Supplementary materials

PICOS

Table S1: PICOS specified per criterium.

Criterium	Description
P: Population of interest	CT scans or thoracic X-rays of patients, that were analyzed for the presence of rib fractures by at least 2 radiologists, which was stated as the reference standard.
I: Intervention	Diagnostic detection by an artificial intelligence tool on its own
C: Comparison	All comparisons
O: Outcome	Number of true positives, true negatives, false positives, and false negatives and/or the sensitivity, and specificity
S: Study type	Diagnostic case-control studies, diagnostic cohort studies, and diagnostic RCTs

Quality assessment tool

Table S2: Quality assessment tool (modified QUADAS-2) [15].

DOMAIN 1: PATIENT SELECTION

A. Risk of Bias

- ❖ Was a consecutive or random sample of patients enrolled? — Yes / No / Unclear
- ❖ Was a case-control design avoided? — Yes / No / Unclear
- ❖ Did the study avoid inappropriate exclusions? — Yes, / No / Unclear

Could the selection of patients have introduced bias? — RISK: LOW / HIGH / UNCLEAR

B. Concerns regarding applicability

Is there concern that the included patients do not match the review question? — CONCERN: LOW / HIGH / UNCLEAR

DOMAIN 2: INDEX TEST(S)

A. Risk of Bias

- ❖ Was the AI trained on a multicenter dataset? — Yes / No / Unclear

Could the conduct or interpretation of the index test have introduced bias? — RISK: LOW / HIGH / UNCLEAR

B. Concerns regarding applicability

Is there concern that the index test, its conduct, or interpretation differ from the review question? — CONCERN: LOW / HIGH / UNCLEAR

DOMAIN 3: REFERENCE STANDARD

A. Risk of Bias

- Is the reference standard likely to correctly classify the target condition?
  - Yes
  - No
  - Unclear

- Were the reference standard results interpreted without knowledge of the results of the index test?
  - Yes
  - No
  - Unclear

Could the reference standard, its conduct, or its interpretation have introduced bias?

RISK:
- LOW
- HIGH
- UNCLEAR

B. Concerns regarding applicability

Is there concern that the target condition as defined by the reference standard does not match the review question?

CONCERN:
- LOW
- HIGH
- UNCLEAR

DOMAIN 4: FLOW AND TIMING

A. Risk of Bias

- Did all patients receive a reference standard?
  - Yes
  - No
  - Unclear

- Did patients receive the same reference standard?
  - Yes
  - No
  - Unclear

- Were all patients included in the analysis?
  - Yes
  - No
  - Unclear

Could the patient flow have introduced bias?

RISK:
- LOW
- HIGH
- UNCLEAR

Text S1: Rationale behind the modifications to the QUADAS-2 tool

- We deleted the guiding question in domain 2A "Were the index test results interpreted without knowledge of the results of the reference standard?", as the AI was only trained to process visual data provided through a CT scan, and as the exact same CT scan was assessed by the reference standard. Thus, we deemed that there was a very small chance of reference standard having influenced the AI.

- We deleted the guiding question in domain 2A "If a threshold was used, was it pre-specified?", as thresholds were irrelevant to our research question.

- We added the guiding question in domain 2A "Was the AI trained on a multicenter dataset?" as training on multiple centers would decrease the chance of the AI containing bias and decrease the chance of it being overfitted.

- We deleted the guiding question in domain 4 "Was there an appropriate interval between index test(s) and reference standard?", as a time interval would not have influenced the CT and therefore would not have influenced the diagnostic accuracy of either the reference standard or the index test.

Study characteristics

Table S3: Study characteristics of the included studies.

The amount of included patients and included CT scans was deemed as being the same, as we believe that the large majority of included patients underwent a single CT scan.

Author, year	Number of patients or CT scans in dataset	Input features	Reference standard	Comparisons	Relevant outcomes	Type of study	Quality
Gipson et al., 2022 [28]	1400	CT	Contemporaneous CT reports	Comparison with reference standard and performance of radiologists using the AI tool	Sensitivity, specificity, TP, FN, FP, and TN	Retrospective diagnostic cohort study	High
Jin et al., 2020 [27]	120	CT	Five radiologists of 3 to 20 years of experience.	Comparison with different AI tools	Sensitivity	Retrospective diagnostic cohort study	High
Kaiume et al., 2021 [26]	39	CT	Two radiologists with 26 and 6 years of image interpretation experience	Diagnostic performance rib fractures of two intern doctors	Sensitivity	Retrospective diagnostic cohort study	High
Niiya et al., 2022 [25]	56	CT	Two radiologists with 6 and 9 years of experience.	Comparison with reference standard	Sensitivity	Retrospective diagnostic case-control study	High
Wang et al., 2022 [24]	1613	CT	Two radiologists with at least 9 years of experience and in case of inconclusion they made consensus with a senior radiologists with at least 20 years of experience.	Comparison with six attending radiologists	Sensitivity, and specificity	Retrospective diagnostic case-control study	High
Wu et al., 2021 [23]	105	CT	Three radiologists with 6, 10, and 14 years of experience and one senior radiologist with 18 years of experience.	Comparison radiologists who used AI to diagnose	Sensitivity	Retrospective diagnostic case-control study	High
Yang et al., 2022 [21]	120	CT	Two experienced musculoskeletal radiologists with at least 10 years of experience and a third radiologist was invited to participate if there was a discussion.	Comparison with the diagnosis of three radiologists, with 5, 7 and 21 years of experience. Those radiologists were not the same as the radiologists who determined the reference standard.	Sensitivity, TP, FP, TN, and FN	Retrospective diagnostic cohort study	High
Yao et al., 2021 [22]	100	CT	Three experienced radiologists (over 10 years experience) and checking by two senior radiologists (over 15 years experience).	Comparison of the performance between AI, radiologist and radiologic-AI collaboration	Sensitivity	Retrospective diagnostic cohort study	High
Zhou et al., 2020 [20]	30	CT	Two experienced musculoskeletal radiologists with 8 and 9 years of experience and two senior radiologists with 20 and 14 years of experience. If the	Comparison with the performance of five attending radiologists with 6–8 years of experience. There was no overlap between	Sensitivity, and FP	Multicentre retrospective diagnostic case-control study	High

			conclusion was inconsistent, one thoracic surgeon was invited to participate in the discussion.	those radiologists and the radiologists who determined the reference standard.			
Zhou et al., 2021 [17]	260	CT	Two experienced musculoskeletal radiologists with 8 and 9 years of experience, two senior radiologists with 20 and 14 years of experience and one thoracic surgeon in case of inconclusion.	Five radiologists with 6 to 8 years of experience with no overlap with the radiologists who determined the reference standard	Sensitivity, and specificity	Multicentre retrospective diagnostic cohort study	High
Zhou et al., 2022 [18]	164	CT	Two musculoskeletal radiologists with five years of experience and one senior musculoskeletal radiologist with more than ten years of experience.	Comparison with different AI tools	Sensitivity	Retrospective diagnostic cohort study	Intermediate
Zhou et al., 2022. [19]	Internal dataset: 90 External dataset: 38	CT	Two experienced musculoskeletal radiologists (9 and 10 years of experience), two senior radiologists (21 and 15 years of experience) and in doubt one thoracic surgeon.	Comparison with the diagnosis of five radiologists with 7–9 years of CT diagnosis experience which were different from the radiologists who determined the reference standard	Sensitivity, TP, FN. and FP	Multicentre retrospective diagnostic cohort study	Intermediate

Quality assessment of the included studies

Table S4: Quality assessment of the included studies.

Author, year	Researchers	Patient selection		Index test		Reference test		Flow and timing	Score	Quality
		1A	1B	2A	2B	3A	3B	4		
Gipson et al., 2022 [28]	MCL, JH, (LFM)	1	1	1	1	0	1	1	6	High
Jin et al., 2020 [27]	LFM, JH, (MCL)	0	1	1	0	1	1	1	5	High
Kaiume et al., 2021 [26]	LFM, JH, (MCL)	1	1	1	1	1	0	1	6	High
Niiya et al., 2022 [25]	MCL, LFM, (JH)	0	1	1	1	1	1	1	6	High
Wang et al., 2022 [24]	LFM, JH, (MCL)	1	1	1	1	1	1	1	7	High
Wu et al., 2021 [23]	LFM, JH, (MCL)	1	1	1	1	0	1	1	6	High
Yang et al., 2022 [21]	LFM, MCL, (JH)	1	1	1	1	1	1	1	7	High
Yao et al., 2021 [22]	LFM, JH, (MCL)	1	1	1	1	1	1	1	7	High
Zhou et al., 2020 [20]	MCL, JH, (LFM)	0	1	1	1	1	1	1	6	High
Zhou et al., 2021 [17]	LFM, MCL, (JH)	1	0	1	0	1	1	1	5	High
Zhou et al., 2022 [18]	LFM, JH, (MCL)	0	0	0	1	1	1	1	4	Intermediate
Zhou et al., 2022. [19]	MCL, LFM, (JH)	0	0	1	1	1	0	1	4	Intermediate

Results of individual studies

Table S5.1: Results of individual studies – Primary outcomes.

- Data was not available directly and could not be calculated.
Italic data has been calculated using other data.
* Estimated using the set of assumptions.

Author, year	Dataset	Total fractures	TP, FP, FN, TN	Sensitivity	Specificity
Gipson et al., 2022 [28]	-	348	143, 75, 205, *977*	0.411 (0,359–0,465)	0.929 (0.912–0.944)
Jin et al., 2020 [27]	Test cohort	882	*819*, *632*, *63*, -	0.929	-
Kaiume et al., 2021 [26]	Validation dataset	256	165, 43, 91, *637**	0.645 (0,586–0,703)	0.937
Niiya et al., 2022 [25]	Evaluation dataset	199	186, *106*, *13*, *1039**	0.935	*0.907**
Wang et al., 2022 [24]	Internal dataset	2096	1915, 788, *181*, *36188**	0.914 (0.901–0.925)	*0.979**
Wang et al., 2022 [24]	External dataset	4144	3521, 565, *623*, *34003**	0.850 (0.838–0.860)	*0.984**
Wu et al., 2021 [23]	Test 1	1545	-	-	-
Wu et al., 2021 [23]	Test 2	491	*417, 90, 75, -*	0.849 (0.803-0.867)	0.872 (0.825-0.887)
Wu et al., 2021 [23]	Test 3	-	-	-	-
Yang et al., 2022 [21]	Cohort 1	2856	-	-	-
Yang et al., 2022 [21]	Cohort 2	397	366, 122, 31, *2361**	0.9219	*0.951**
Yang et al., 2022 [21]	Cohort 3	309	288, 21, 111, *1470**	0.932	*0.986**
Yao et al., 2021 [22]	Testing set	436	398, *60, 38, 1188*	0.913	*0.952*
Zhou et al., 2020 [20]	Validation Set	494	*417, 134, 77, -*	0.845	-
Zhou et al., 2021 [17]	Validation set	525	-	-	-
Zhou et al., 2022 [18]	Test set	*627*	*510, -, 117, -*	0.8128	-
Zhou et al., 2022. [19]	Testing dataset	*427*	*391, 45, 36, 1672**	*0.916*	*0.974**
Zhou et al., 2022. [19]	External dataset	*163*	*153, 32, 10, -*	*0.939*	-
Zhou et al., 2022. [19]	Competition dataset	*241*	*217, 23, 24, 808**	*0.900*	*0.972**

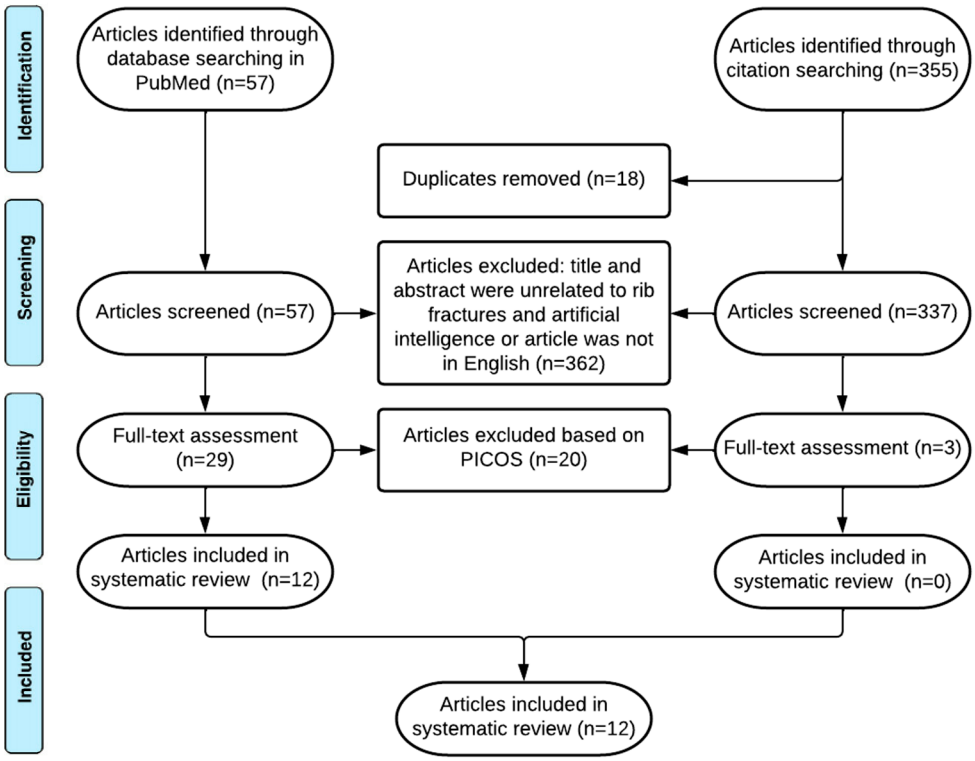
Table S5.2: Results of individual studies – All outcomes.

* Estimated using the set of assumptions
- Data was not available directly and could not be calculated.
Italic data has been calculated using other data

Author, year	Dataset	Year	TP	FP	FN	TN	Total positives	Total negatives	Total non-fractures	Total rib fractures	Sensitivity	Specificity	F1-score	PPV	NPV	Time
Gipson et al., 2022 [28]	-	2022	143	75	205	977	*218*	*1182*	*1052*	*348*	0.411	0.929	*0.505*	*0.656*	*0.827*	-
Jin et al., 2020 [27]	Test cohort	2020	*819*	*632*	*63*	-	*1451*	-	-	882	0.929	-	*0.702*	*0.564*	-	31
Kaiume et al., 2021 [26]	Validation dataset	2021	165	43	91	*637**	208	*728**	*680**	256	0.645	0.937	0.711	0.793	*0.875**	-
Niiya et al., 2022 [25]	Evaluation dataset	2022	186	*106*	*13*	*1039**	*292*	*1052**	*1145**	199	0.935	*0.907**	*0.758*	*0.637*	*0.988**	-
Wang et al., 2022 [24]	Internal dataset	2022	1915	788	*181*	*36188**	*2703*	*36369**	*36976**	2096	0.914	*0.979**	*0.798*	*0.708*	*0.995**	-
Wang et al., 2022 [24]	External dataset	2022	3521	565	*623*	*34003**	*4086*	*34626**	*34568*	4144	0.850	*0.984**	*0.856*	*0.862*	*0.982**	-
Wu et al., 2021 [23]	Test 1	2021	-	-	-	-	-	-	-	1545	-	-	-	-	-	-
Wu et al., 2021 [23]	Test 2	2021	*417*	*90*	*75*	-	*507*	-	-	491	0.849	0.872	0.833	0.822	-	-
Wu et al., 2021 [23]	Test 3	2021	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Yang et al., 2022 [21]	Cohort 1	2022	-	-	-	-	-	-	-	2856	-	-	-	-	-	-
Yang et al., 2022 [21]	Cohort 2	2022	366	122	31	*2361**	*488*	*2392**	*2483**	397	0.9219	*0.951**	0.8271	*0.750*	*0.987**	49.13
Yang et al., 2022 [21]	Cohort 3	2022	288	21	111	*1470**	*309*	*1581**	*1491**	309	0.932	*0.986**	0.8135	*0.932*	*0.930**	50.29
Yao et al., 2021 [22]	Testing set	2021	398	*60*	*38*	*1188*	458	*1226*	*1248*	436	0.913	*0.952*	0.890	0.869	0.969	20
Zhou et al., 2020 [20]	Validation Set	2020	*417*	*134*	*77*	-	*551*	-	-	494	0.845	-	*0.798*	0.757	-	-
Zhou et al., 2021 [17]	Validation set	2021	-	-	-	-	-	-	-	525	-	-	-	-	-	-
Zhou et al., 2022 [18]	Test set	2022	*510*	-	*117*	-	-	-	-	*627*	0.8128	-	-	-	-	-
Zhou et al., 2022. [19]	Testing dataset	2022	*391*	*45*	*36*	1672*	*436*	*1708**	*1717**	*427*	*0.916*	*0.974**	*0.906*	*0.897*	*0.979**	-
Zhou et al., 2022. [19]	External dataset	2022	*153*	*32*	*10*	-	*185*	-	-	*163*	*0.939*	-	*0.879*	*0.827*	-	-
Zhou et al., 2022. [19]	Competition dataset	2022	*217*	*23*	*24*	*808**	*240*	*831**	*831**	*241*	*0.900*	*0.972**	*0.902*	*0.904*	*0.971**	12.638

Study selection

Figure S1: PRISMA flow chart showing study selection from databases and citation searching.



Synthesis of results

Figure S2.1: Forest plot of the sensitivity of AI in rib fracture detection [18–28].

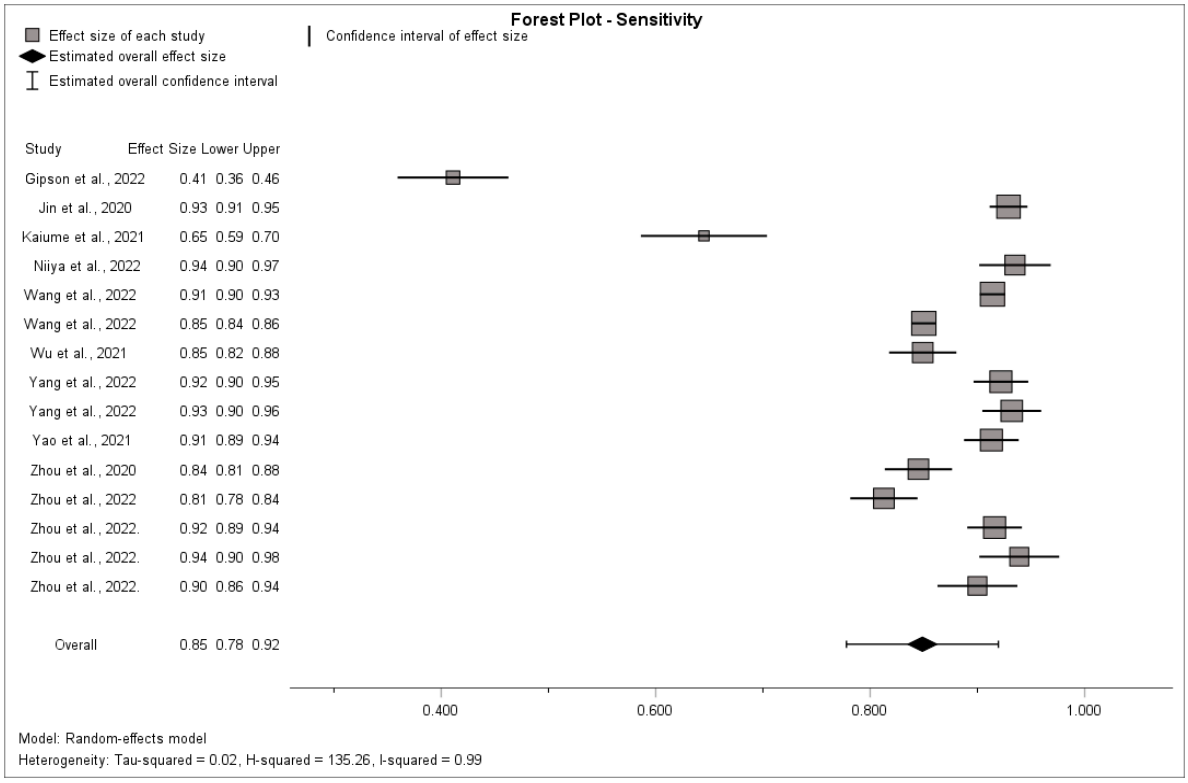


Figure S2.2: Forest plot of the specificity of AI in rib fracture detection [19,21,22,24–26,28].

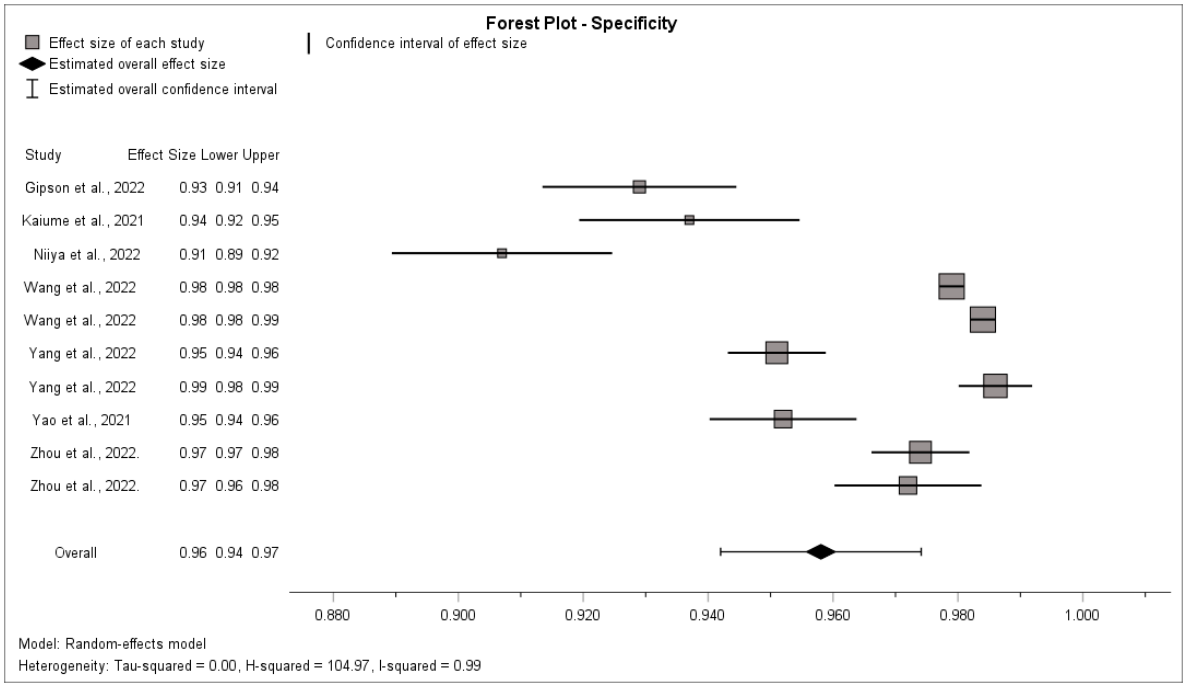


Figure S2.3: Forest plot of the Positive Predictive Value (PPV) [19–28].

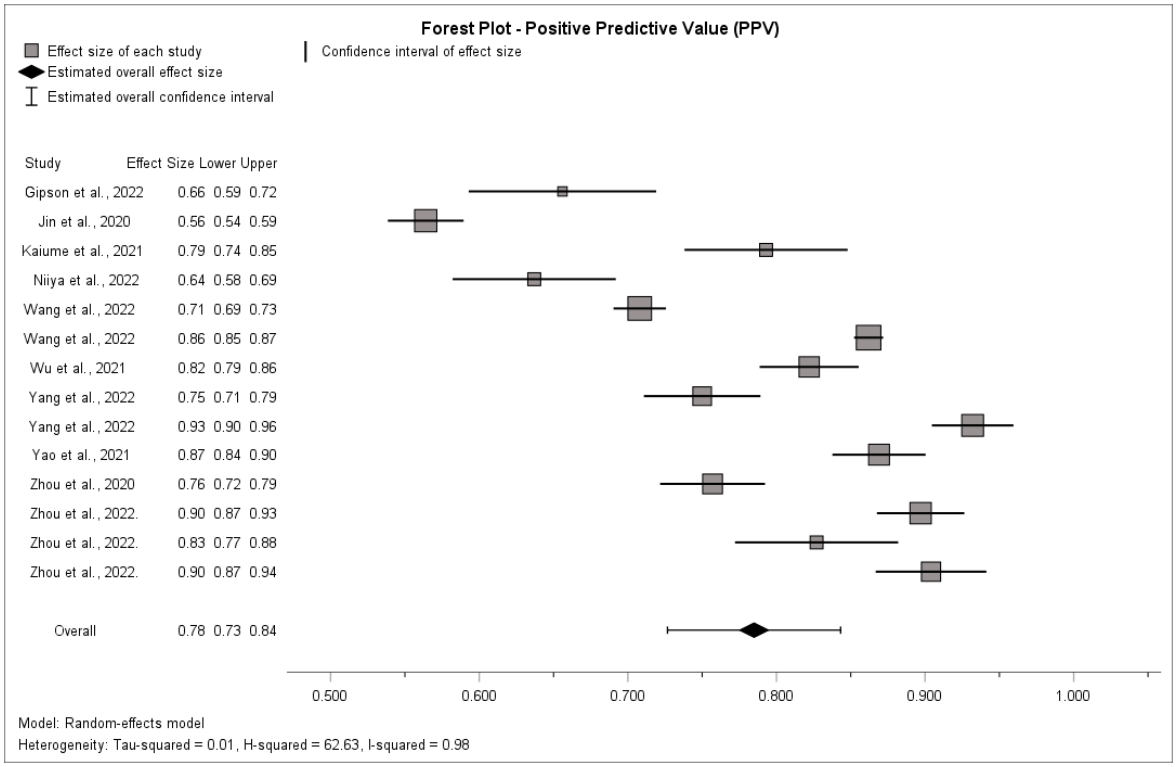


Figure S2.4: Forest plot of the Negative Predictive Value (NPV) [19,21–22,24–26,28].

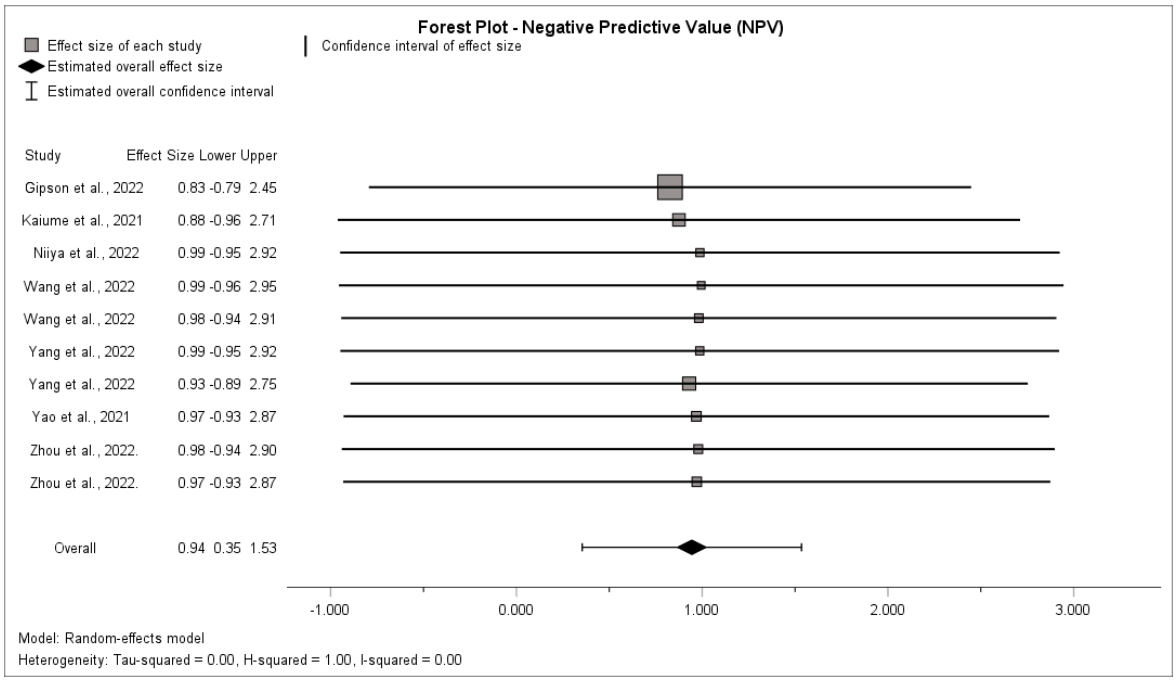


Figure S2.5: Forest plot of the F1-score [19–28].

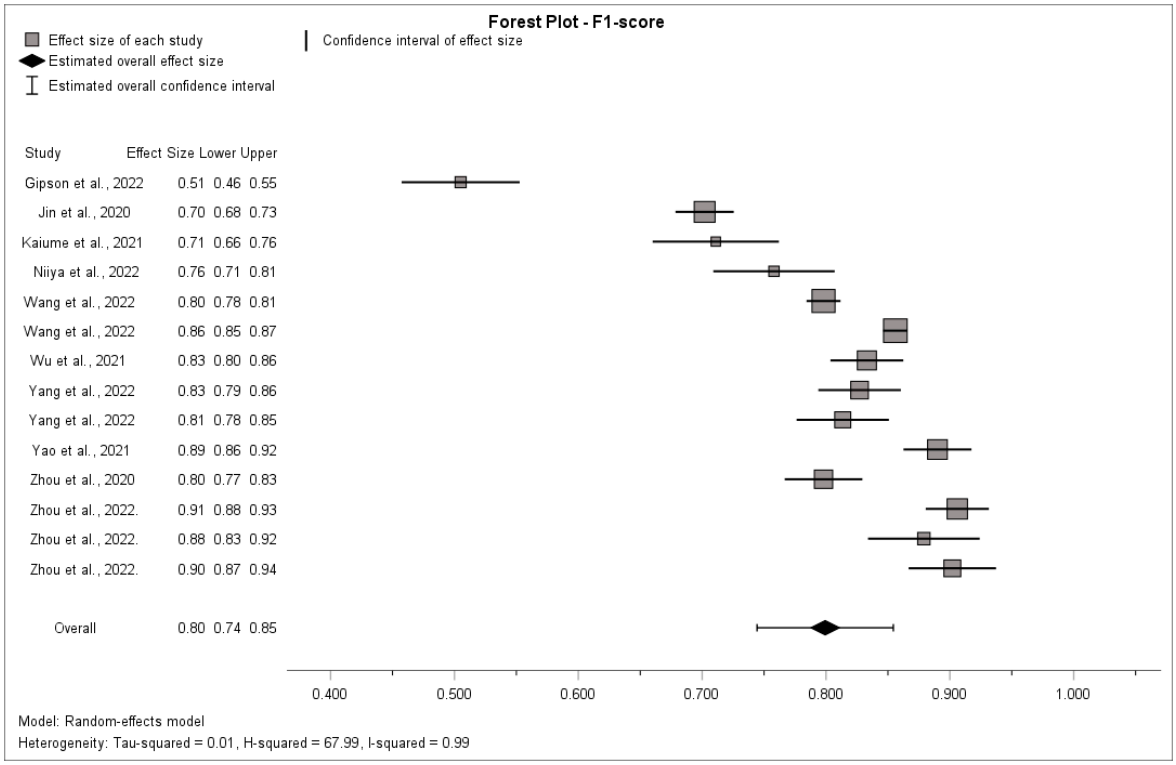
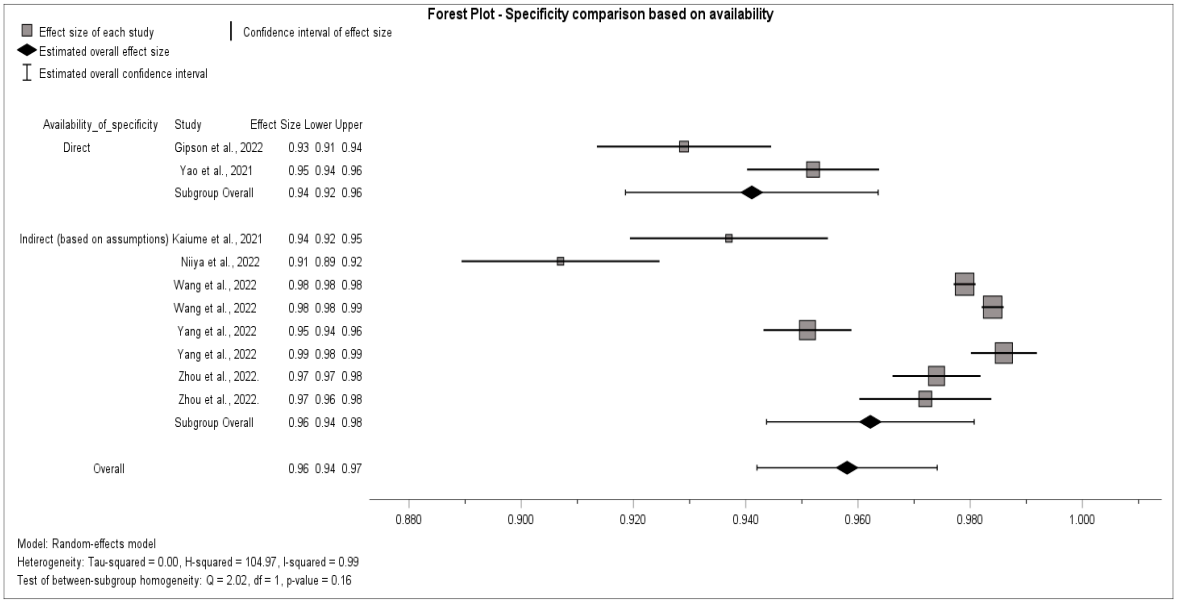


Figure S2.6: Forest plot comparing directly and indirectly available data [19,21,22,24–26,28].



Risk of bias across studies

Figure S3.1: Funnel plot of the Sensitivity [18–28].

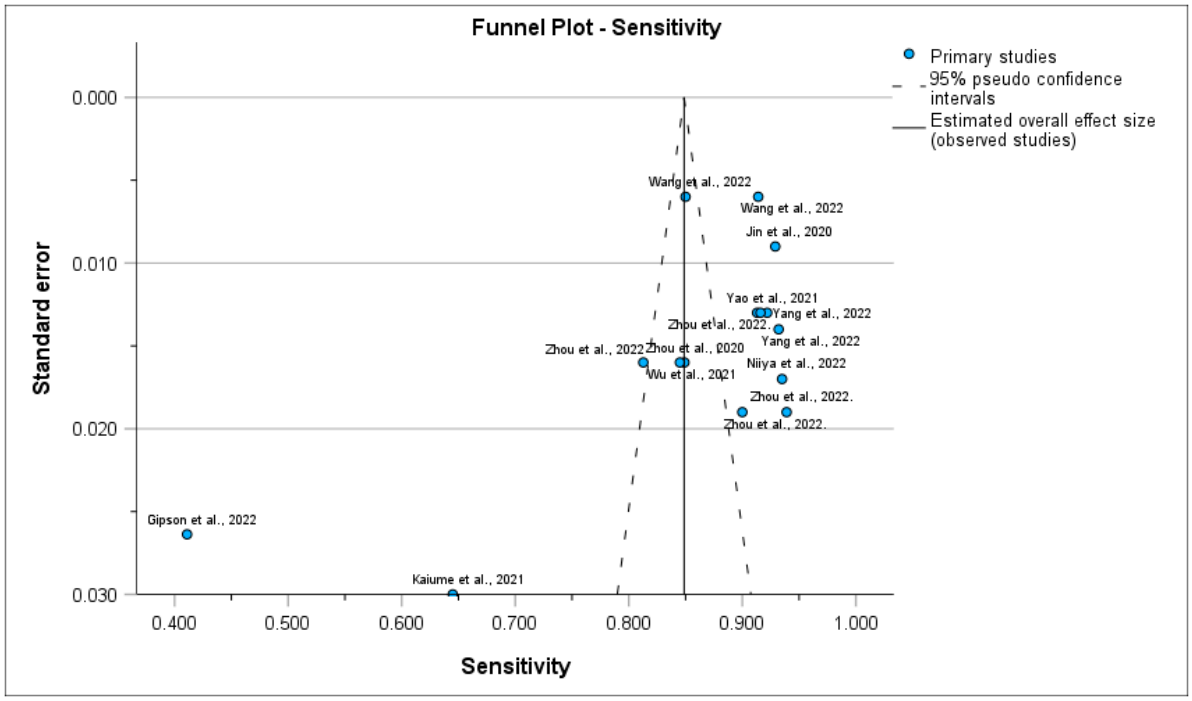


Figure S3.2: Funnel plot of the Specificity [19,21,22,24–26,28].

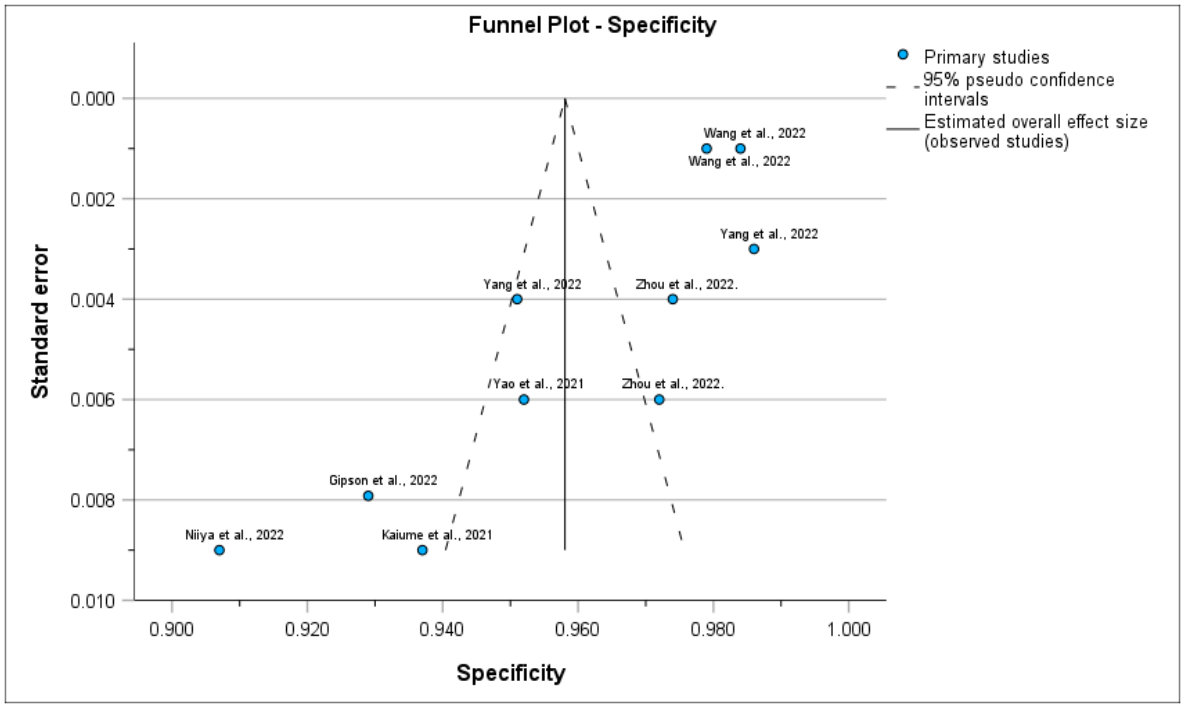


Figure S3.3: Funnel plot of the F1-score [19–28].

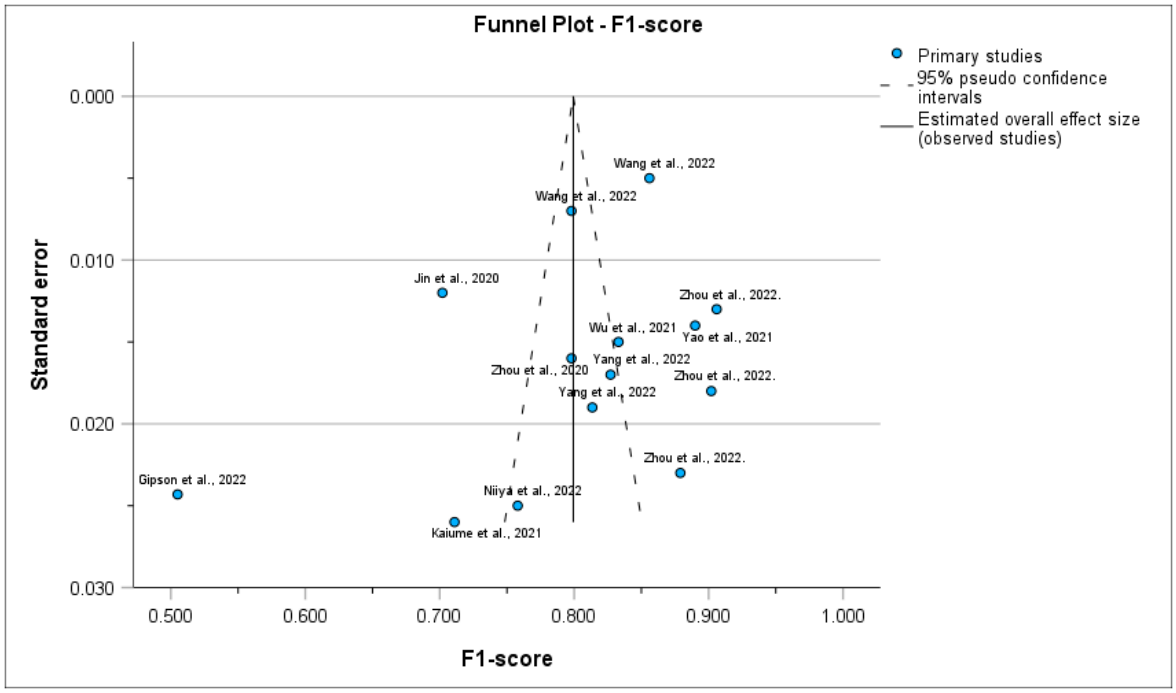


Figure S3.4: Funnel plot of the Positive Predictive Value (PPV) [19–28].

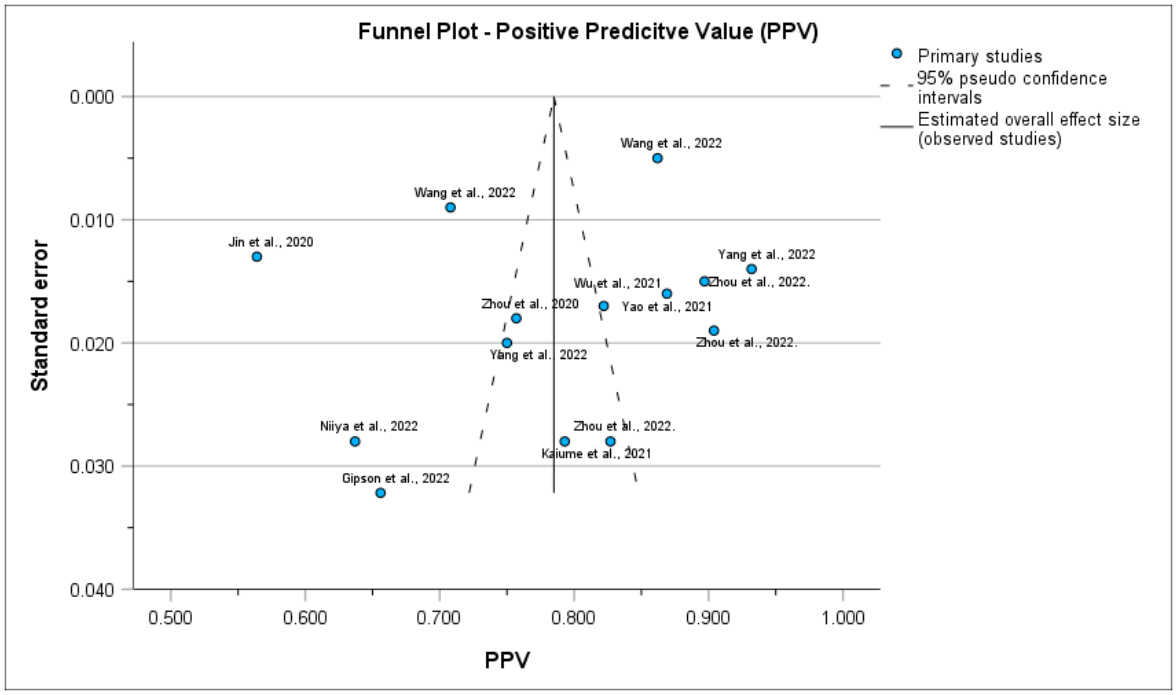


Figure S3.5: Funnel plot of the Negative Predictive Value (NPV) [19,21–22,24–26,28].

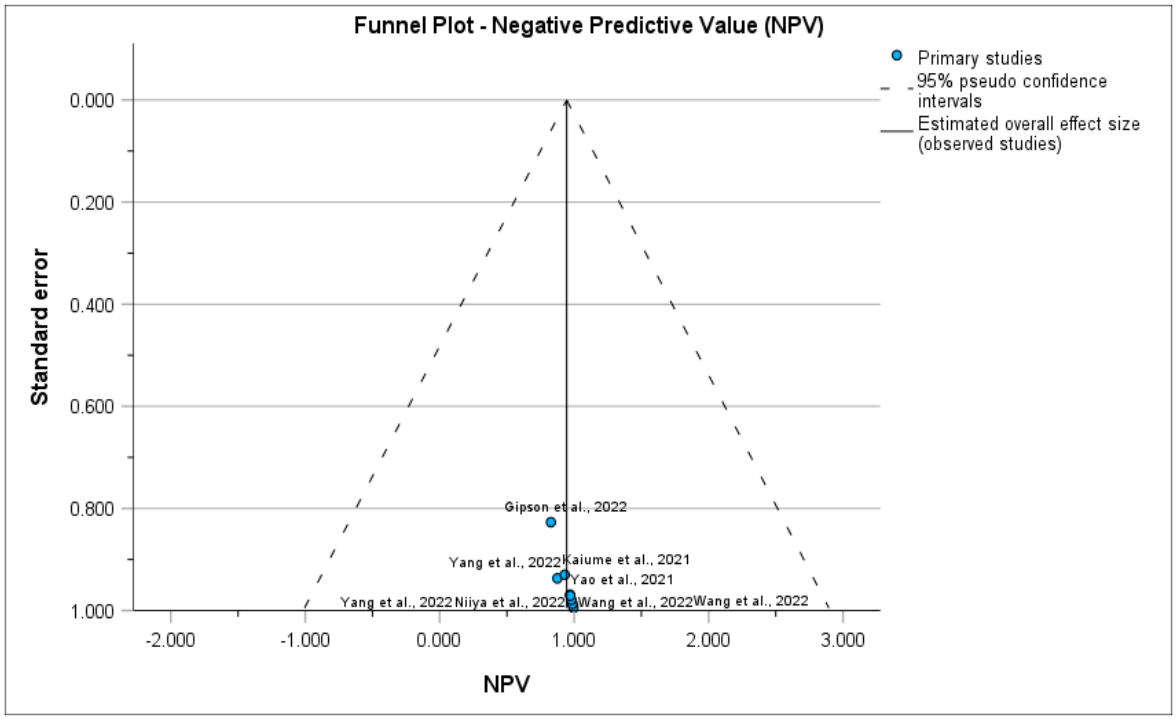


Figure S3.6: Assessment of risk of within-study selective reporting.

Author, year	Relevant outcomes stated in methods	Outcomes also reported as result?	Score	Risk of within-study selective reporting
Gipson et al., 2022 [28]	Sensitivity, specificity, TP, FN, FP, and TN	Yes	0	Low
Jin et al., 2020 [27]	Sensitivity	Yes	0	Low
Kaiume et al., 2021 [26]	Sensitivity	Yes	0	Low
Niiya et al., 2022 [25]	Sensitivity	Yes	0	Low
Wang et al., 2022 [24]	Sensitivity, and specificity	Yes (specificity was given, but not on a per-fracture level)	0	Low
Wu et al., 2021 [23]	Sensitivity	Yes	0	Low
Yang et al., 2022 [21	Sensitivity, TP, FP, TN, and FN	Yes	0	Low
Yao et al., 2021 [22]	Sensitivity	Yes	0	Low
Zhou et al., 2020 [20]	Sensitivity and FP	Yes	0	Low
Zhou et al., 2021 [17]	Sensitivity and specificity	Yes (but not per-fracture level)	0	Low
Zhou et al., 2022 [18]	Sensitivity	Yes	0	Low
Zhou et al., 2022. [19]	Sensitivity, TP, FN, and FP	Yes (but not per-fracture level)	0	Low

Additional analysis on quality scores

Figure S4.1: Forest plot of the sensitivity, comparison on domain 1A [18–28].

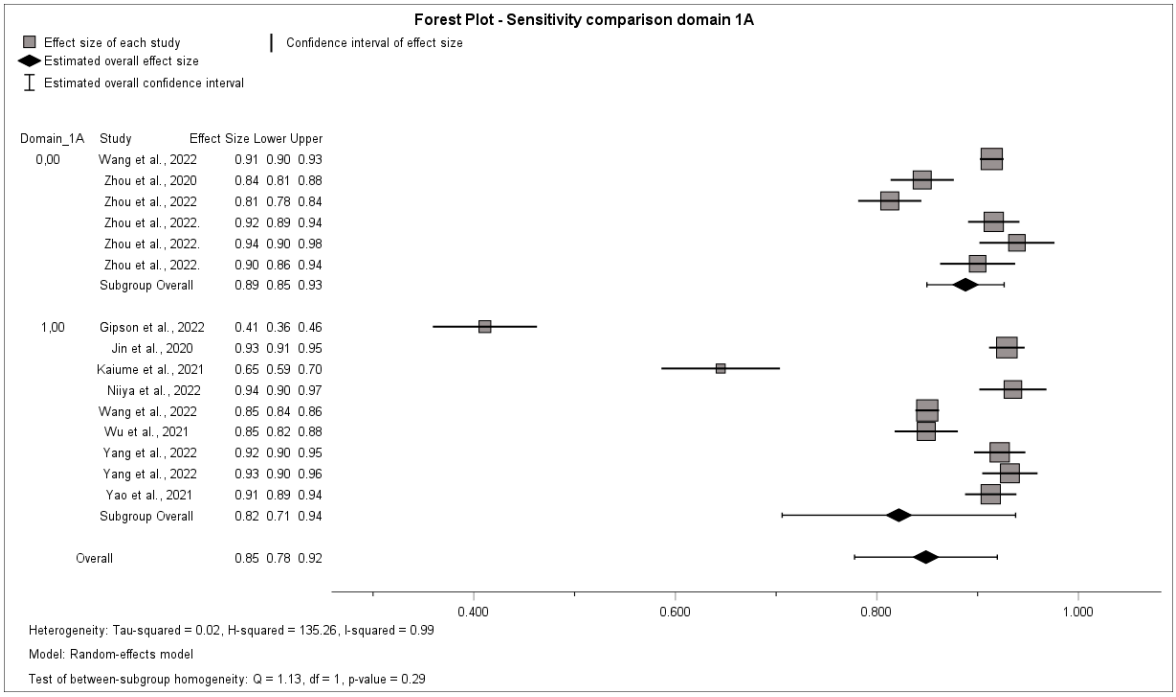


Figure S4.2: Forest plot of the sensitivity, comparison on domain 1B [18–28].

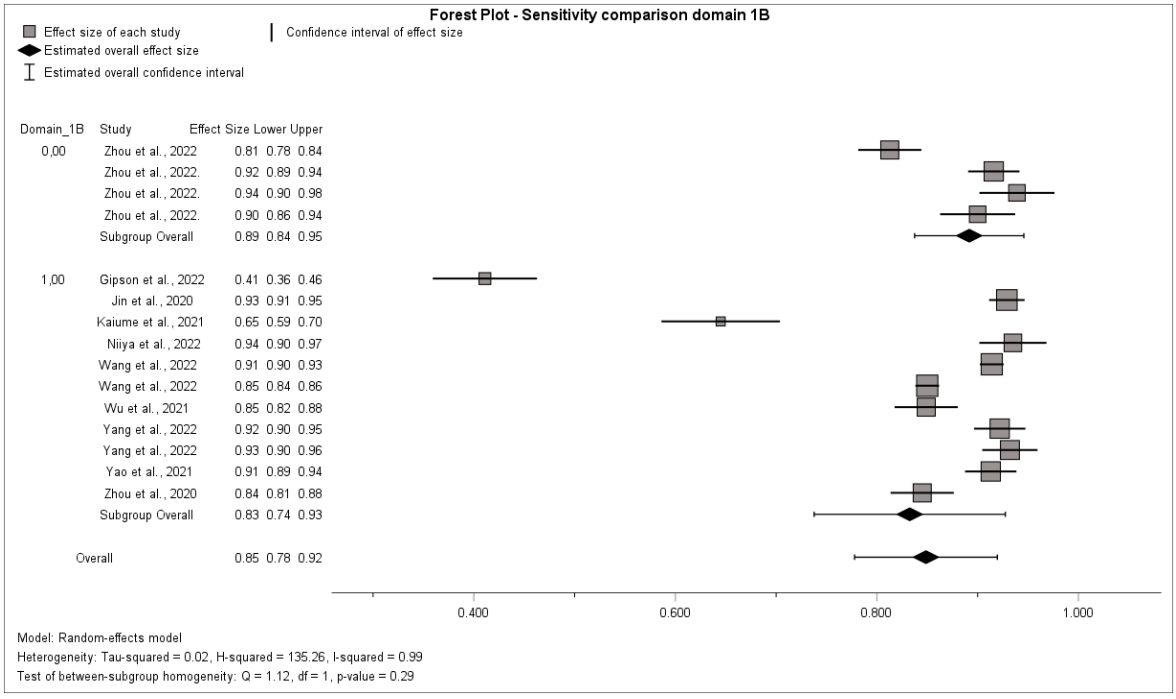


Figure S4.3: Forest plot of the sensitivity, comparison on domain 2A [18–28].

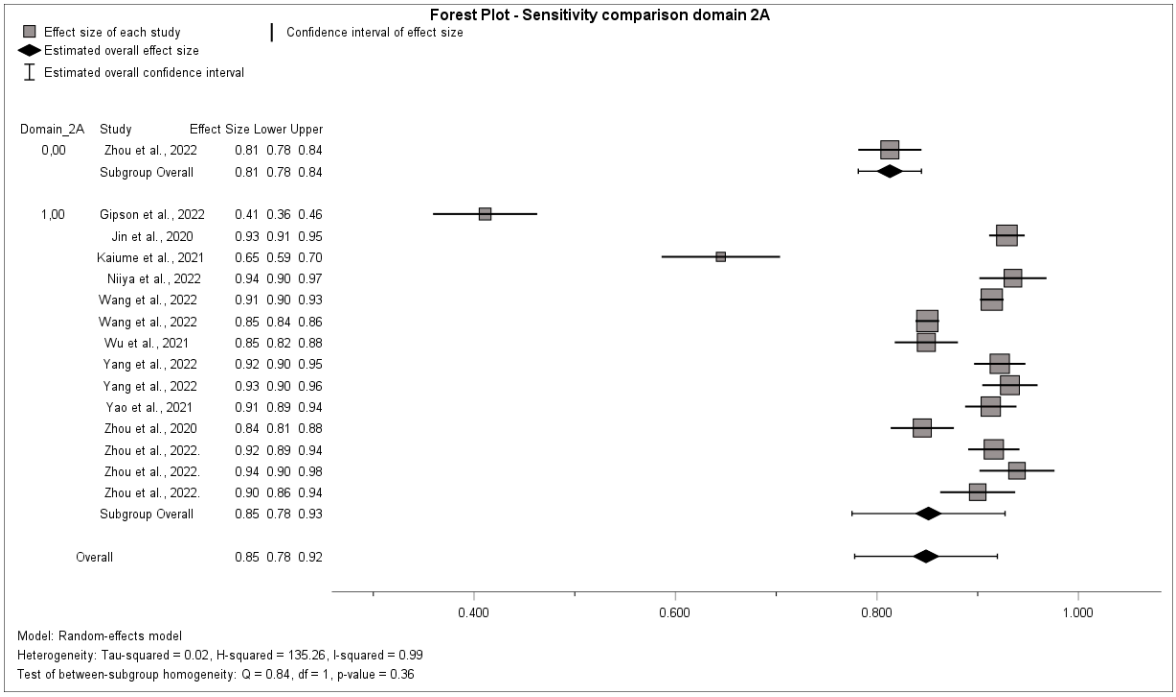


Figure S4.4: Forest plot of the sensitivity, comparison on domain 2B [18–28].

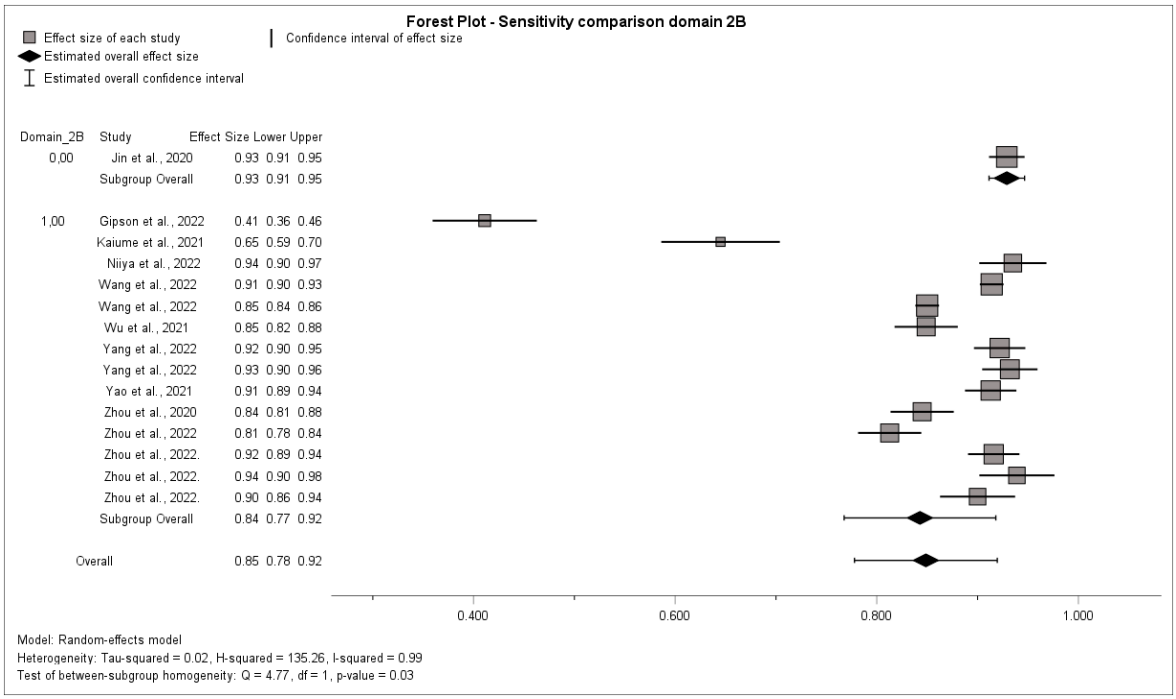


Figure S4.5: Forest plot of the sensitivity, comparison on domain 3A [18–28].

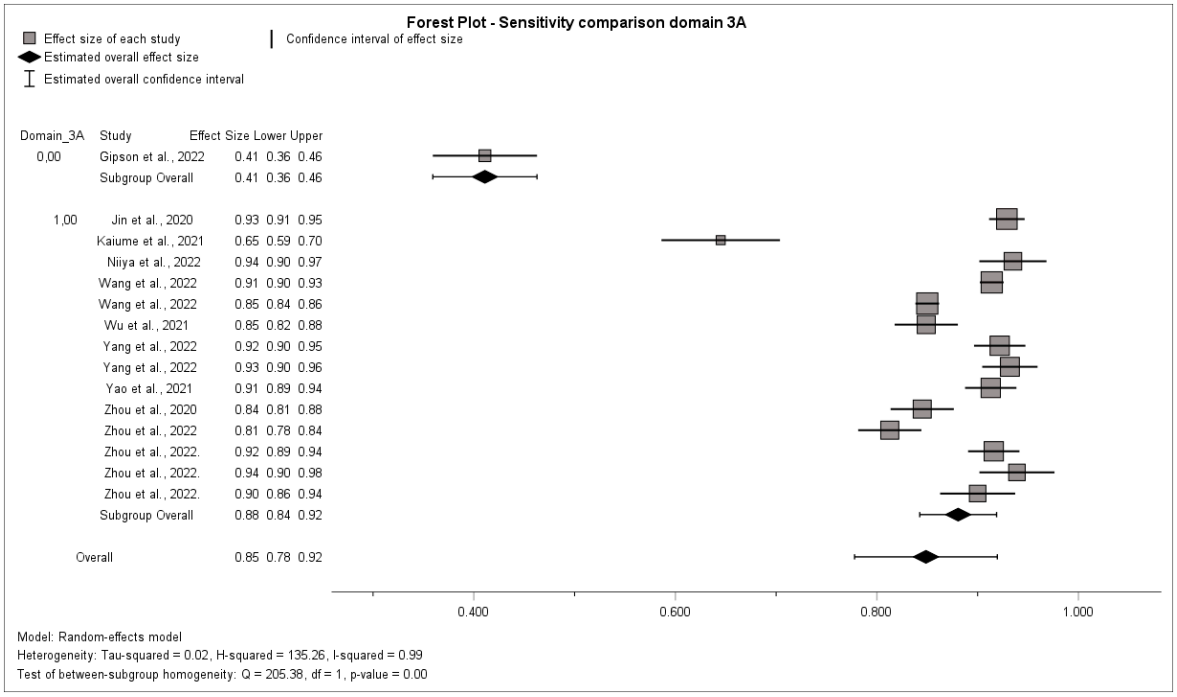


Figure S4.6: Forest plot of the sensitivity, comparison on domain 3B [18–28].

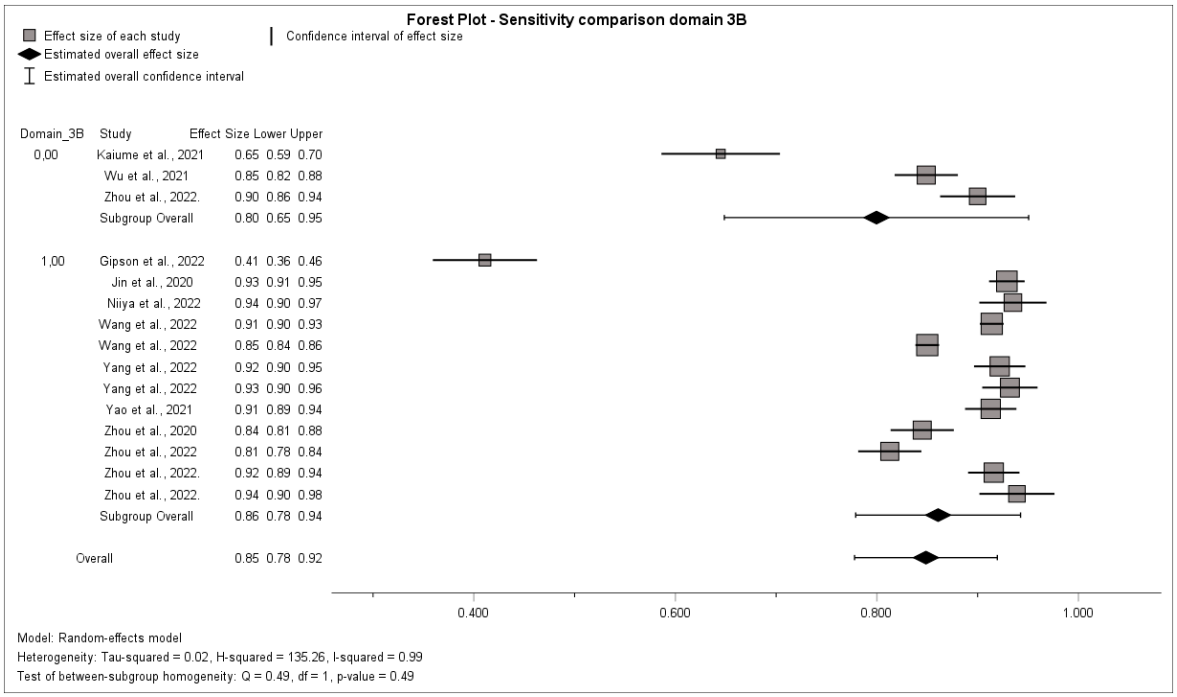


Figure S4.7: Forest plot of the sensitivity, comparison on domain 4 [18–28].

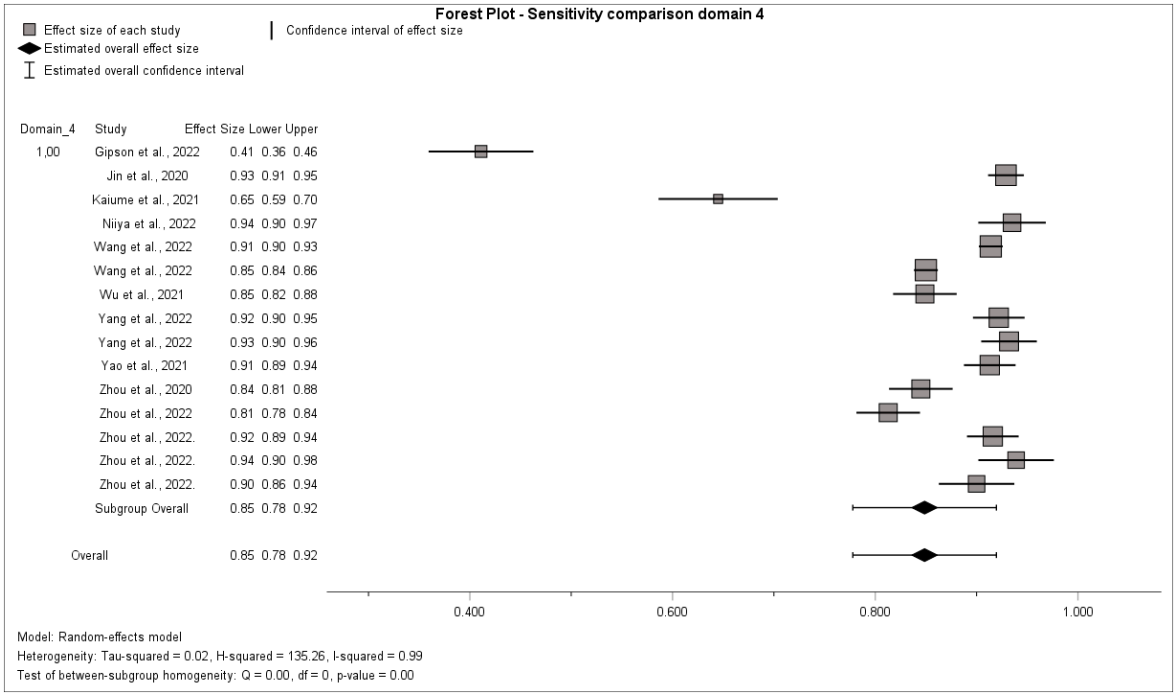


Figure S4.8: Forest plot of the specificity, comparison on domain 1A [19,21,22,24–26,28].

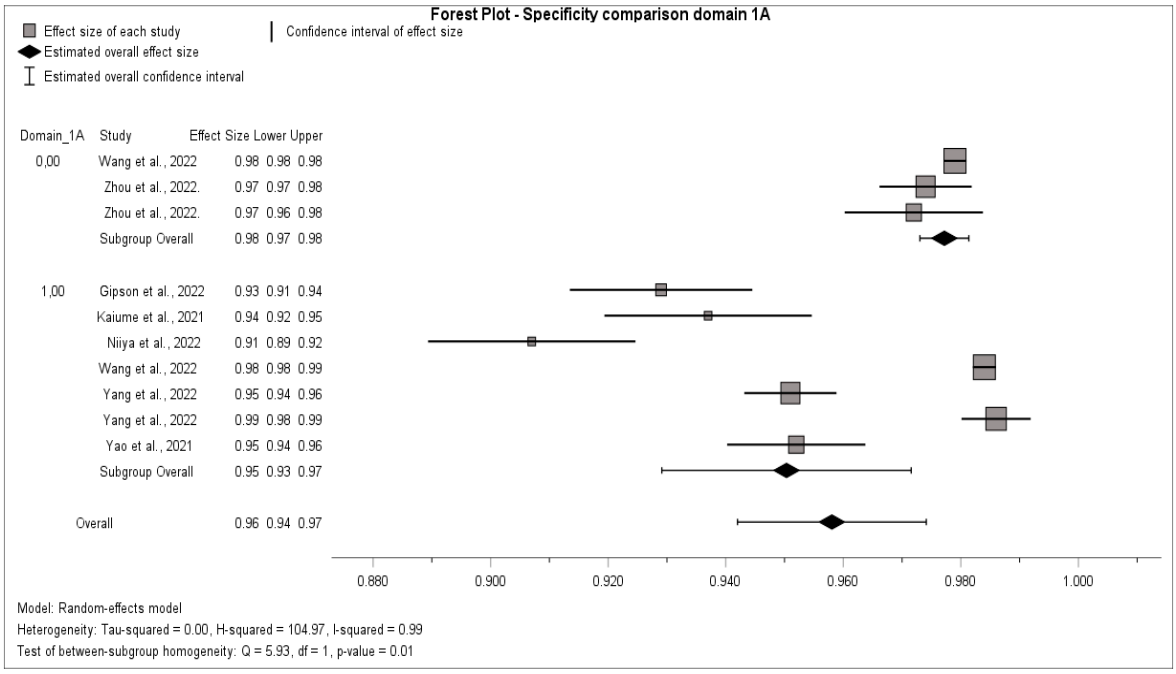


Figure S4.9: Forest plot of the specificity, comparison on domain 1B [19,21,22,24–26,28].

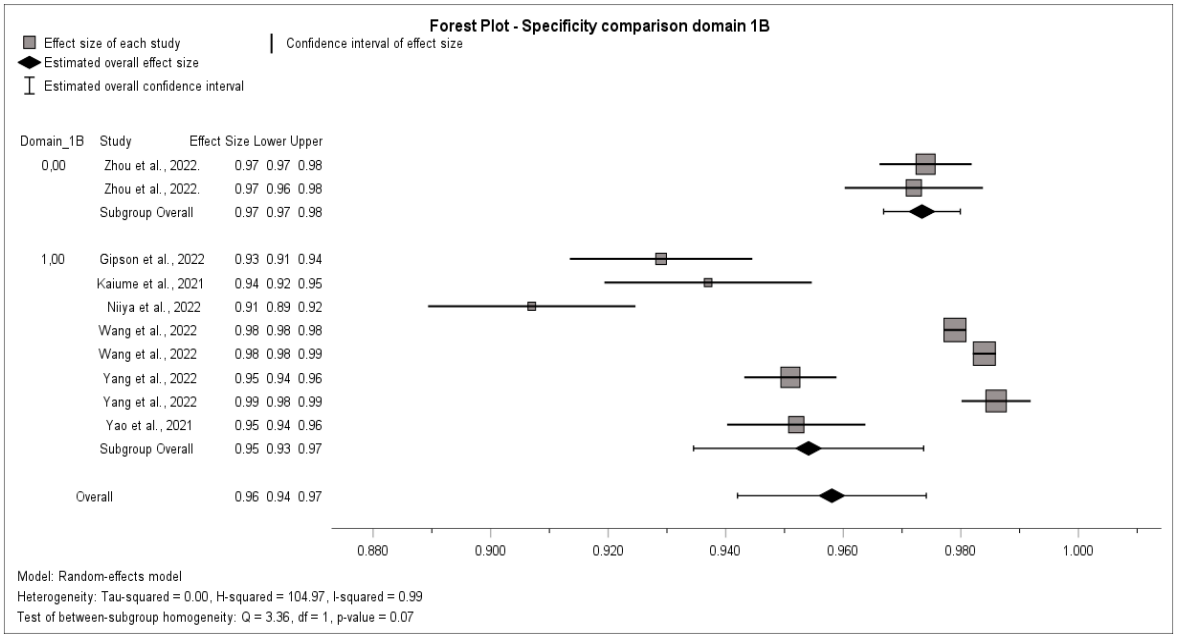


Figure S4.10: Forest plot of the specificity, comparison on domain 2A [19,21,22,24–26,28].

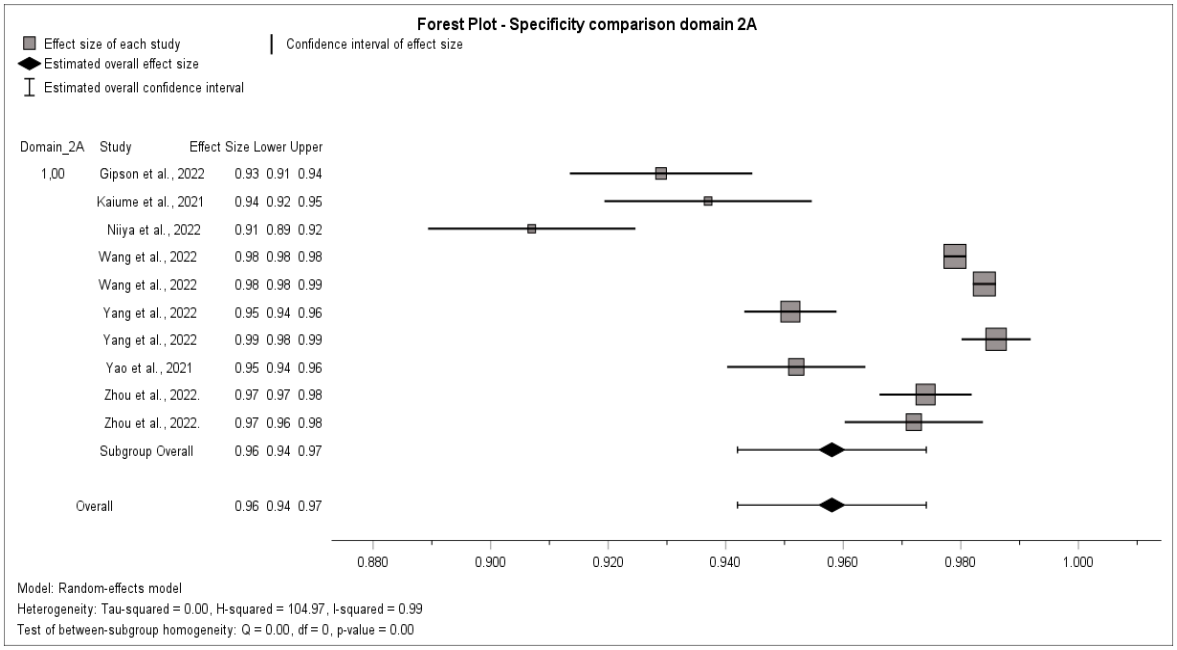


Figure S4.11: Forest plot of the specificity, comparison on domain 2B [19,21,22,24–26,28].

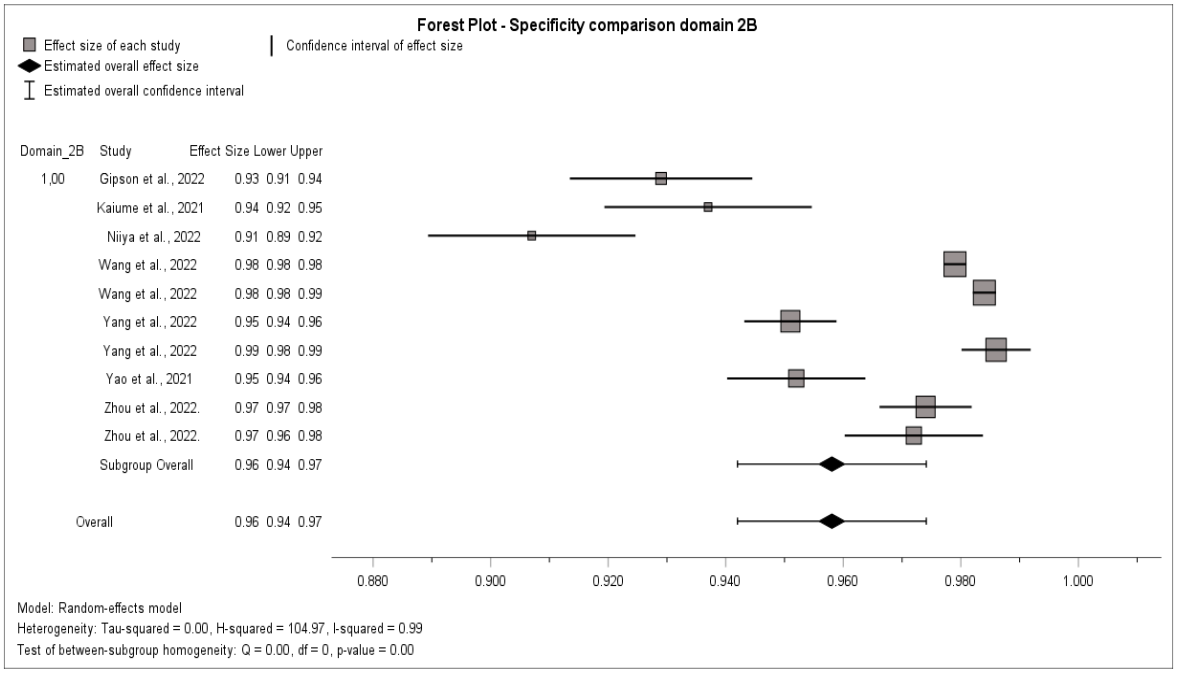


Figure S4.12: Forest plot of the specificity, comparison on domain 3A [19,21,22,24–26,28].

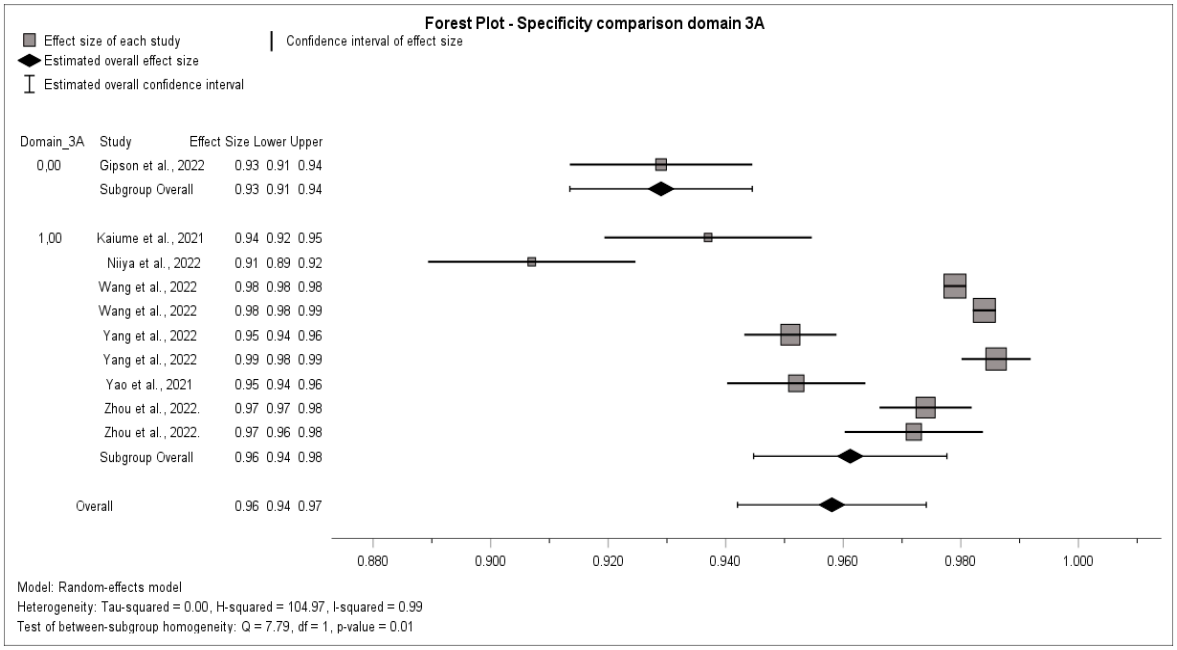


Figure S4.13: Forest plot of the specificity, comparison on domain 3B [19,21,22,24–26,28].

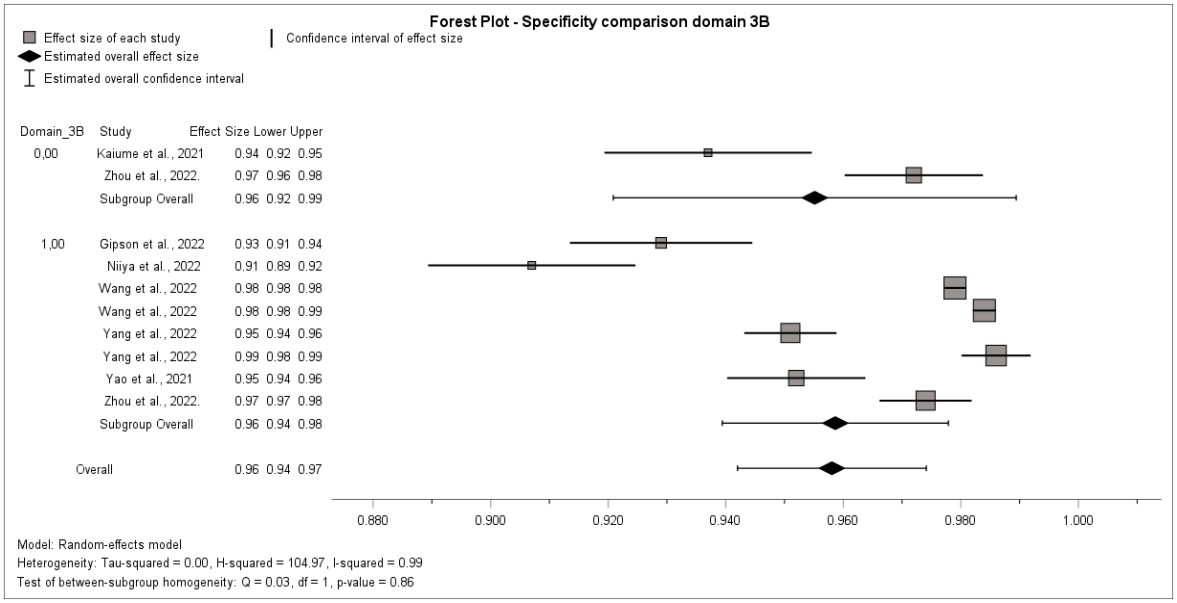


Figure S4.14: Forest plot of the specificity, comparison on domain 4 [19,21,22,24–26,28].

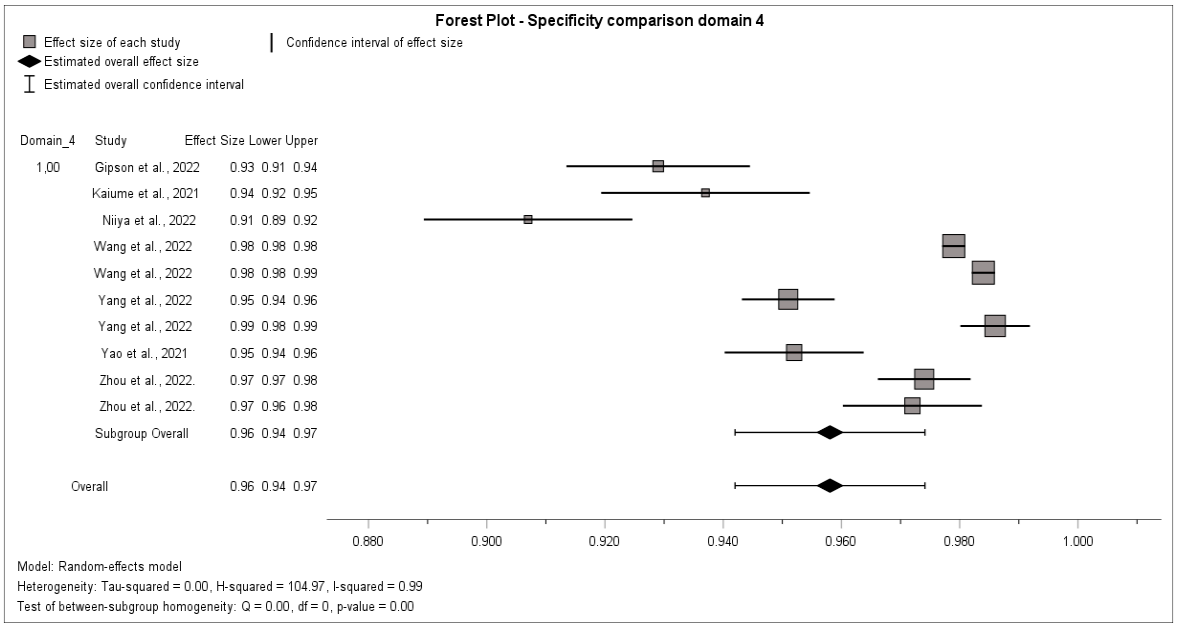


Figure S4.15: Forest plot of the sensitivity, comparison based on quality score [18–28].