Supplemental materials for:

Summanwar D, Fowler NR, Hammers DB, Perkins AJ, Brosch JR, Willis DR. Agile implementation of a digital cognitive assessment for dementia in primary care. *Ann Fam Med*. 2025;online. doi: https://doi.org/10.1370/afm.240294

Supplemental Appendix 1.

Section & Topic	No	Item	Reported on page #
TITLE OR ABSTRACT			
	1	Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)	2
ABSTRACT			
	2	Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)	2
INTRODUCTION			
	3	Scientific and clinical background, including the intended use and clinical role of the index test	3-4
	4	Study objectives and hypotheses	4
METHODS			
Study design	**5**	Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)	4-5
Participants	**6**	Eligibility criteria	4
	7	On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)	4
	8	Where and when potentially eligible participants were identified (setting, location and dates)	4
	9	Whether participants formed a consecutive, random or convenience series	4
Test methods	**10a**	Index test, in sufficient detail to allow replication	4-5
	10b	Reference standard, in sufficient detail to allow replication	5-6
	11	Rationale for choosing the reference standard (if alternatives exist)	5-6
	12a	Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory	7
	12b	Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory	5-6
	13a	Whether clinical information and reference standard results were available to the performers/readers of the index test	5
	13b	Whether clinical information and index test results were available to the assessors of the reference standard	5
Analysis	**14**	Methods for estimating or comparing measures of diagnostic accuracy	N/A
	15	How indeterminate index test or reference standard results were handled	7
	16	How missing data on the index test and reference standard were handled	7
	17	Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory	N/A
	18	Intended sample size and how it was determined	N/A
RESULTS			
Participants	**19**	Flow of participants, using a diagram	22
	20	Baseline demographic and clinical characteristics of participants	18
	21a	Distribution of severity of disease in those with the target condition	18, 19
	21b	Distribution of alternative diagnoses in those without the target condition	N/A
	22	Time interval and any clinical interventions between index test and reference standard	5
Test results	**23**	Cross tabulation of the index test results (or their distribution) by the results of the reference standard	24-25
	24	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	20, 24, 25
	25	Any adverse events from performing the index test or the reference standard	N/A
DISCUSSION			
	26	Study limitations, including sources of potential bias, statistical uncertainty, and generalisability	11-12
	27	Implications for practice, including the intended use and clinical role of the index test	12
OTHER INFORMATION			
	28	Registration number and name of registry	N/A
	29	Where the full study protocol can be accessed	N/A
	30	Sources of funding and other support; role of funders	12-13

STARD 2015

AIM

STARD stands for "Standards for Reporting Diagnostic accuracy studies". This list of items was developed to contribute to the completeness and transparency of reporting of diagnostic accuracy studies. Authors can use the list to write informative study reports. Editors and peer-reviewers can use it to evaluate whether the information has been included in manuscripts submitted for publication.

EXPLANATION

A **diagnostic accuracy study** evaluates the ability of one or more medical tests to correctly classify study participants as having a **target condition.** This can be a disease, a disease stage, response or benefit from therapy, or an event or condition in the future. A medical test can be an imaging procedure, a laboratory test, elements from history and physical examination, a combination of these, or any other method for collecting information about the current health status of a patient.

The test whose accuracy is evaluated is called **index test.** A study can evaluate the accuracy of one or more index tests. Evaluating the ability of a medical test to correctly classify patients is typically done by comparing the distribution of the index test results with those of the **reference standard**. The reference standard is the best available method for establishing the presence or absence of the target condition. An accuracy study can rely on one or more reference standards.

If test results are categorized as either positive or negative, the cross tabulation of the index test results against those of the reference standard can be used to estimate the **sensitivity** of the index test (the proportion of participants *with* the target condition who have a positive index test), and its **specificity** (the proportion *without* the target condition who have a negative index test). From this cross tabulation (sometimes referred to as the contingency or "2x2" table), several other accuracy statistics can be estimated, such as the positive and negative **predictive values** of the test. Confidence intervals around estimates of accuracy can then be calculated to quantify the statistical **precision** of the measurements.

If the index test results can take more than two values, categorization of test results as positive or negative requires a **test positivity cut-off**. When multiple such cut-offs can be defined, authors can report a receiver operating characteristic (ROC) curve which graphically represents the combination of sensitivity and specificity for each possible test positivity cut-off. The **area under the ROC curve** informs in a single numerical value about the overall diagnostic accuracy of the index test.

The **intended use** of a medical test can be diagnosis, screening, staging, monitoring, surveillance, prediction or prognosis. The **clinical role** of a test explains its position relative to existing tests in the clinical pathway. A replacement test, for example, replaces an existing test. A triage test is used before an existing test; an add-on test is used after an existing test.

Besides diagnostic accuracy, several other outcomes and statistics may be relevant in the evaluation of medical tests. Medical tests can also be used to classify patients for purposes other than diagnosis, such as staging or prognosis. The STARD list was not explicitly developed for these other outcomes, statistics, and study types, although most STARD items would still apply.

DEVELOPMENT

This STARD list was released in 2015. The 30 items were identified by an international expert group of methodologists, researchers, and editors. The guiding principle in the development of STARD was to select items that, when reported, would help readers to judge the potential for bias in the study, to appraise the applicability of the study findings and the validity of conclusions and recommendations. The list represents an update of the first version, which was published in 2003.

More information can be found on http://www.equator-network.org/reporting-guidelines/stard.

SUPPLEMENTAL TABLES & FIGURES

Contents

SUPPLEMENTAL TABLES

Summary: In the primary analysis presented in the manuscript, we selected an LVEF cut-off value of <50% based on the recommended criteria for defining heart failure with associated LV systolic dysfunction, as outlined in the most recent universal definition and classification of heart failure. The classification of LV systolic dysfunction includes heart failure with mildly reduced ejection fraction (HFmrEF: LVEF 41 – 49%) and heart failure with reduced ejection fraction (HFrEF: LVEF < 40%). The diagnostic cut-off value for peripartum cardiomyopathy falls within the heart failure category, HFmrEF. It is important to note that the goal of this study and the intended use of the screening tool is not to diagnose peripartum cardiomyopathy alone, but rather to identify all forms of LV systolic dysfunction that may be pre-existing (prior to conception) or occur de novo in the perinatal period.

In the supplemental tables below, we present sensitivity analysis for the performance of the AI-ECG and AI-Stethoscope for detection of LVEF < 45% (cut-of value for peripartum cardiomyopathy), Table S1. In Table S2, we present a sensitivity analysis that includes missing values/poor-quality recordings for the digital stethoscope recordings; to be conservative, we assumed missing or poor-quality data to be negative screens.

Supplemental Table 1. Sensitivity Analysis – Performance of the AI-ECG and AI-Stethoscope for detection of LVEF <45%

	N	AUC	Sensitivity	Specificity	Accuracy	F1 Score	Negative Predictive Value	Odds Ratio	Positive Predictive Value
12-lead ECG (LVEF < 45%)									
AI-ECG	100	0.953 (0.900, 1.000)	50.0% (6.8%, 93.2%)	95.8% (89.7%, 98.9%)	94.0% (87.4%, 97.8%)	40.0	97.9% (92.5%, 99.7%)	23.0 (2.5, 207.7)	33.3% (4.3%, 77.7%)
Digital Stethoscope ECG+PCG (LVEF < 45%)									
Angled	99	0.983 (0.955, 1.000)	80.0% (28.4%, 99.5%)	93.6% (86.6%, 97.6%)	92.9% (86.0%, 97.1%)	53.3	98.9% (93.9%, 100.0%)	58.7 (5.6, 610.4)	40.0% (12.2%, 73.8%)
Subclavicular	96	0.857 (0.672, 1.000)	60.0% (14.7%, 94.7%)	90.1% (82.1%, 95.4%)	88.5% (80.4%, 94.1%)	35.3	97.6% (91.7%, 99.7%)	13.7 (2.0, 92.9)	25.0% (5.5%, 57.2%)
V2	100	0.949 (0.871, 1.000)	80.0% (28.4%, 99.5%)	91.6% (84.1%, 96.3%)	91.0% (83.6%, 95.8%)	47.1	98.9% (93.8%, 100.0%)	43.5 (4.3, 437.3)	33.3% (9.9%, 65.1%)
Mean Prediction	100	0.971 (0.929, 1.000)	60.0% (14.7%, 94.7%)	93.7% (86.8%, 97.6%)	92.0% (84.8%, 96.5%)	42.9	97.8% (92.3%, 99.7%)	22.2 (3.1, 159.7)	33.3% (7.5%, 70.1%)
Maximum Prediction	100	0.979 (0.950, 1.000)	100.0% (47.8%, 100.0%)	82.1% (72.9%, 89.2%)	83.0% (74.2%, 89.8%)	37.0	100.0% (95.4%, 100.0%)	49.3 (2.6, 934.3)	22.7% (7.8%, 45.4%)

Abbreviations: AUC- area under the curve, ECG- electrocardiogram, PCG- phonocardiogram

*Results provided represent available AI prediction results based on available diagnostic quality ECG/phonocardiogram recordings. Missing or recorded ECG/phonocardiogram data deemed to be of poor quality were excluded from the analysis resulting in a sample size less than 100 for some of the digital stethoscope recording locations as shown in the table.

Supplemental Table 2. Sensitivity Analysis – Performance of the AI-ECG and AI-Stethoscope for detection of LVEF <50% (including missing data)

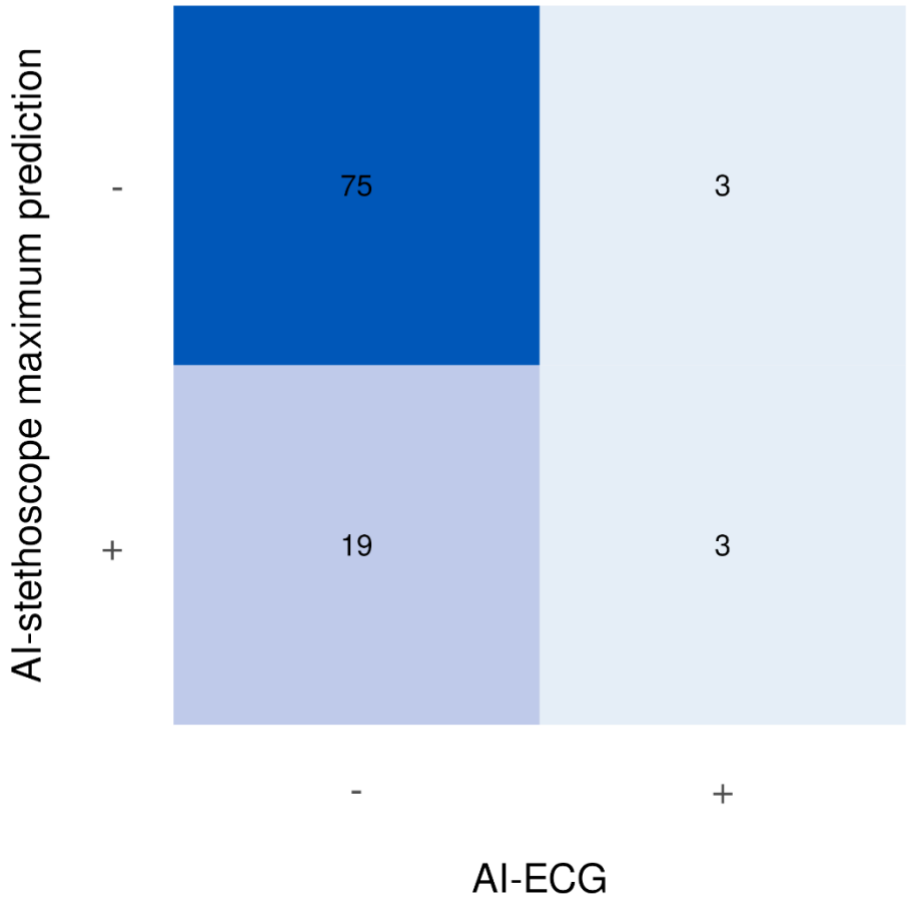
	N	Sensitivity	Specificity	Accuracy	F1 Score	Negative Predictive Value	Odds Ratio	Positive Predictive Value
12-lead ECG (LVEF < 45%)								
AI-ECG	100	40.0% (5.3%, 85.3%)	95.8% (89.6%, 98.8%)	93.0% (86.1%, 97.1%)	36.4	96.8% (91.0%, 99.3%)	15.2 (2.0, 117.9)	33.3% (4.3%, 77.7%)
Digital Stethoscope ECG+PCG (LVEF < 45%)								
Angled	100	80.0% (28.4%, 99.5%)	93.7% (86.8%, 97.6%)	93.0% (86.1%, 97.1%)	53.3	98.9% (94.0%, 100.0%)	59.3 (5.7, 617.2)	40.0% (12.2%, 73.8%)
Subclavicular	100	60.0% (14.7%, 94.7%)	90.5% (82.8%, 95.6%)	89.0% (81.2%, 94.4%)	35.3	97.7% (92.0%, 99.7%)	14.3 (2.1, 97.4)	25.0% (5.5%, 57.2%)
V2	100	80.0% (28.4%, 99.5%)	91.6% (84.1%, 96.3%)	91.0% (83.6%, 95.8%)	47.1	98.9% (93.8%, 100.0%)	43.5 (4.3, 437.3)	33.3% (9.9%, 65.1%)
Mean Prediction	100	60.0% (14.7%, 94.7%)	93.7% (86.8%, 97.6%)	92.0% (84.8%, 96.5%)	42.9	97.8% (92.3%, 99.7%)	22.2 (3.1, 159.7)	33.3% (7.5%, 70.1%)
Maximum Prediction	100	100.0% (47.8%, 100.0%)	82.1% (72.9%, 89.2%)	83.0% (74.2%, 89.8%)	37.0	100.0% (95.4%, 100.0%)	49.3 (2.6, 934.3)	22.7% (7.8%, 45.4%)

Abbreviations: ECG- electrocardiogram, PCG- phonocardiogram

*Results provided represent all available AI prediction results based on diagnostic quality ECG/phonocardiogram recordings in addition to missing/poor-quality recordings. Missing or recorded ECG/phonocardiogram data deemed to be of poor quality were conservatively assumed to be negative, as such AUC values were not computed.

SUPPLEMENTAL FIGURES

Supplemental Figure 1. Agreement between AI-ECG predictions and AI-stethoscope maximum predictions



Supplemental Figure 2. Agreement between AI-ECG predictions and AI-stethoscope angled position