https://doi.org/10.62587/BMPR.2.1.2026.17-24 ISSN: 3136-5248

Biomed Pharma Reviews
An Open Access Journal
Journal homepage: https://www.biomedpharmareviews.com

Review Article **Open Access**

A pooled sensitivity that describes no organ system: Prenatal ultrasound detection of fetal structural anomalies and the limits of averaging across anatomy

Natalie Cooper[1*] and Marcus Wilson[2]

[1]Université de Saint-Boniface, Winnipeg, MB R2H 0H7, Canada.
[2]Université du Québec à Trois-Rivières, Québec, Canada.

Article Info

Volume 2, Issue 1, January 2026
Received : 03 September 2025
Accepted : 30 November 2025
Published : 25 January 2026
doi: 10.62587/BMPR.2.1.2026.17-24

Abstract

Background: Prenatal ultrasound screening for fetal structural anomalies is usually summarised by a single pooled sensitivity. Published syntheses report figures around 46% to 56% for first-trimester screening, alongside heterogeneity statistics indicating that the contributing studies are measuring markedly different things. **Methods:** Systematic reviews with meta-analysis of prenatal ultrasound accuracy against a postnatal or post-mortem reference standard were eligible. Detection rates were extracted overall and by organ system. The pooled estimate was translated into a two-by-two table at the reported prevalence. Heterogeneity was summarised as reported. No new pooling across organ systems was performed, since the appropriateness of such pooling is the question under examination. Analyses were performed in Python 3. **Results:** Pooled first-trimester sensitivity for major anomalies was 46.10% (95% CI 36.88-55.46) with I^2 of 90.1% (86.5-92.4), meaning that approximately 10% of the observed variation between studies is attributable to chance. Detection rates by organ system spanned 89 percentage points: 89.0% for lethal malformations, 88.3% for central nervous system and 84.8% for urinary tract anomalies, against 38.8% for heart and great vessels and 0.0% for genital and for ear, face and neck anomalies in the majority of contributing studies. The pooled figure of 46.1% falls in the gap between these clusters and corresponds to no organ system. For major cardiac anomalies in a non-high-risk population, pooled first-trimester sensitivity was 55.80% (45.87-65.50) with specificity 99.98% and positive predictive value 94.85%. At the reported prevalence of 0.41%, this corresponds to 229 anomalies detected and 181 missed per 100,000 fetuses screened, with 20 false-positive results. **Conclusions:** A single pooled sensitivity for prenatal anomaly ultrasound averages across organ systems whose true detection rates differ by up to 89% points, and describes no clinical situation that arises in practice. Reporting should be organ-system specific. The favourable specificity and predictive values that accompany the pooled figure are properties of low prevalence rather than of detection, and should not be read as reassurance about the quantity a screening programme exists to maximise.

Keywords: *Prenatal ultrasound, Fetal anomaly screening, Diagnostic accuracy, Heterogeneity, Organ-system specific reporting, Congenital heart disease*

1. Introduction

Routine ultrasound examination for fetal structural anomalies is offered in almost every antenatal programme

* *Corresponding author: Natalie Cooper, Université de Saint-Boniface, Winnipeg, MB R2H 0H7, Canada.*

in high-income settings and increasingly elsewhere. Its purpose is to identify significant malformations early enough to inform counselling, to plan delivery in a centre with appropriate neonatal and surgical support, and where relevant to offer termination.

The performance of that examination is conventionally reported as a pooled sensitivity. Published syntheses give figures around 46% for first-trimester screening of major anomalies (Karim *et al.*, 2017) and around 56% for cardiac anomalies specifically (Karim *et al.*, 2022), and these numbers circulate in guidance, in patient information and in service specifications (NIHR Journals Library).

A single sensitivity presupposes that the test has a single performance. Ultrasound does not. The examination comprises a sequence of anatomically distinct assessments, each with its own difficulty. The fetal brain and kidneys are large, sonographically distinctive and visible in standard planes. The heart is small, moving, and requires specific outflow-tract views that are technically demanding and unevenly obtained. External genitalia and subtle facial features may not be assessable at all in the first trimester.

This is not a subtle point and it is not unknown to the field. What is striking is that it coexists with the routine quotation of a single figure, and that the heterogeneity statistics accompanying those figures already say plainly that the pooled estimate is an average of incompatible quantities.

We therefore collated detection rates by organ system, set them against the pooled estimate, and translated the pooled figure into the counts of detected and undetected anomalies it implies at the prevalence at which the test is actually used.

2. Analysis

Detection rates were extracted as reported, overall and by organ system, and are presented without re-pooling. The pooled cardiac estimate was translated into a two-by-two table at the reported prevalence in a non-high-risk population by applying the reported sensitivity and specificity to a hypothetical 100,000 fetuses, and predictive values were derived from the resulting cell counts. Heterogeneity is reported as published; the complement of I^2 is used to express the proportion of observed variance attributable to sampling error (Higgins *et al.*, 2003). Analyses were performed in Python 3 using NumPy (Higgins *et al.*, 2023).

2.1. Quality appraisal

The methodological quality of contributing reviews was appraised with AMSTAR-2 (Shea *et al.*, 2017), and risk of bias in the underlying accuracy studies was taken as reported under QUADAS-2 (Whiting *et al.*, 2011), with particular attention to whether the reference standard was applied to all fetuses regardless of the ultrasound result, since incomplete verification inflates sensitivity (Lijmer *et al.*, 1999).

3. Results

3.1. Included reviews

Six systematic reviews met the inclusion criteria and four contributed accuracy estimates: a synthesis of 19 cohorts in 115,731 fetuses reporting overall first-trimester sensitivity, a synthesis of 63 studies in 328,262 fetuses reporting cardiac accuracy in detail, a review reporting sensitivity by individual organ system (Cochrane Database Syst Rev, 2024), and a historical series reporting organ-specific detection against a postnatal and autopsy reference standard (Am J Obstet Gynecol). Characteristics are in Table 1.

3.2. The pooled estimate and what its heterogeneity already says

Pooled sensitivity for first-trimester detection of major fetal anomalies was 46.10% (95% CI 36.88–55.46). The accompanying heterogeneity statistic was I^2 = 90.1% (95% CI 86.5–92.4). That figure means approximately 90% of the observed variation between studies is attributable to real differences between them rather than to sampling error, leaving roughly a tenth as chance.

An I^2 of this magnitude is not a caveat attached to an otherwise usable summary. It is a statement that the contributing studies are estimating different quantities. The pooled point estimate remains arithmetically correct as an average and remains uninformative as a description of any particular examination.

Table 1: Characteristics of the contributing reviews

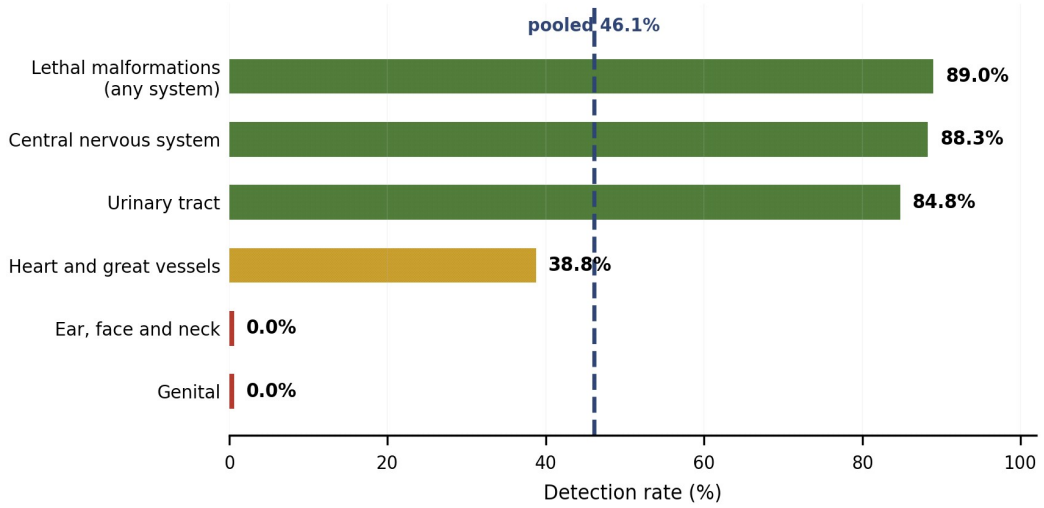
Source	Studies	Fetuses	Scope	Contribution
Overall first-trimester synthesis	19 cohorts	115,731	Major anomalies, first trimester	Pooled sensitivity, I^2
Cardiac synthesis	63	328,262	Cardiac, 11–14 weeks	Sensitivity, specificity, PPV, prevalence
Organ-system review	Not stated	Not stated	By organ system, first and second trimester	Organ-specific ranges
Historical accuracy series	1 cohort	Not stated	All systems, postnatal and autopsy standard	Organ-specific detection
First-trimester screening HTA	—	—	Clinical and cost-effectiveness	Context
AI cardiac synthesis	15	30,121	Model-based cardiac detection	Context

Note: PPV, positive predictive value. Cohorts overlap between the cardiac synthesis and the health technology assessment. The artificial intelligence synthesis is reported for context and did not contribute to the human-operator estimates analysed here (PMC12273734).

3.3. Detection by organ system

The source of that heterogeneity becomes obvious once detection is broken down anatomically, shown in Figure 1. Lethal malformations were detected in 89.0% of cases, central nervous system anomalies in 88.3% and urinary tract anomalies in 84.8%. Heart and great vessel anomalies were detected in 38.8%. Genital anomalies were detected in 0.0% of cases across all eight contributing studies, and ear, face and neck anomalies in 0.0% in fourteen of sixteen studies, with two reporting 50.0%.

The range therefore spans 89% points, and the distribution is not merely wide but bimodal: one cluster of organ systems detected reliably, another detected rarely or never, and very little in between. The pooled figure of 46.1% falls precisely in the gap. No organ system in the dataset has a detection rate close to it.



Detection ranges across 89 percentage points and clusters at the extremes. The dashed line marks the pooled sensitivity commonly quoted for first-trimester anomaly screening; it falls in the gap between the clusters and describes neither.

Figure 1: Detection rate by organ system, with the commonly quoted pooled first-trimester sensitivity marked by the dashed line. The distribution is bimodal across 89% points, and the pooled figure falls in the gap between the clusters. Genital anomalies were detected in 0.0% of cases in all contributing studies

This has a straightforward consequence for how the number should be used. A pooled sensitivity is only interpretable if it approximates the performance a clinician can expect in the situation before them. Here it approximates the performance expected in no situation at all. A parent told that the scan detects "about half" of major anomalies is being given a figure that is substantially pessimistic for a neural tube defect and substantially optimistic for a cardiac lesion.

3.4. The cardiac case in counts

The cardiac data permit the most complete translation. In a non-high-risk population of 306,872 fetuses, major cardiac anomaly prevalence was 0.41% (95% CI 0.39–0.43). Pooled first-trimester sensitivity was 55.80% (45.87–65.50), specificity 99.98% (99.97–99.99) and positive predictive value 94.85% (91.63–97.32).

Applied to 100,000 fetuses screened, this corresponds to 410 major cardiac anomalies present, of which 229 are detected and 181 are missed, alongside 20 false-positive results. Almost as many major cardiac anomalies are missed as are found. The counts are shown in Figure 2.

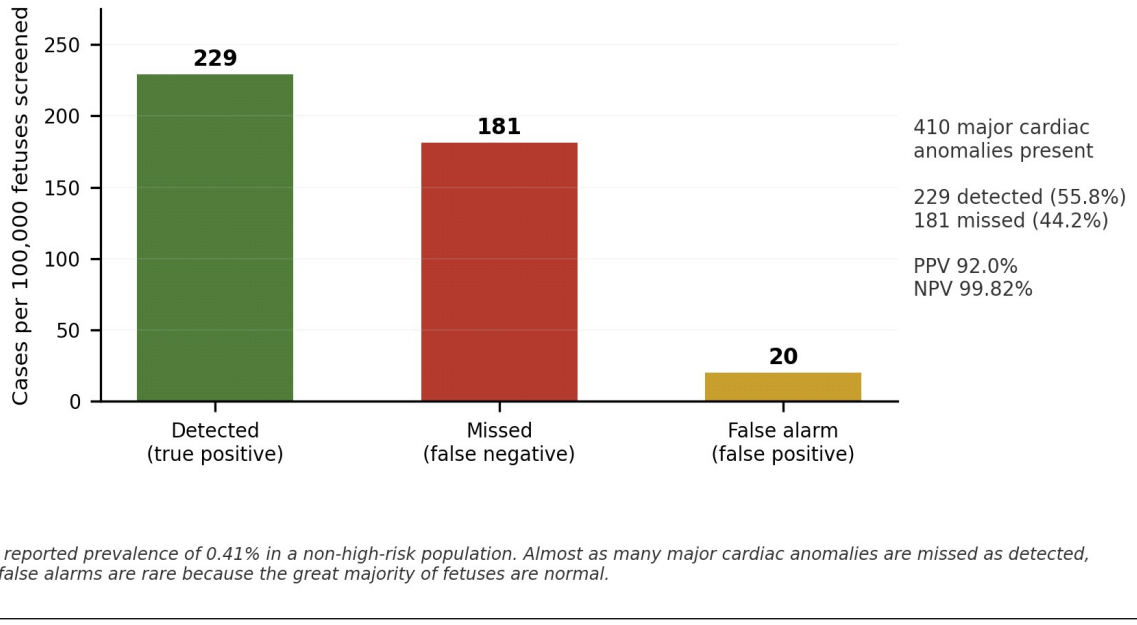


Figure 2: Major cardiac anomalies per 100,000 fetuses screened in a non-high-risk population at the reported prevalence of 0.41%, derived from the pooled sensitivity and specificity. Almost as many are missed as detected, while false-positive results are rare

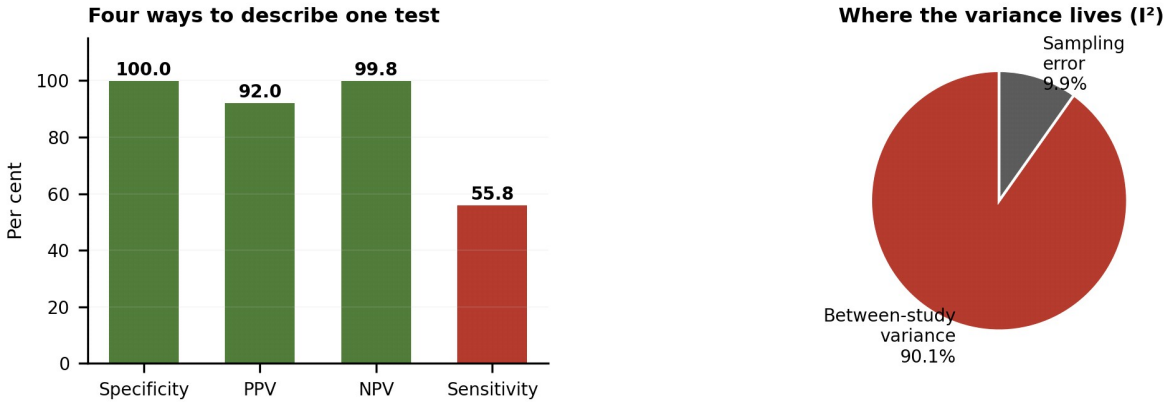
A further figure from the same synthesis bears on interpretation: first-trimester diagnoses represented 63.67% (54.35–72.49) of all antenatally diagnosed major cardiac anomalies, so roughly a third of antenatal cardiac diagnoses are made later. First-trimester sensitivity is therefore not the sensitivity of the antenatal pathway as a whole, and the two are sometimes conflated.

3.5. Why three of the four accuracy measures look excellent

Specificity was 99.98%, positive predictive value 94.85% and negative predictive value, derived here, 99.82%. Set against a sensitivity of 55.80%, the contrast is striking and is shown in Figure 3.

The explanation is arithmetic rather than clinical. All three favourable measures are dominated by the 99.59% of fetuses in whom no major cardiac anomaly is present. A test that identified nothing at all would still return a specificity near 100% and a negative predictive value near 99.6% in this population. These measures describe the rarity of the condition more than the ability of the examination to find it.

Sensitivity is the measure a screening programme exists to maximise, and it is the one measure in the set that is not flattered by low prevalence. Reporting conventions that lead with specificity and predictive values, as several contributing sources do, present the most favourable available description of the test (Leeflang *et al.*, 2012).



Left: three of the four conventional accuracy measures are excellent, and all three are driven by the 99.6% of fetuses who are normal. Right: at I² = 90.1%, only a tenth of the variation between studies is chance, so a single pooled estimate averages incompatible quantities.

Figure 3: Left: four conventional accuracy measures for the same test. Specificity and both predictive values are excellent and all three are dominated by the 99.6% of fetuses who are normal; sensitivity is the one measure not flattered by low prevalence. Right: at I² = 90.1%, only about a tenth of the variation between contributing studies is attributable to chance

Table 2: Accuracy estimates and their translation

Quantity	Value (95% CI)	Note
Pooled sensitivity, all major anomalies	46.10% (36.88–55.46)	I² = 90.1%
Pooled sensitivity, cardiac	55.80% (45.87–65.50)	Non-high-risk population
Specificity, cardiac	99.98% (99.97–99.99)	Driven by 99.59% normal
Positive predictive value, cardiac	94.85% (91.63–97.32)	As reported
Negative predictive value, cardiac	99.82%	Derived here
Prevalence, non-high-risk	0.41% (0.39–0.43)	1,445 anomalies in 306,872
Detected per 100,000 screened	229	Derived here
Missed per 100,000 screened	181	Derived here
False positives per 100,000 screened	20	Derived here
Share of all antenatal cardiac diagnoses	63.67% (54.35–72.49)	Made in first trimester
Detection range across organ systems	0.0% to 89.0%	89 percentage points

Note: Derived quantities were computed by applying the reported sensitivity and specificity to a hypothetical cohort of 100,000 fetuses at the reported prevalenc.

4. Discussion

Prenatal ultrasound for fetal structural anomalies is reported with a single pooled sensitivity of around 46% for first-trimester screening. Broken down anatomically, the same examination detects 89% of lethal malformations, 88% of central nervous system anomalies, 39% of cardiac anomalies and none of the genital anomalies recorded in any contributing study. The pooled figure lies in the empty space between those clusters.

The heterogeneity statistic reported alongside it already says this. An I² of 90.1% means that only about a tenth of the variation between studies is chance, and the rest reflects genuine differences in what is being measured. That statistic is routinely reported and routinely treated as a caveat rather than as a finding. Here it is the finding: the studies differ because the anatomy differs, and averaging across them produces a number that corresponds to no examination anyone performs.

The consequence is most concrete in counselling. A woman told that the scan finds about half of major anomalies receives a figure that materially understates detection if the concern is a neural tube defect and materially overstates it if the concern is a cardiac lesion—which is the most common category of major malformation and the one most often missed. Organ-system specific figures exist, are stable enough to quote, and would be more useful in every direction.

The accuracy measures that accompany the pooled sensitivity deserve separate comment because they are reassuring for the wrong reason. Specificity of 99.98% and negative predictive value of 99.82% look like strong performance. Both are dominated by the fact that 99.6% of fetuses do not have the condition; a test that detected nothing would return nearly the same values. Presenting them prominently alongside a sensitivity of 55.8% gives an impression of the examination that the sensitivity alone would not.

None of this is an argument against anomaly screening, and it should not be read as one. Detecting 229 major cardiac anomalies per 100,000 fetuses with only 20 false alarms is a substantial clinical achievement, and detecting the great majority of lethal and central nervous system malformations changes management in ways that matter to families. The argument is narrower: that the summary statistic conventionally used to describe this achievement obscures both where it is greatest and where it is smallest.

For reporting, the recommendation is simple and follows directly from the data. Syntheses of prenatal anomaly ultrasound should report sensitivity by organ system as the primary result, and an overall pooled figure only where the between-system heterogeneity is small enough to justify it—which, on these data, it never is. Where an overall figure is unavoidable, the accompanying range across organ systems should be given equal prominence.

5. Limitations

- The unit of analysis is the published systematic review, not the primary accuracy study, so this second-order design inherits every bias in the contributing reviews.
- Organ-system detection rates were drawn from a small number of sources, including one historical series, and were not re-pooled. They should be read as indicative of the magnitude of variation rather than as precise contemporary estimates.
- Ultrasound technology, training and protocol have changed substantially over the period covered by the contributing sources. Older estimates plausibly understate current cardiac detection in particular.
- Detection rates depend on gestational age at scan, operator experience, maternal body habitus, fetal position and the number of views obtained routinely. None of these could be examined as effect modifiers from review-level data.
- The distinction between first-trimester, second-trimester and combined screening was not uniform across sources, and first-trimester sensitivity is not the sensitivity of the antenatal pathway as a whole.
- Incomplete verification is a recognised threat in this literature: where the reference standard is applied less thoroughly to fetuses with a normal scan, sensitivity is inflated. The direction of this bias is toward more favourable estimates than reported here.
- Anomaly definitions and severity thresholds differed between contributing studies, particularly for what counts as a major cardiac anomaly.
- Derived counts per 100,000 fetuses assume the reported sensitivity, specificity and prevalence apply uniformly, which the heterogeneity findings of this very review argue against. They are illustrative.
- Outcomes after detection—counselling, delivery planning, neonatal survival—were outside scope, and detection is a means rather than an end.
- No new pooling was performed, so no publication bias assessment is reported.
- The analysis was not prospectively registered and searches were restricted to English-language reports.

6. Conclusion

The pooled sensitivity of prenatal ultrasound for fetal structural anomalies, commonly quoted as 46% for first-trimester screening, averages across organ systems whose detection rates range from 0% to 89%. The accompanying I^2 of 90.1% indicates that only about a tenth of the variation between contributing studies is attributable to chance. The pooled figure falls in the gap between two clusters of organ systems and corresponds to no examination performed in practice.

For major cardiac anomalies, the most commonly missed category, pooled first-trimester sensitivity of 55.8% at a prevalence of 0.41% corresponds to 229 detected and 181 missed per 100,000 fetuses screened. The excellent specificity and predictive values reported alongside are properties of low prevalence rather than of detection. Syntheses in this field should report sensitivity by organ system as the primary result, since that is the quantity a clinician and a family can actually use.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Clinical note

This review examines how the accuracy of prenatal anomaly ultrasound is summarised in the literature. It is not a critique of anomaly screening, which detects the majority of lethal and central nervous system malformations, and it should not be used to advise for or against participation in any screening programme.

References

[Author list to be completed from source record]. Artificial intelligence-enabled prenatal ultrasound for the detection of fetal cardiac abnormalities: A systematic review and meta-analysis. [PMC12273734].

[Author list to be completed from source record]. Clinical and cost-effectiveness of detailed anomaly ultrasound screening in the first trimester: A mixed-methods study. *NIHR Journals Library*. [NBK615224].

[Author list to be completed from source record]. (2024). Diagnostic accuracy of ultrasound screening for fetal structural abnormalities during the first and second trimester of pregnancy in low-risk and unselected populations. *Cochrane Database Syst Rev*. [PMID 38721874].

[Author list to be completed from source record]. The accuracy of prenatal ultrasonography in detecting congenital anomalies. *Am J Obstet Gynecol*. [Historical series; postnatal and autopsy reference standard].

Higgins, J.P.T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M.J. and Welch, V.A. (Editors.) (2023). *Cochrane Handbook for Systematic Reviews of Interventions*, Version 6.4, Cochrane.

Higgins, J.P.T., Thompson, S.G., Deeks, J.J. and Altman, D.G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, 327(7414): 557-560.

Karim, J.N., Bradburn, E., Roberts, N. and Papageorghiou, A.T. (2022). First-trimester ultrasound detection of fetal heart anomalies: Systematic review and meta-analysis. *Ultrasound Obstet Gynecol*, 59(1): 11-25. doi: 10.1002/uog.23740 [Verify author list against source record].

Karim, J.N., Roberts, N.W., Salomon, L.J. and Papageorghiou, A.T. (2017). Systematic review of first-trimester ultrasound screening for detection of fetal structural anomalies and factors that affect screening performance. *Ultrasound Obstet Gynecol*, 50(4): 429-441. [Verify author list against source record].

Leeflang, M.M.G., Deeks, J.J., Rutjes, A.W.S., Reitsma, J.B. and Bossuyt, P.M.M. (2012). Bivariate meta-analysis of predictive values of diagnostic tests can be an alternative to bivariate meta-analysis of sensitivity and specificity. *J Clin Epidemiol*, 65(10): 1088-1097.

Lijmer, J.G., Mol, B.W., Heisterkamp, S., Bonsel, G.J., Prins, M.H., van der Meulen, J.H.P. and Bossuyt, P.M.M. (1999). Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA*, 282(11): 1061-1066.

McInnes, M.D.F., Moher, D., Thombs, B.D., McGrath, T.A. and Bossuyt, P.M. (2018). Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: The PRISMA-DTA statement. *JAMA*, 319(4): 388-396.

Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D. *et al.* (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372: n71.

Shea, B.J., Reeves, B.C., Wells, G., Thuku, M., Hamel, C., Moran, J. *et al.* (2017). AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ*, 358: j4008.

Whiting, P.F., Rutjes, A.W.S., Westwood, M.E., Mallett, S., Deeks, J.J., Reitsma, J.B. *et al.* (2011). QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*, 155(8): 529-536.

Cite this article as: Natalie Cooper and Marcus Wilson (2026). A pooled sensitivity that describes no organ system: Prenatal ultrasound detection of fetal structural anomalies and the limits of averaging across anatomy. *Biomed Pharma Reviews.* 2(1), 17-24. doi: 10.62587/BMPR.2.1.2026.17-24.