking. Requiring registration of study protocols prior to data collection would help address these most common deficiencies in diagnostic accuracy studies. Multiple registries welcome diagnostic accuracy studies. For example, ClinicalTrials.gov "accepts registration of all studies that 'are in conformance with applicable human subjects or ethics review regulations (or equivalent) and applicable regulations of the national (or regional) health authority (or equivalent)" [20]. Safety reporting is critical in randomized trials, but may have less relevance in diagnostic accuracy testing.

To calculate overall adherence to STARD reporting recommendations, reviewers excluded three STARD items from a majority of the articles due to lack of relevance. STARD item 25 (adverse events caused by the index or reference test) was excluded in 168/171 (98%). Pathology tests are performed in vitro and unlikely to cause direct harm. STARD item 11 (rationale for the choice of reference test) was excluded in 156/171 (91%). The reference test was predominantly histopathology, which is the most common gold standard for a broad range of diagnostic tests and represents a core element of pathology practice. STARD item 22 (time interval between index and reference test) was excluded in 112/171 (66%) of the articles, primarily because index and reference tests were performed on the same specimen (for example, immunohistochemical staining of a sample in which the reference test was histopathologic diagnosis of the same tissue sample). These items may have less relevance for reporting diagnostic accuracy test studies in pathology than in other medical specialties.

Compliance with STARD item 4 (study objectives and hypotheses) was much lower in this survey than has been reported in ophthalmology, radiology, and urology (6% vs 37%, 95-100%, and 95% respectively). This finding was consistent with the low frequency of hypothesis reporting in laboratory medicine articles (12%) and consistent with a survey of 126 diagnostic accuracy articles from high-impact journals which found 88% did not include a test hypothesis [21]. Item 4 encompasses both study objectives and hypotheses. The articles in this survey included general exploratory objectives but rarely included prespecified hypotheses. Future iterations of STARD should consider bifurcating item 4 into separate items on objectives and hypotheses to allow readers to differentiate exploratory from hypothesis-testing research. Authors should clearly state whether statistical tests of diagnostic accuracy measures were planned a priori or by post-hoc analysis.

Only 38/171 (22%) of articles explicitly described blinding the test readers to clinical information and/or reference test results. Failure to disclose blinding of test readers occurred in two contexts. In the first instance laboratory assays performed by a machine were scored as noncompliant due to the strict interpretation of the guideline adopted by the reviewers. It may be unreasonable to expect authors of studies on automated machine tests to explicitly state that such tests were performed blindly. The second instance concerned tissue based tests such as immunohistochemistry (IHC) reactions in which a pathologist examined a hematoxylincounterstained slide and subjectively scored the IHC reaction on an ordinal scale (e.g. 1+ to 4+). This latter setting has potential for verification bias. In studies where the reference test was histopathologic diagnosis, use of hematoxylin counter-stained sections with IHC testing effectively unblinds the test reader. Use of subjective ordinal scoring further exacerbates potential for verification bias. To mitigate verification bias in diagnostic accuracy IHC tests, use of tissue microarrays and automated image analysis with continuous data measurement scales should be encouraged.

Authors are expected to acknowledge the limitations of their work to help readers interpret their findings and to inform future research. Thirty-nine percent of the 171 articles in this survey reported study limitations. A survey of 30 general medical and specialty journals found 73% of 300 articles contained a median of three self-reported limitations [22]. This finding suggests reporting of research limitations in pathology is less frequent than in other medical specialties. Authors may be reluctant to acknowledge study limitations out of concern it will reduce their odds of manuscript acceptance. Expert consensus recommendations are that limitations should be included in both abstract and discussion, include all factors that may have impacted the quality of evidence, discuss the direction of any contingent bias, consider both internal and external validation, list strengths that counterbalance these limitations, and discuss implications for future research [23]. The reviewers in this study accepted any mention of a study limitation as compliant with STARD, but very few articles in this survey would meet the more stringent expert recommendations. None of the journals in this survey required dedicated subsections for limitations, unlike non-pathology journals such as the Annals of Internal Medicine or the Journal of the American Medical Association. Adopting this requirement might improve reporting of study limitations in pathology journals.

There was excellent inter-rater reliability in evaluating STARD item reporting (Cohen Kappa coefficient = 0.8773) with > 90% concordance between independent grading of 26/34 items. STARD item 22, description of the time interval and clinical interventions between index and reference tests, had the lowest inter-observer concordance (77%). Delays or clinical interventions between the index test and the reference test can bias test performance measurements. Pathology authors could acknowledge inter-test delays and interventions more explicitly to help readers recognize their potential for bias. Concordance in grading of STARD item 30a (reporting sources of funding) was less than expected for a categorical variable (87%). Industry-sponsored studies have more favorable results and conclusions than independent studies [24]. Transparent disclosure of funding sources is imperative to make readers aware of potential bias. Disagreements between raters for this item were attributed to inconsistent placement of funding source disclosures and variable reporting language. Journals should consider more visible placement of study funding statements and encourage more forthright language from authors on presence or absence of external funding.

Only 12/171 (8%) of articles included a flow chart of case selection. Flow diagrams convey the basic study design to readers. They can highlight selection bias if only a subset of eligible subjects undergo index or reference testing. Diagnostic test performance measurements depend upon the prevalence of disease in the study population, and flow diagrams allow readers to determine if the reported populations are similar to theirs [25]. Likewise, summary clinical demographics were rarely reported in tables, and cross-tabulations of index and reference test results often had to be reconstructed from the text of the results section. Articles were scored as reporting these STARD items if the results text included minimum suggested elements, but it was often very difficult to construct contingency tables of index test results by the results of reference test standards from text data. STARD recommends reporting clinical demographics in the study results, but raters found them frequently included in the methods sections of articles instead. Greater use of standardized demographic and contingency tables modeled on the examples in the STARD recommendations would improve reader comprehension and completeness of reporting.

Studies on predictive/prognostic biomarkers were the most common cause for exclusion in the case selection process (220/779 excluded). Distinction between pathology prognostic/predictive biomarkers and diagnostic tests is not always straightforward. By some definitions, biomarkers are a measurable element of diagnostic tests [26]. Some biomarkers are both prognostic and diagnostic. For example, p53 mutation can be both a marker of aggressive behavior and a diagnostic marker. The REMARK statement is a checklist for tumor prognostic biomarker studies [27]. This checklist consists of 20 items which were proposed in 2005 and remain unchanged [28]. The 20 items overlap with the 30 STARD checklist items, but also include descriptions of treatment received, clinical endpoints, a list of all candidate prognostic variables studied, and advocate reporting hazard ratios, survival probabilities, and time-to-event plots instead of diagnostic accuracy measures. Creating a pathology-specific reporting checklist that would encompass both prognostic and diagnostic biomarkers would simplify reporting and might improve compliance. However the primary outcome measures in prognostic and diagnostic biomarkers are different, and a joint checklist might have less precision.

In summary, this survey found 50% overall adherence to STARD 2015 reporting recommendations in the recently published pathology literature on diagnostic accuracy tests. There was substantial variation in individual item reporting, with > 75% reporting of 8/34 items and < 25% reporting of 11/34 items. Journal endorsement of STARD usage was associated with significantly more frequent reporting of STARD items, but the difference in overall adherence was small (52 vs. 48%). More prevalent endorsement and enforcement of STARD 2015, requirements for completed STARD checklists in the manuscript submission process, and adjustment of some STARD items to pathology-specific needs are areas for future improvement.

Limitations

The primary limitation of this study is potential information bias due to measurement error. Scoring for STARD 2015 reporting was performed by two raters who resolved disagreements by consensus. STARD 2015 items range from categorical assessments (e.g. use of a participant flow chart to demonstrate subject selection) to subjective assessments (e.g. reference test procedure in sufficient detail for reproduction). The raters adopted restrictive interpretations of some of the more ambiguous items to enhance inter-rater concordance. This produced lower STARD item reporting frequency for some items such as prespecified hypotheses and prospective vs. retrospective study designs than have been reported by other investigations of STARD 2015 compliance in the radiology, ophthalmology, and urology scientific literature. The strict interpretation of reporting items would bias our results against STARD compliance. Design bias may limit comparability with other studies. STARD 2015 includes several criteria with subitems (e.g. item 10, explanation of index and reference tests in sufficient detail to allow replication, is reported individually for each test as item 10a and 10b). We scored each subitem as a single item. This facilitated inter-rater concordance, but could overweight these items in comparison to studies which graded these items as a solitary unit.

These weaknesses are counterbalanced by certain strengths. The sample size of 171 articles is the largest survey of STARD 2015 adherence to date. The raters strict interpretation of STARD 2015 reporting recommendations was associated with high inter-rater agreement (Cohen Kappa coefficient = 0.8773).

Considerations for Future Research

While checklists such as STARD and CONSORT are known to improve the consistency of scientific reporting, it is unknown if they improve study reproduction. Further research is needed to determine if enforcement of checklist reporting will translate into higher research reproducibility, more frequent negative studies, and more frequent hypothesis-testing research. Editors may be concerned that additional demands on authors will discourage submissions. Research into the effect of checklist enforcement on the frequency of manuscript submissions is needed.

References

1. Fontelo P, Liu F. A review of recent publication trends from top publishing countries. *Syst Rev*. 2018;7:147.
2. Begley CG, Ioannidis JP. Reproducibility in science: improving the standard for basic and preclinical research. *Circ Res*. 2015;116:116-126.
3. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. *Lancet*. 2009;374:86-89.
4. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ*. 2010;340:c869.
5. EQUATOR (Enhancing the QUAlity and Transparency Of health Research) Network. Available at http://www.equator-network.org/. Accessed April 11, 2020.
6. Whiting PF, Rutjes AW, Westwood ME, et al. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. *J Clin Epidemiol*. 2013;66:1093–1104.
7. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open*. 2016;6:e012799.
8. Zarei F, Zeinali-Rafsanjani B. Assessment of adherence of diagnostic accuracy studies published in radiology journals to STARD statement indexed in Web of Science, PubMed & Scopus in 2015. *J Biomed Phys Eng*. 2018;8:311-324.
9. Michelessi M, Lucenteforte E, Miele A, et al. Diagnostic accuracy research in glaucoma is still incompletely reported: An application of Standards for Reporting of Diagnostic Accuracy Studies (STARD) 2015. *PLoS One*. 2017;12:e0189716.

10. Hong PJ, Korevaar DA, McGrath TA, et al. Reporting of imaging diagnostic accuracy studies with focus on MRI subgroup: Adherence to STARD 2015. *J Magn Reson Imaging*. 2018;47:523-544.

11. Smith DW, Gandhi S, Dahm P. The reporting quality of studies of diagnostic accuracy in the urologic literature. *World J Urol.* 2019;37:969-974.

12. Jang MA, Kim B, Lee YK. Reporting Quality of Diagnostic Accuracy Studies in Laboratory Medicine: Adherence to Standards for Reporting of Diagnostic Accuracy Studies (STARD) 2015. *Ann Lab Med.* 2020;40:245-252.

13. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: Building an international community of software platform partners. *J Biomed Inform.* 2019;95:103208.

14. ICMJE (International Committee of Medical Journal Editors) available at http://www.icmje.org/journals-following-the-icmje-recommendations/. Accessed April 11, 2020.

15. Fleiss, J.L. (1981). Statistical methods for rates and proportions (2nd ed.). New York: John Wiley.

16. Caron JE, March JK, Cohen MB, Schmidt RL. A survey of the prevalence and impact of reporting guideline endorsement in pathology journals. *Am J Clin Pathol.* 2017;148:314-322.

17. Toews I, Binder N, Wolff RF, Toprak G, von Elm E, Meerpohl JJ. Guidance in author instructions of hematology and oncology journals: A cross sectional and longitudinal study. *PLoS One.* 2017;12:e0176489.

18. Smidt N, Overbeke J, de Vet H, Bossuyt P. Endorsement of the STARD Statement by biomedical journals: survey of instructions for authors. *Clin Chem.* 2007;53:1983-1985.

19. Simonsohn U, Nelson LD, Simmons JP. P-curve: a key to the file-drawer. *J Exp Psychol Gen*. 2014;143:534-547.

20. Korevaar DA, Hooft L, Askie LM, et al. Facilitating prospective registration of diagnostic accuracy studies: A STARD Initiative. *Clin Chem.* 2017;63:1331-1341.

21. Ochodo EA, de Haan MC, Reitsma JB, Hooft L, Bossuyt PM, Leeflang MM. Overinterpretation and misreporting of diagnostic accuracy studies: evidence of "spin". *Radiology*. 2013;267:581-588.

22. Ter Riet G, Chesley P, Gross AG, et al. All that glitters isn't gold: a survey on acknowledgment of limitations in biomedical studies. *PLoS One*. 2013;8:e73623.

23. Puhan MA, Akl EA, Bryant D, Xie F, Apolone G, ter Riet G. Discussing study limitations in reports of biomedical studies - the need for more transparency. *Health Qual Life Outcomes.* 2012;10:23-26.

24. Lundh A, Lexchin J, Mintzes B, Schroll JB, Bero L. Industry sponsorship and research outcome. *Cochrane Database Syst Rev.* 2017 Feb 16;2:MR000033.

25. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med*. 1978;299:926-930.

26. Strimbu K, Tavel JA. What are biomarkers? *Curr Opin HIV AIDS*. 2010;5:463-466.

27. Altman DG, McShane LM, Sauerbrei W, Taube SE. Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK): explanation and elaboration. *PLoS Med.* 2012;9:e1001216.

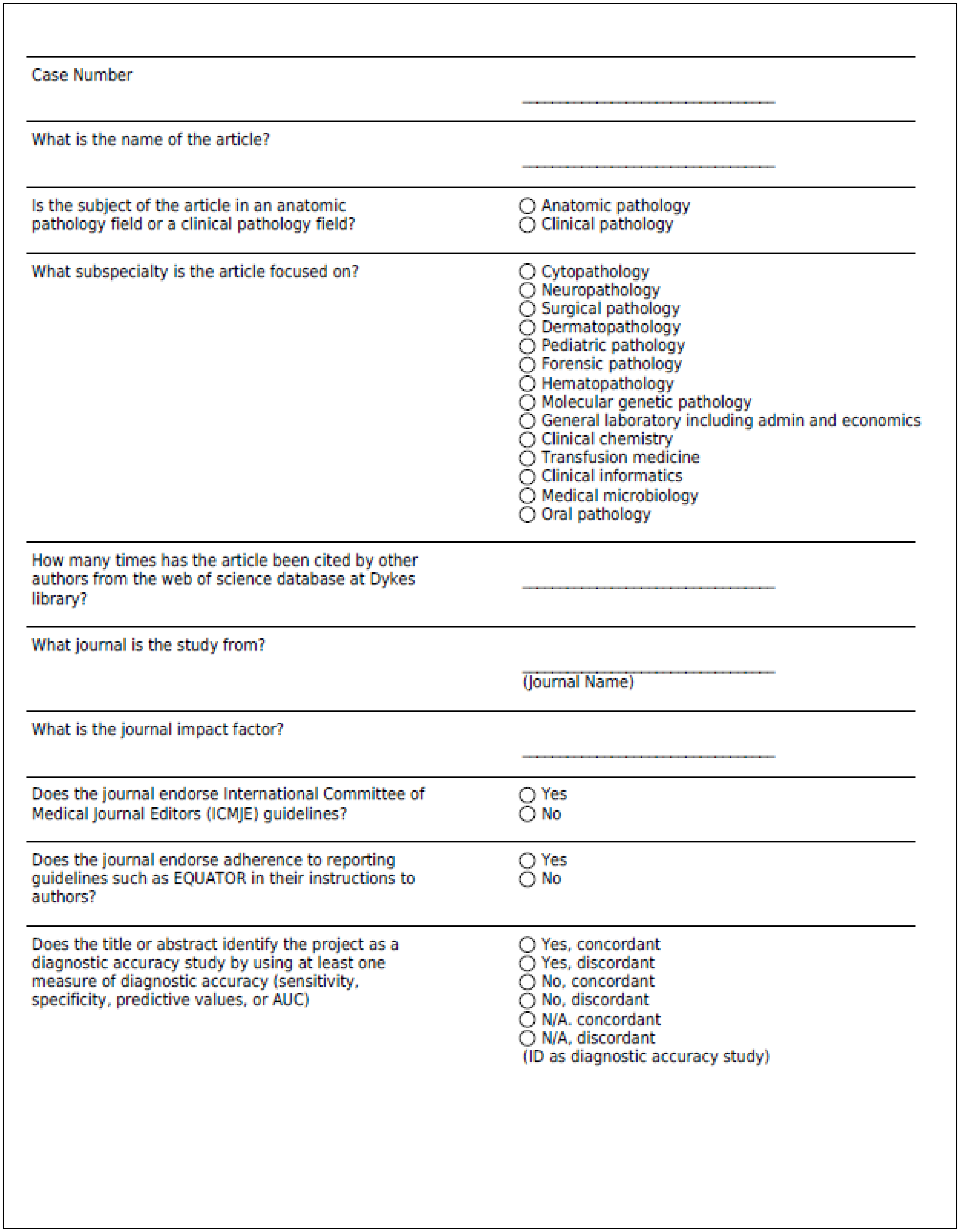
28. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM; Statistics Subcommittee of the NCI-EORTC Working Group on Cancer Diagnostics. Reporting

recommendations for tumor marker prognostic studies (REMARK). *J Natl Cancer Inst.* 2005;97:1180-4.

Appendix A: PubMed Search String

((((((((((("Acta Neuropathol"[Journal] OR "Am J Pathol"[Journal]) OR "Dis Model Mech"[Journal]) OR "Annu Rev Pathol"[Journal]) OR "Lab Invest"[Journal]) OR "J Clin Pathol"[Journal]) OR "J Mol Diagn"[Journal]) OR "J Neuropathol Exp Neurol"[Journal]) OR "Semin Immunopathol"[Journal]) OR "Brain Pathol"[Journal]) OR "Dis Markers"[Journal] OR "Diagn Pathol"[Journal] OR ("Int J Clin Exp Pathol"[jour] OR "Am J Surg Pathol"[Journal]) OR "J Pathol."[jour] OR "Mod Pathol"[Journal] OR "Hum Pathol"[Journal] OR "Histopathology"[Journal] OR "Am J Clin Pathol"[Journal] OR "Arch Pathol Lab Med"[Journal] OR "Virchows Arch."[jour] OR “Exp Mol Pathol“ [journal] OR “Neuropathol Appl Neurobiol” [journal] OR “APMIS” [JOURNAL] OR “Histol Histopathol”[Journal] OR “Alzheimer Dis Assoc Disord”[jour] OR “Cancer Cytopathol”[Jour] OR “J Cutan Pathol”[Jour] OR “J Oral Pathol Med”[Jour] OR “Pathol Oncol Res”[Jour] OR “Toxicol Pathol”[jour] OR “Appl Immunohistochem Mol Morphol”[jour] OR “Cardiovasc Pathol”[jour] OR “Diagn Cytopathol”[jour] OR “Exp Pathol”[jour] OR “Neuropathology”[jour] OR “Pathol Int”[jour]) AND (("sensitivity and specificity"[MeSH Terms] OR "probability"[MeSH Terms] OR ppv[All Fields] OR npv[All Fields] OR (false[All Fields] AND "positivity"[All Fields])) OR (false[All Fields] AND negativity[All Fields]) OR (true[All Fields] AND "positivity"[All Fields]) OR (true[All Fields] AND negativity[All Fields]) OR "Area Under Curve"[MeSH Terms] OR "odds ratio"[MeSH Terms] OR "Predictive Value of Tests"[MeSH Terms] AND ("2017/01/01"[PDAT] : "2018/12/31"[PDAT]))).

Appendix B. Data Collection Sheet



Case Number

____

What is the name of the article?

____

Is the subject of the article in an anatomic pathology field or a clinical pathology field?

○ Anatomic pathology
○ Clinical pathology

What subspecialty is the article focused on?

○ Cytopathology
○ Neuropathology
○ Surgical pathology
○ Dermatopathology
○ Pediatric pathology
○ Forensic pathology
○ Hematopathology
○ Molecular genetic pathology
○ General laboratory including admin and economics
○ Clinical chemistry
○ Transfusion medicine
○ Clinical informatics
○ Medical microbiology
○ Oral pathology

How many times has the article been cited by other authors from the web of science database at Dykes library?

____

What journal is the study from?

____
(Journal Name)

What is the journal impact factor?

____

Does the journal endorse International Committee of Medical Journal Editors (ICMJE) guidelines?

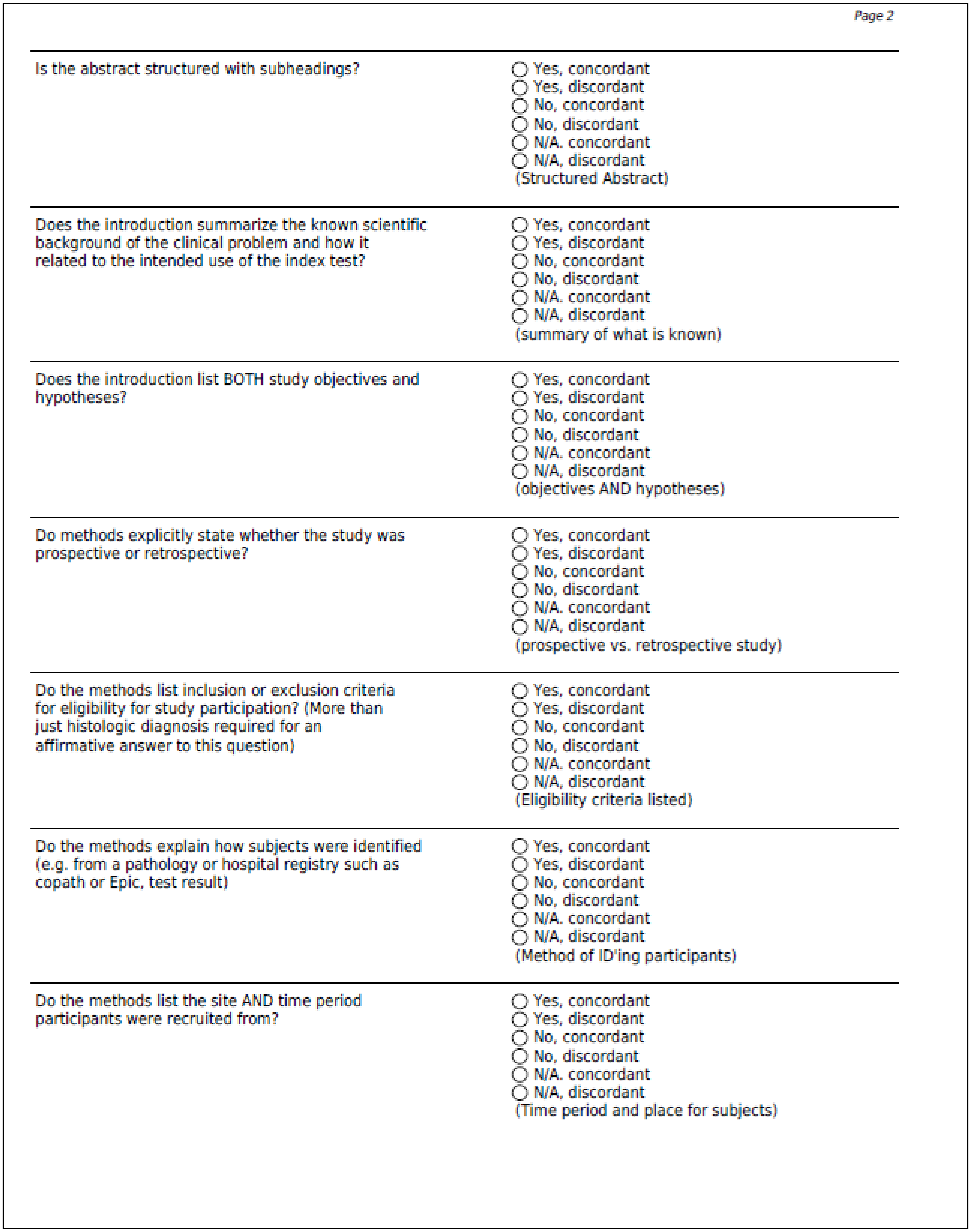
○ Yes
○ No

Does the journal endorse adherence to reporting guidelines such as EQUATOR in their instructions to authors?

○ Yes
○ No

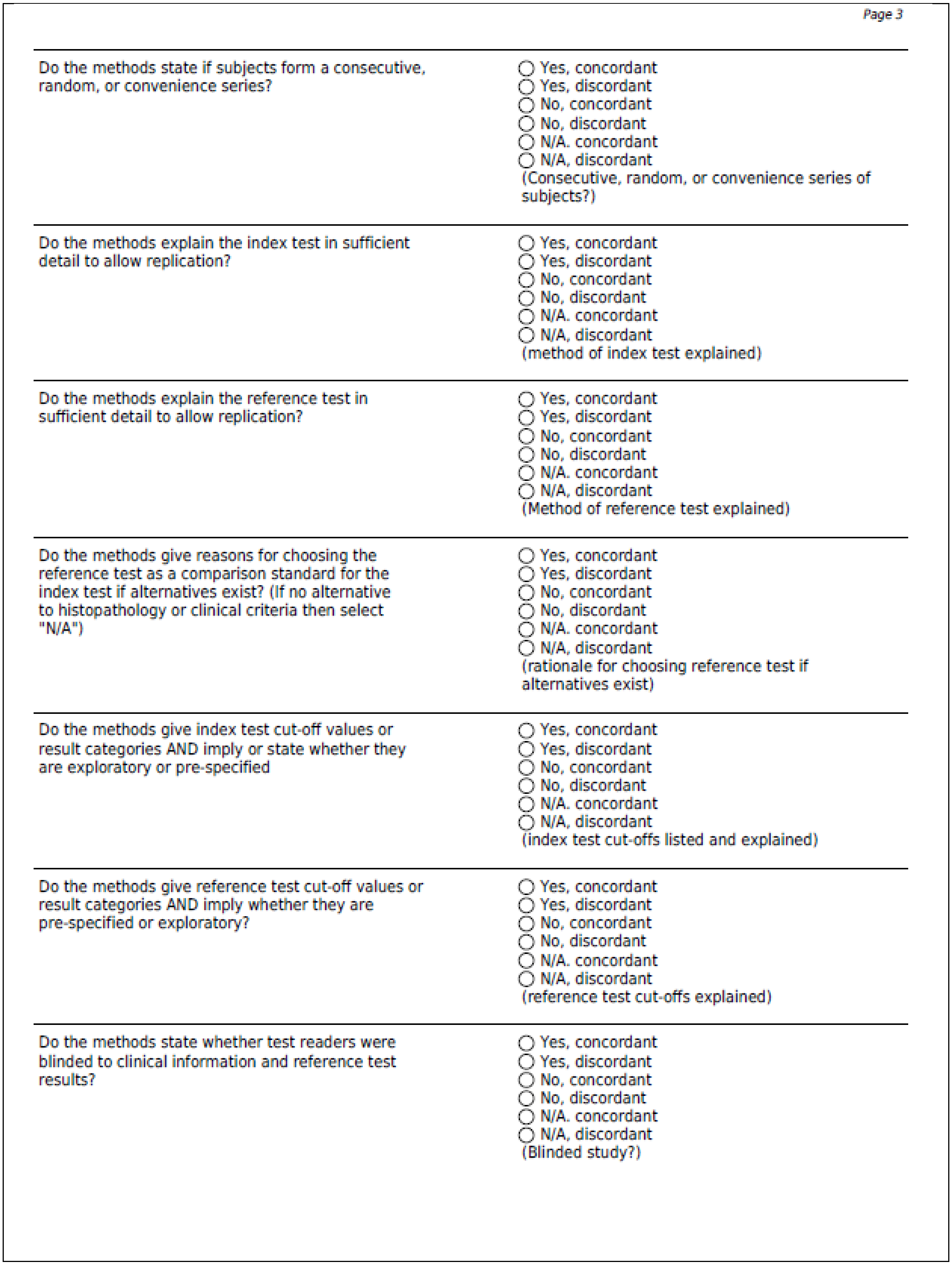
Does the title or abstract identify the project as a diagnostic accuracy study by using at least one measure of diagnostic accuracy (sensitivity, specificity, predictive values, or AUC)

○ Yes, concordant
○ Yes, discordant
○ No, concordant
○ No, discordant
○ N/A. concordant
○ N/A, discordant
(ID as diagnostic accuracy study)



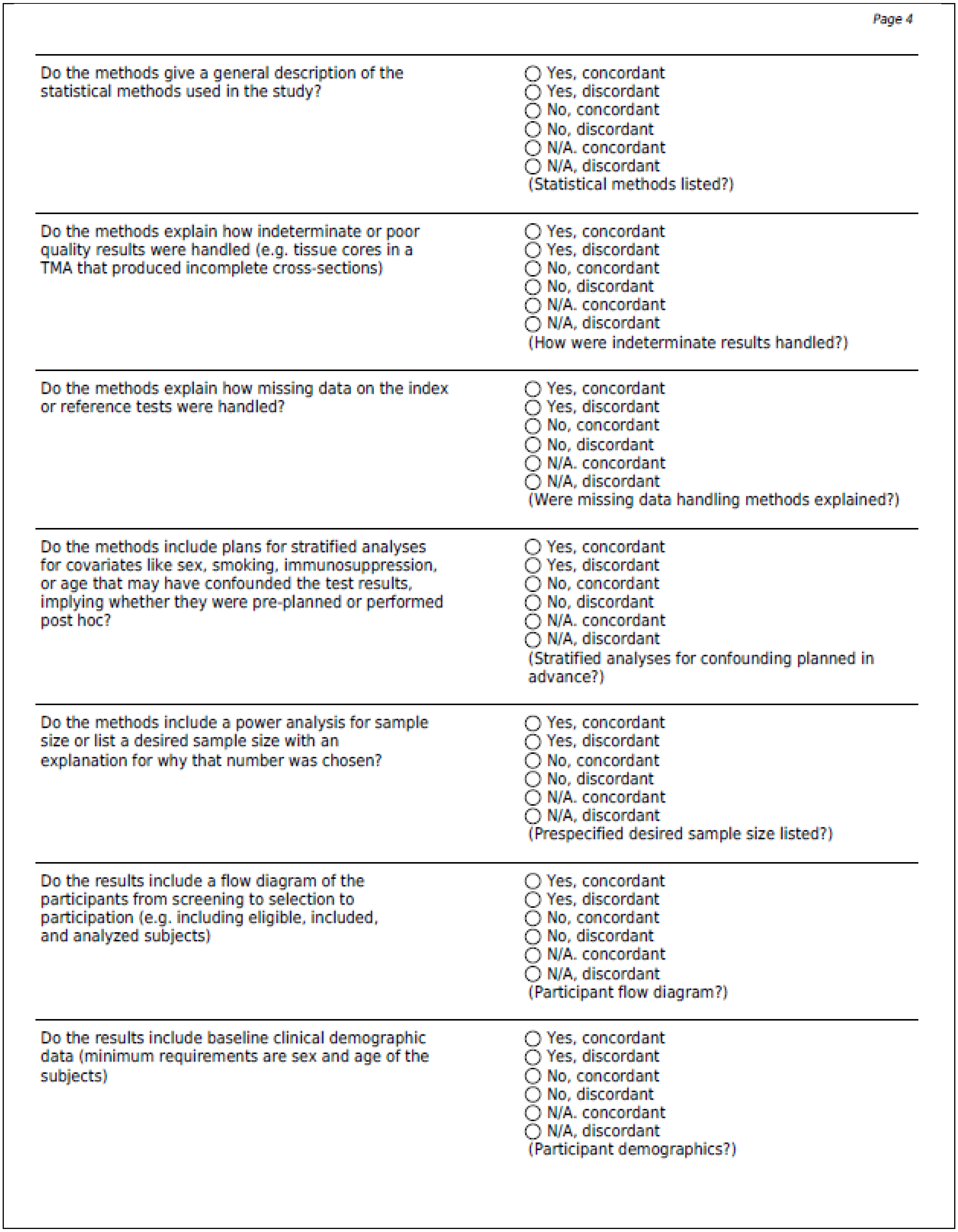
Page 2

Is the abstract structured with subheadings?	○ Yes, concordant ○ Yes, discordant ○ No, concordant ○ No, discordant ○ N/A. concordant ○ N/A, discordant (Structured Abstract)
Does the introduction summarize the known scientific background of the clinical problem and how it related to the intended use of the index test?	○ Yes, concordant ○ Yes, discordant ○ No, concordant ○ No, discordant ○ N/A. concordant ○ N/A, discordant (summary of what is known)
Does the introduction list BOTH study objectives and hypotheses?	○ Yes, concordant ○ Yes, discordant ○ No, concordant ○ No, discordant ○ N/A. concordant ○ N/A, discordant (objectives AND hypotheses)
Do methods explicitly state whether the study was prospective or retrospective?	○ Yes, concordant ○ Yes, discordant ○ No, concordant ○ No, discordant ○ N/A. concordant ○ N/A, discordant (prospective vs. retrospective study)
Do the methods list inclusion or exclusion criteria for eligibility for study participation? (More than just histologic diagnosis required for an affirmative answer to this question)	○ Yes, concordant ○ Yes, discordant ○ No, concordant ○ No, discordant ○ N/A. concordant ○ N/A, discordant (Eligibility criteria listed)
Do the methods explain how subjects were identified (e.g. from a pathology or hospital registry such as copath or Epic, test result)	○ Yes, concordant ○ Yes, discordant ○ No, concordant ○ No, discordant ○ N/A. concordant ○ N/A, discordant (Method of ID'ing participants)
Do the methods list the site AND time period participants were recruited from?	○ Yes, concordant ○ Yes, discordant ○ No, concordant ○ No, discordant ○ N/A. concordant ○ N/A, discordant (Time period and place for subjects)



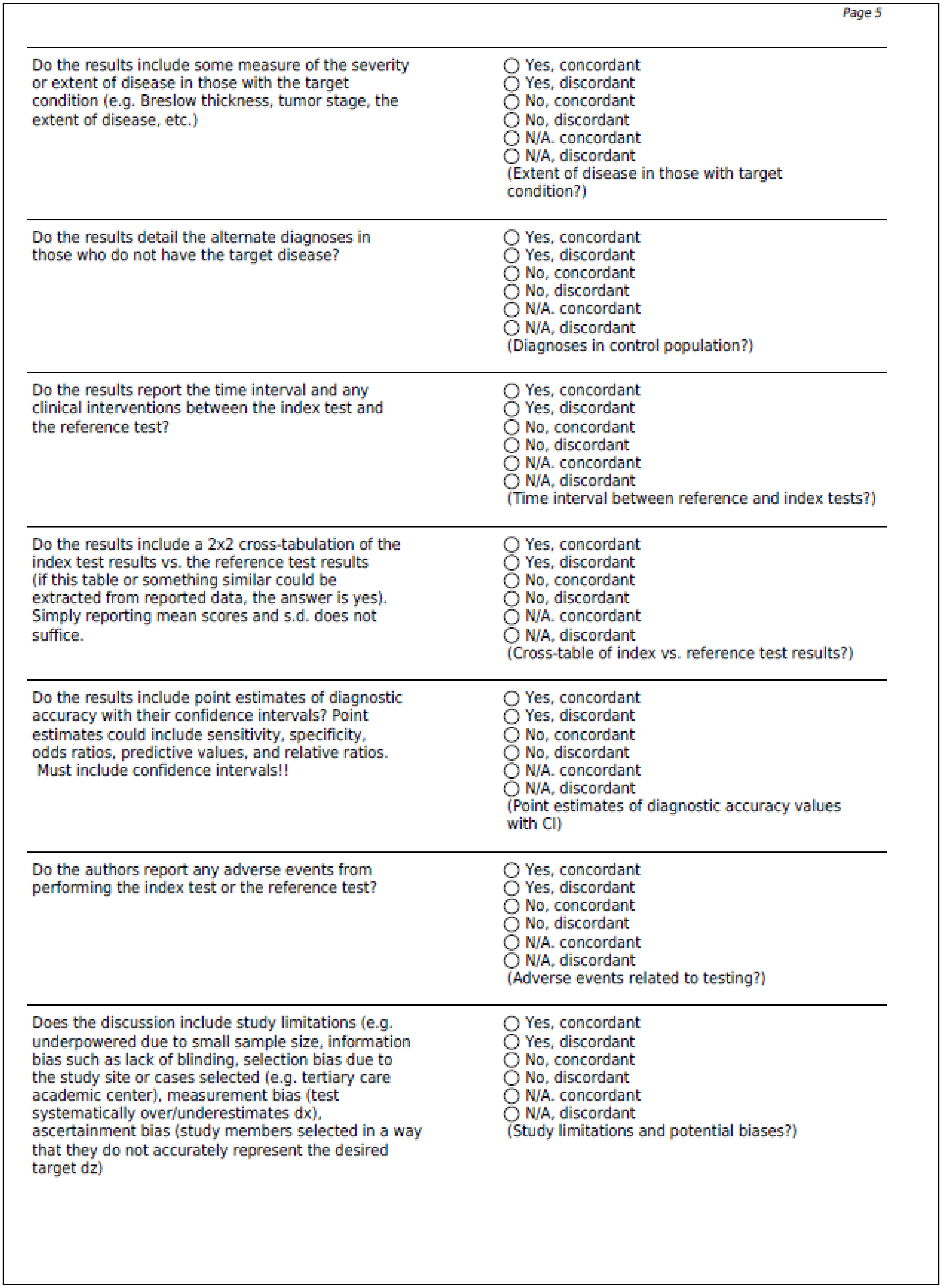
Page 3

Do the methods state if subjects form a consecutive, random, or convenience series?	○ Yes, concordant ○ Yes, discordant ○ No, concordant ○ No, discordant ○ N/A. concordant ○ N/A, discordant (Consecutive, random, or convenience series of subjects?)
Do the methods explain the index test in sufficient detail to allow replication?	○ Yes, concordant ○ Yes, discordant ○ No, concordant ○ No, discordant ○ N/A. concordant ○ N/A, discordant (method of index test explained)
Do the methods explain the reference test in sufficient detail to allow replication?	○ Yes, concordant ○ Yes, discordant ○ No, concordant ○ No, discordant ○ N/A. concordant ○ N/A, discordant (Method of reference test explained)
Do the methods give reasons for choosing the reference test as a comparison standard for the index test if alternatives exist? (If no alternative to histopathology or clinical criteria then select "N/A")	○ Yes, concordant ○ Yes, discordant ○ No, concordant ○ No, discordant ○ N/A. concordant ○ N/A, discordant (rationale for choosing reference test if alternatives exist)
Do the methods give index test cut-off values or result categories AND imply or state whether they are exploratory or pre-specified	○ Yes, concordant ○ Yes, discordant ○ No, concordant ○ No, discordant ○ N/A. concordant ○ N/A, discordant (index test cut-offs listed and explained)
Do the methods give reference test cut-off values or result categories AND imply whether they are pre-specified or exploratory?	○ Yes, concordant ○ Yes, discordant ○ No, concordant ○ No, discordant ○ N/A. concordant ○ N/A, discordant (reference test cut-offs explained)
Do the methods state whether test readers were blinded to clinical information and reference test results?	○ Yes, concordant ○ Yes, discordant ○ No, concordant ○ No, discordant ○ N/A. concordant ○ N/A, discordant (Blinded study?)



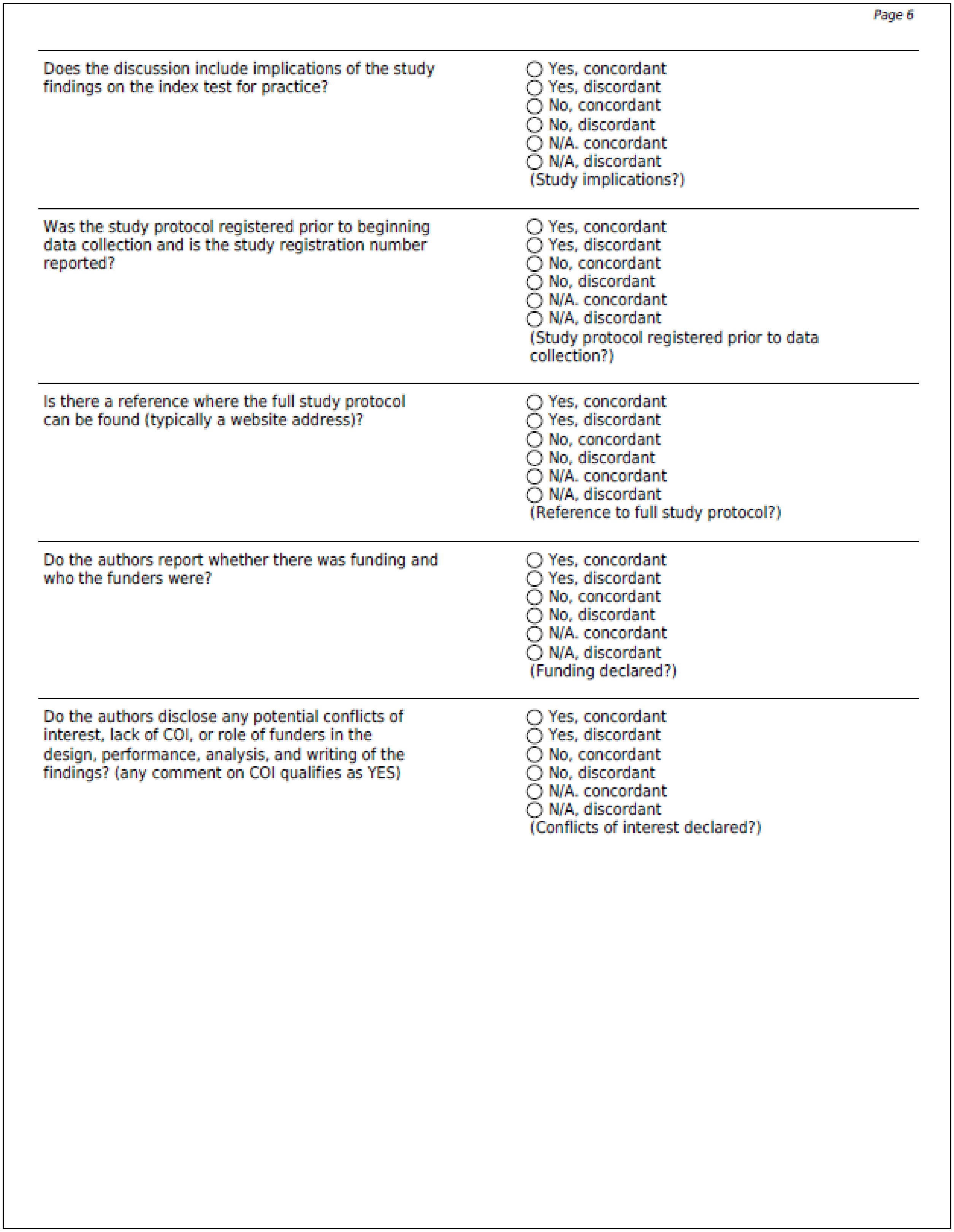
Page 4

Do the methods give a general description of the statistical methods used in the study?	○ Yes, concordant ○ Yes, discordant ○ No, concordant ○ No, discordant ○ N/A. concordant ○ N/A, discordant (Statistical methods listed?)
Do the methods explain how indeterminate or poor quality results were handled (e.g. tissue cores in a TMA that produced incomplete cross-sections)	○ Yes, concordant ○ Yes, discordant ○ No, concordant ○ No, discordant ○ N/A. concordant ○ N/A, discordant (How were indeterminate results handled?)
Do the methods explain how missing data on the index or reference tests were handled?	○ Yes, concordant ○ Yes, discordant ○ No, concordant ○ No, discordant ○ N/A. concordant ○ N/A, discordant (Were missing data handling methods explained?)
Do the methods include plans for stratified analyses for covariates like sex, smoking, immunosuppression, or age that may have confounded the test results, implying whether they were pre-planned or performed post hoc?	○ Yes, concordant ○ Yes, discordant ○ No, concordant ○ No, discordant ○ N/A. concordant ○ N/A, discordant (Stratified analyses for confounding planned in advance?)
Do the methods include a power analysis for sample size or list a desired sample size with an explanation for why that number was chosen?	○ Yes, concordant ○ Yes, discordant ○ No, concordant ○ No, discordant ○ N/A. concordant ○ N/A, discordant (Prespecified desired sample size listed?)
Do the results include a flow diagram of the participants from screening to selection to participation (e.g. including eligible, included, and analyzed subjects)	○ Yes, concordant ○ Yes, discordant ○ No, concordant ○ No, discordant ○ N/A. concordant ○ N/A, discordant (Participant flow diagram?)
Do the results include baseline clinical demographic data (minimum requirements are sex and age of the subjects)	○ Yes, concordant ○ Yes, discordant ○ No, concordant ○ No, discordant ○ N/A. concordant ○ N/A, discordant (Participant demographics?)



Page 5

Do the results include some measure of the severity or extent of disease in those with the target condition (e.g. Breslow thickness, tumor stage, the extent of disease, etc.)	○ Yes, concordant ○ Yes, discordant ○ No, concordant ○ No, discordant ○ N/A. concordant ○ N/A, discordant (Extent of disease in those with target condition?)
Do the results detail the alternate diagnoses in those who do not have the target disease?	○ Yes, concordant ○ Yes, discordant ○ No, concordant ○ No, discordant ○ N/A. concordant ○ N/A, discordant (Diagnoses in control population?)
Do the results report the time interval and any clinical interventions between the index test and the reference test?	○ Yes, concordant ○ Yes, discordant ○ No, concordant ○ No, discordant ○ N/A. concordant ○ N/A, discordant (Time interval between reference and index tests?)
Do the results include a 2x2 cross-tabulation of the index test results vs. the reference test results (if this table or something similar could be extracted from reported data, the answer is yes). Simply reporting mean scores and s.d. does not suffice.	○ Yes, concordant ○ Yes, discordant ○ No, concordant ○ No, discordant ○ N/A. concordant ○ N/A, discordant (Cross-table of index vs. reference test results?)
Do the results include point estimates of diagnostic accuracy with their confidence intervals? Point estimates could include sensitivity, specificity, odds ratios, predictive values, and relative ratios. Must include confidence intervals!!	○ Yes, concordant ○ Yes, discordant ○ No, concordant ○ No, discordant ○ N/A. concordant ○ N/A, discordant (Point estimates of diagnostic accuracy values with CI)
Do the authors report any adverse events from performing the index test or the reference test?	○ Yes, concordant ○ Yes, discordant ○ No, concordant ○ No, discordant ○ N/A. concordant ○ N/A, discordant (Adverse events related to testing?)
Does the discussion include study limitations (e.g. underpowered due to small sample size, information bias such as lack of blinding, selection bias due to the study site or cases selected (e.g. tertiary care academic center), measurement bias (test systematically over/underestimates dx), ascertainment bias (study members selected in a way that they do not accurately represent the desired target dz)	○ Yes, concordant ○ Yes, discordant ○ No, concordant ○ No, discordant ○ N/A. concordant ○ N/A, discordant (Study limitations and potential biases?)



Page 6

Does the discussion include implications of the study findings on the index test for practice?	○ Yes, concordant ○ Yes, discordant ○ No, concordant ○ No, discordant ○ N/A. concordant ○ N/A, discordant (Study implications?)
Was the study protocol registered prior to beginning data collection and is the study registration number reported?	○ Yes, concordant ○ Yes, discordant ○ No, concordant ○ No, discordant ○ N/A. concordant ○ N/A, discordant (Study protocol registered prior to data collection?)
Is there a reference where the full study protocol can be found (typically a website address)?	○ Yes, concordant ○ Yes, discordant ○ No, concordant ○ No, discordant ○ N/A. concordant ○ N/A, discordant (Reference to full study protocol?)
Do the authors report whether there was funding and who the funders were?	○ Yes, concordant ○ Yes, discordant ○ No, concordant ○ No, discordant ○ N/A. concordant ○ N/A, discordant (Funding declared?)
Do the authors disclose any potential conflicts of interest, lack of COI, or role of funders in the design, performance, analysis, and writing of the findings? (any comment on COI qualifies as YES)	○ Yes, concordant ○ Yes, discordant ○ No, concordant ○ No, discordant ○ N/A. concordant ○ N/A, discordant (Conflicts of interest declared?)

Appendix C: Journal Attributes

Journal	**n (%)**	**2018 Impact Factor**	**Endorses STARD?**	**Endorses ICMJE?**
Am J Clin Pathol	12 (7.02)	1.962	No	Yes
Am J Surg Pathol	15 (8.77)	6.155	No	Yes
Arch Pathol Lab Med	7 (4.09)	4.151	No	Yes
APMIS	3 (1.75)	2.225	No	Yes
Am J Pathol	1 (0.58)	3.762	No	No
AIMM	3 (1.75)	1.863	No	No
Cancer Cytopathol	12 (7.02)	4.425	Yes	Yes
Diagn Cytopathol	26 (15.20)	1.402	No	No
Diagn Pathol	2 (1.17)	2.528	Yes	Yes
Dis Markers	11 (6.43)	2.761	Yes	No
Exp Mol Pathol	2 (1.17)	2.350	No	No
Histopathol	5 (2.92)	3.294	Yes	No
Hum Pathol	9 (5.26)	2.740	No	No
J Clin Pathol	28 (16.37)	2.346	Yes	Yes
J Cutan Pathol	5 (2.92)	1.524	Yes	No
J Mol Diagn	7 (4.09)	4.426	Yes	Yes
J Oral Path Med	5 (2.92)	2.030	Yes	No
Mod Pathol	10 (5.85)	6.365	No	No
Pathol Onc Res	3 (1.75)	2.433	Yes	No
Virchow Arch	5 (2.92)	2.868	No	No