Inter-rater and intra-rater reliability and agreement of echocardiographic diagnosis of rheumatic heart disease using the World Heart Federation evidence-based criteria

Bo Remenyi,[1,2] Jonathan Carapetis,[3] John W Stirling,[4] Beatrice Ferreira,[5] Krishnan Kumar,[6] John Lawrenson,[7,8] Eloi Marijon,[9] Mariana Mirabel,[10] A O Mocumbi,[11] Cleonice Mota,[12] John Paar,[13] Anita Saxena,[14] Janet Scheel,[15] Satu Viali,[16] I B Vijayalakshmi,[17] Gavin R Wheaton,[18] Liesl Zuhlke,[19] Karishma Sidhu,[2] Eliazar Dimalapang,[2] Thomas L Gentles,[20] Nigel J Wilson[2,21]

For numbered affiliations see end of article.

Correspondence to
Dr Bo Remenyi, Menzies School of Health Research, Casuarina, NT 0810, Australia; Bo.Remenyi@menzies.edu.au

Received 1 June 2019
Revised 5 June 2019
Accepted 6 June 2019

ABSTRACT

Objective Different definitions have been used for screening for rheumatic heart disease (RHD). This led to the development of the 2012 evidence-based World Heart Federation (WHF) echocardiographic criteria. The objective of this study is to determine the intra-rater and inter-rater reliability and agreement in differentiating no RHD from mild RHD using the WHF echocardiographic criteria.

Methods A standard set of 200 echocardiograms was collated from prior population-based surveys and uploaded for blinded web-based reporting. Fifteen international cardiologists reported on and categorised each echocardiogram as no RHD, borderline or definite RHD. Intra-rater and inter-rater reliability was calculated using Cohen's and Fleiss' free-marginal multirater kappa (κ) statistics, respectively. Agreement assessment was expressed as percentages. Subanalyses assessed reproducibility and agreement parameters in detecting individual components of WHF criteria.

Results Sample size from a statistical standpoint was 3000, based on repeated reporting of the 200 studies. The inter-rater and intra-rater reliability of diagnosing definite RHD was substantial with a kappa of 0.65 and 0.69, respectively. The diagnosis of pathological mitral and aortic regurgitation was reliable and almost perfect, kappa of 0.79 and 0.86, respectively. Agreement for morphological changes of RHD was variable ranging from 0.54 to 0.93 κ.

Conclusions The WHF echocardiographic criteria enable reproducible categorisation of echocardiograms as definite RHD versus no or borderline RHD and hence it would be a suitable tool for screening and monitoring disease progression. The study highlights the strengths and limitations of the WHF echo criteria and provides a platform for future revisions.

Key messages

What is already known about this subject?

- Different definitions have been used for screening for rheumatic heart disease (RHD). This led to the development of the 2012 evidence-based World Heart Federation (WHF) echocardiographic criteria.

What does this study add?

- This study demonstrates that if the WHF echocardiographic criteria are strictly applied to screening echocardiograms, then no RHD can be reliably differentiated from mild RHD. Physiological regurgitation can usually be differentiated from mild pathological regurgitation; however, the agreement over the presence of morphological features is more variable.

How might this impact on clinical practice?

- The WHF echocardiographic criteria enable reproducible categorisation of echocardiograms as no RHD, borderline and definite RHD. The criteria are a suitable tool for RHD screening programmes and can be used in the clinical setting for the undifferentiated valve disease and when a diagnosis of RHD is being considered.

INTRODUCTION

Rheumatic heart disease (RHD), a sequel of acute rheumatic fever (ARF), remains a major global health problem affecting an estimated 33.4 million people worldwide and leads to substantial morbidity and 319 400 deaths per year.[1] ARF may go undetected if symptoms are mild or atypical, patients may not seek medical care or medical staff may not be equipped to make diagnoses. On a global basis, most patients with RHD who seek medical attention do not have a history of ARF.[2]

Asymptomatic patients with mild to moderate RHD likely benefit the most from secondary prophylaxis.[3 4] Auscultation does not have sufficient sensitivity (just 20%) and specificity to be useful in diagnostic testing for RHD and is no longer recommended as a screening tool.[5 6] Echocardiography is the gold standard for the diagnosis of both acute and chronic RHD.[7 8]

To allow for rapid and consistent case identification of patients with mild RHD without a prior history of ARF, in 2012 the evidence-based World Heart Federation (WHF) echocardiographic criteria for RHD were developed (table 1).[7] The criteria

To cite: Remenyi B, Carapetis J, Stirling JW, *et al*. *Heart Asia* 2019;**11**:e011233. doi:10.1136/heartasia-2019-011233

Table 1 2012 WHF criteria for echocardiographic diagnosis of RHD for individuals aged ≤20 years[7]

Definite RHD (either A, B, C or D)	
(A) Pathological MR and at least two morphological features of RHD of the MV	
(B) MS mean gradient ≥4 mm Hg*	
(C) Pathological AR and at least two morphological features of RHD of the AV†	
(D) Borderline disease of both the AV and MV‡	
Borderline RHD (either A, B or C)	
(A) At least two morphological features of RHD of the MV without pathological MR or MS	
(B) Pathological MR	
(C) Pathological AR	
Normal echocardiographic findings (all of A, B, C and D)	
(A) MR that does not meet all four Doppler echocardiographic criteria (physiological MR)	
(B) AR that does not meet all four Doppler echocardiographic criteria (physiological AR)	
(C) An isolated morphological feature of RHD of the MV (eg, valvular thickening) without any associated pathological stenosis or regurgitation	
(D) Morphological feature of RHD of the AV (eg, valvular thickening) without any associated pathological stenosis or regurgitation	
Pathological mitral regurgitation (all four Doppler criteria must be met)	**Pathological aortic regurgitation (all four Doppler criteria must be met)**
Seen in two views	Seen in two views
In at least one view, jet length ≥2 cm	In at least one view, jet length ≥1 cm
Velocity ≥3 m/s for one complete envelope	Velocity ≥3 m/s in early diastole
Pan-systolic jet in at least one envelope	Pan-diastolic jet in at least one envelope
Morphological features of the MV	Morphological features of the AV
AMVL thickening ≥3 mm	Irregular or focal thickening
Chordal thickening	Coaptation defect
Restricted leaflet motion	Restricted leaflet motion
Excessive leaflet tip motion during systole	Prolapse

*Congenital MV anomalies must be excluded.
†Bicuspid AV, dilated aortic root and hypertension must be excluded.
‡Combined AR and MR in high-prevalence regions and in the absence of congenital heart disease is regarded as rheumatic.
AMVL, anterior mitral valve leaflet; AR, aortic regurgitation; AV, aortic valve; MR, mitral regurgitation; MS, mitral stenosis; MV, mitral valve; RHD, rheumatic heart disease; WHF, World Heart Federation.

were developed to discriminate at the milder end of the spectrum of RHD. The echocardiography of severe RHD has been well characterised.[9]

Since its publication, the 2012 WHF echocardiographic criteria for RHD have proven to be highly sensitive compared with auscultation[5 10] and highly specific in the school-aged population.[11–13] Three large population-based surveys showed that no 'low-risk' children were labelled with 'definite RHD' using the WHF definitions.[11–13] Importantly, the criteria have been widely adopted for use since 2012 and have in essence become the gold standard.[8 10 14 15]

Concerns have been raised that the use of WHF criteria may be too complex for population-based screening.[14 16] The interpretation of echocardiograms and specifically grading of severity of valvular regurgitation is known to have variable reproducibility.[17 18] If echocardiography is to be used for population-based screening of school-aged children or for monitoring of disease progression and regression, then it is essential to ensure that the diagnosis of mild RHD is reproducible. This has not been formally evaluated to date.

The primary objective of this study is to assess the intra-rater and inter-rater reliability and the agreement parameters associated with the 2012 WHF echocardiographic criteria in terms of differentiating no RHD from borderline and definite RHD.

METHODOLOGY

This study is reported on in accordance with guidelines for reporting on reliability and agreement studies—GRRASS 2011.[19]

Sample size

Sample size of 200 was chosen based on consideration of prevalence of disease and precision to be expected in estimates in kappa index and agreement parameters. Sample size calculations were performed using nQuery software. Using nQuery, if kappa (κ)=0.8, precision of ±0.1 can be expected with n=200 if prevalence of RHD is 0.25.

Study participants

Members of the WHF Advisory Group on echocardiographic screening of RHD participated as raters or reporters in the study: 15 cardiologists from 9 countries (Australia, Brazil, France, India, Mozambique, New Zealand, Samoa, South Africa and USA).

Echocardiograms

Two hundred de-identified digital echocardiographic studies were uploaded onto a secure website for viewing and reporting. Images were obtained prospectively from two large echocardiographic epidemiologic RHD screening studies conducted between 2008 and 2010 in New Zealand[20] and Australia.[10] Echocardiography was performed by qualified echocardiographers on Vivid E and Vivid I machines. From each site, 100 studies were selected. Normal case distribution during echocardiographic screening is 97% no RHD, 1%–2% borderline RHD and 1% definite RHD.[10] In order to attain case distribution ideal for the evaluation of the reliability of the WHF criteria with kappa statistics, a non-probabilistic sampling methodology was used. The target distribution was 1/3 no RHD, 1/3 borderline RHD and 1/3 definite RHD. To achieve this, from each site consecutive abnormal studies (borderline and definite RHD as judged by the original reporting team) were enrolled as well as consecutive subtly abnormal studies that did not meet WHF definitions for RHD. Completely normal echocardiograms were excluded. Subtly abnormal studies included those with physiological mitral or aortic regurgitation, isolated morphological feature of RHD such as valvular or chordal thickening, and minor congenital defects such as a bicuspid valve. Excluding completely normal studies decreased the sample size required for statistical validity and made the study feasible with a large number of reporters.

Echocardiographic studies included the following moving images: parasternal-long-axis, parasternal-short-axis, apical-four-chamber and apical-five-chamber views (2D and colour Doppler). Still-frame images included in studies were continuous wave (CW) Doppler, image of the anterior mitral valve leaflet (AMVL) in diastole with measurement, and images of aortic and mitral regurgitant jets with measurements. The study participants were directed to re-measure these parameters using strict protocols as per WHF guidelines.[7]

Reporting

Reporting cardiologist independently reviewed all 200 echocardiographic studies and entered reports in a standardised secure website that was specifically designed to view echocardiograms, perform measurements and report on echocardiograms, based

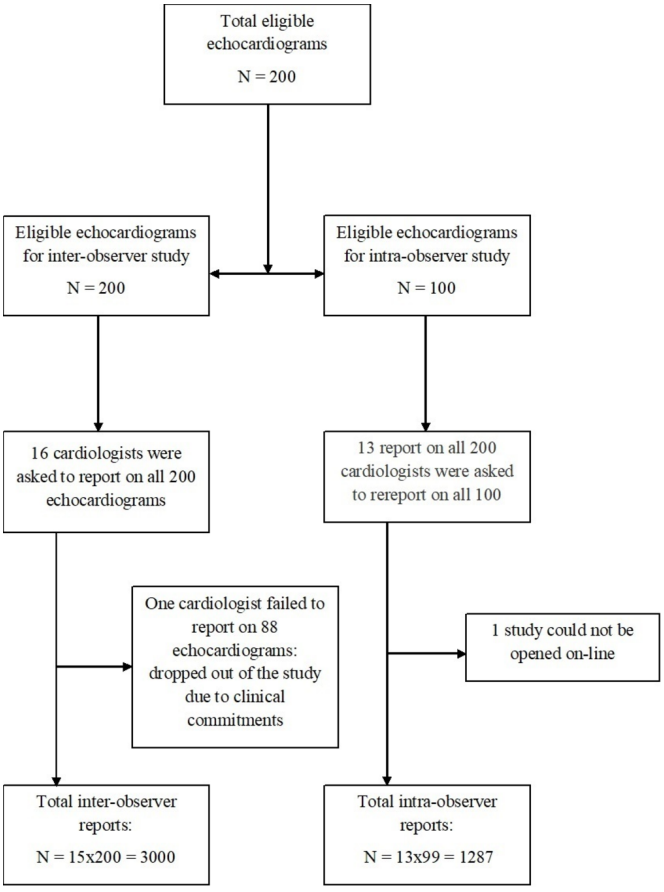

Figure 1 Flow of echocardiogram reports.

on the 2012 WHF criteria. Cardiologists were blinded to all clinical information and case distribution. The flow of echocardiogram reports are depicted in figure 1.

To measure intra-observer variability, 100 images were re-coded and randomly re-uploaded to the website for re-reporting. Cardiologists were blinded to their original reading. Thirteen out of the 15 cardiologists participated in the intra-observer component of the study. The interval between first and second reading was >6 months.

Endpoints

The primary outcomes were to assess intra-rater and inter-rater reliability and proportion of agreement in categorising echocardiograms as no RHD, borderline or definite RHD, as per 2012 WHF criteria.[7] Secondary outcomes were to assess agreement in identifying individual components of the 2012 WHF criteria such as pathological regurgitation, valvular thickening and chordal thickening as detailed in table 1.

The interpretation of kappa values was based on the Landis and Koch guidelines[21]:

< 0 κ	Poor agreement
0.01 – 0.20 κ	Slight agreement
0.21 – 0.40 κ	Fair agreement
0.41 – 0.60 κ	Moderate agreement
0.61 – 0.80 κ	Substantial agreement
0.81 – 1.00 κ	Almost perfect agreement

Ethics

Ethics approvals were obtained from Australia and New Zealand and individual patient consent was waived. All patients had previously provided formal written consent for the echocardiographic screening programmes.[10 20] This study used de-identified and non-re-identifiable images for secondary research use.

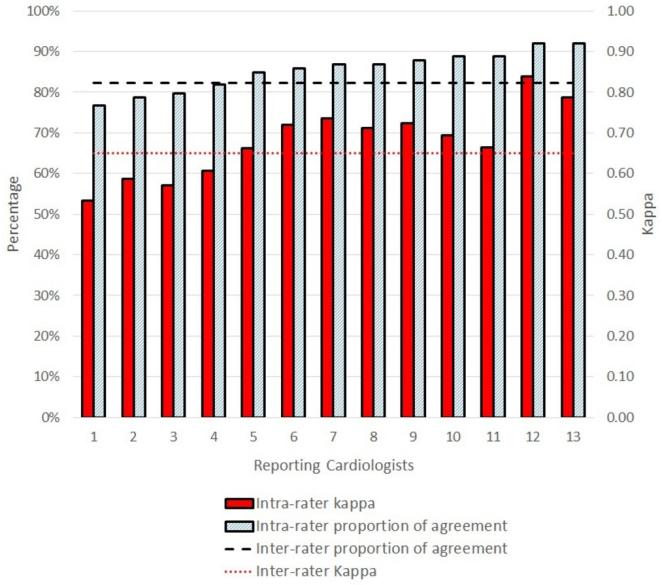

Figure 2 Definite rheumatic heart disease: inter-rater and intra-rater reliability and agreement.

Statistical analysis

Data were exported in Excel format from the designated research website. Statistical calculations were performed with the Statistical Package SAS software V.9.4 (SAS Institute, Cary, North Carolina, USA).

Inter-rater reliability was calculated using Fleiss' free-marginal multi-rater kappa, as this was deemed to be the most appropriate statistics when marginals are not fixed and hence raters are unaware of case distribution.[22] Intra-rater reliability were measured using Cohen's kappa coefficient for dichotomous variables and linearly weighted Cohen's kappa for trichotomous variables (no RHD, borderline RHD and definite RHD). Inter-rater reliability was expressed as mean kappa values and reported with a 95% CI. Intra-rater measurements were expressed as median kappa values with an IQR. The proportion of agreements were reported as mean percentages with a 95% CI for inter-rater agreement and as median with IQR for intra-rater agreement. Individual intra-rater reliability and agreement parameters are depicted in figures.figures 2–6 In the absence of a gold standard, it was not statistically possible to provide individual inter-rater results.

Prevalence of many of the secondary endpoints, morphological features of RHD, were low. Both kappa values and proportions of agreement were reported. Kappa values were not adjusted for disease prevalence as per standard reporting requirements. When disease prevalence is very high or very low (rather than intermediate), the κ values decrease relative to the percentage of agreement, as κ is a relative measure of reliability and is heavily influenced by disease prevalence.[23]

RESULTS

Echocardiograms were obtained from RHD screening studies conducted at schools in children aged 5–15 years in Australia[10] and 11–13 years in New Zealand.[20] In those studies, 79%

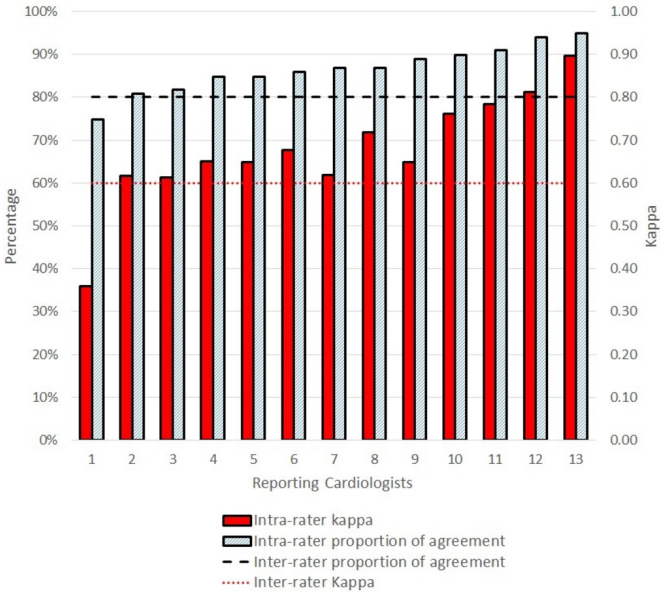

Figure 3 Any rheumatic heart disease (borderline and definite): inter-rater and intra-rater reliability and agreement.

individuals identified as indigenous Australian, Maori or Pacific Islander and 49% were female.

A total of 3000 reports by 15 cardiologists were analysed for the inter-observer assessment. One cardiologist only reported on final diagnosis and not on subcategories. Thirteen cardiologists participated in the intra-rater assessment. Each reported on 99 echocardiograms as one study was uploaded to the website erroneously and hence 1287 reports were analysed. The flow of echocardiogram reports is depicted in figure 1.

In those without the target conditions of RHD, 13 had congenital heart disease as per original reports for the original screening programme where images were obtained from: 7 had bicuspid aortic valve (AV), 4 MV prolapse disease, 1 ventricular septal defect and 1 had atrioventricular septal defect.

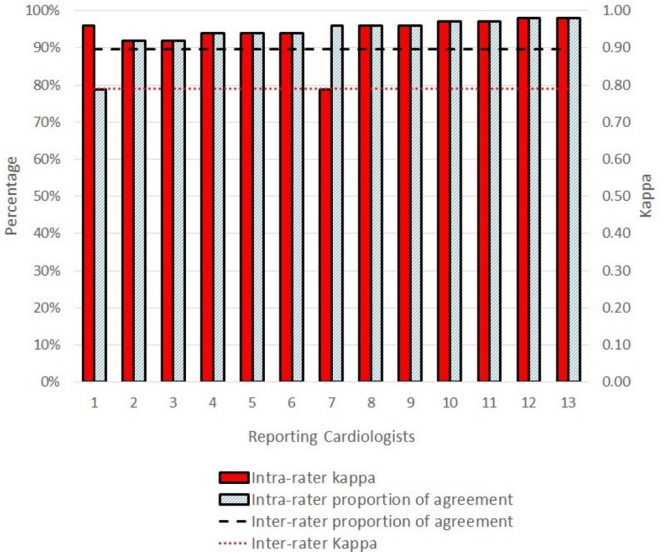

Figure 4 Mitral regurgitation: inter-rater and intra-rater reliability and agreement.

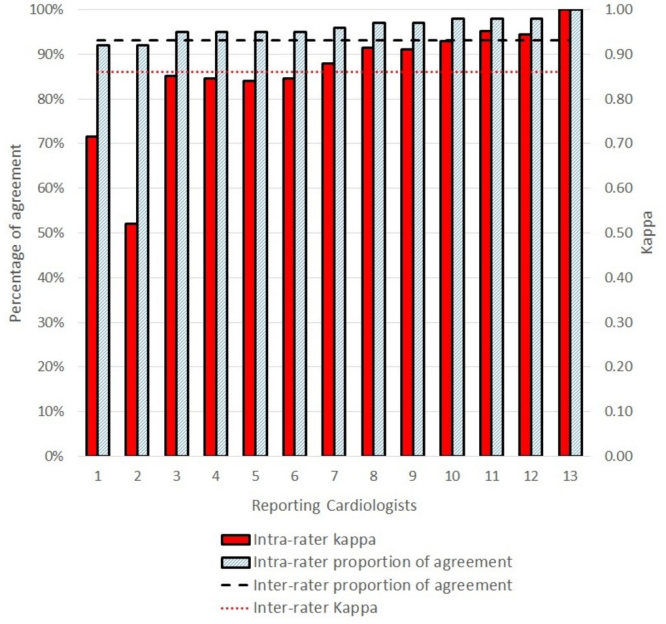

Figure 5 Aortic regurgitation: inter-rater and intra-rater reliability and agreement.

Primary endpoint: RHD

Overall, the inter-rater reproducibility in categorising echocardiograms as no RHD, borderline and definite RHD (primary endpoint) was moderate with mean Fleiss' free-marginal multi-rater kappa of 0.49 (95% CI 0.45 to 0.54) figure 2. When inter-rater reproducibility readings were dichotomised, is there definite RHD or is there any RHD, the agreement was substantial with of κ 0.65 (95% CI 0.59 to 0.70) and κ 0.6 (95% CI 0.55 to 0.65), respectively figures 3 and 4. Total proportion of agreement was highest when results were dichotomised to answer the question "Is there definite RHD?" with a total agreement of 82.27% (95% CI 79.54% to 84.99%) figure 3. Table 2 details reliability and agreement parameters inter-rater and intra-rater reproducibility.

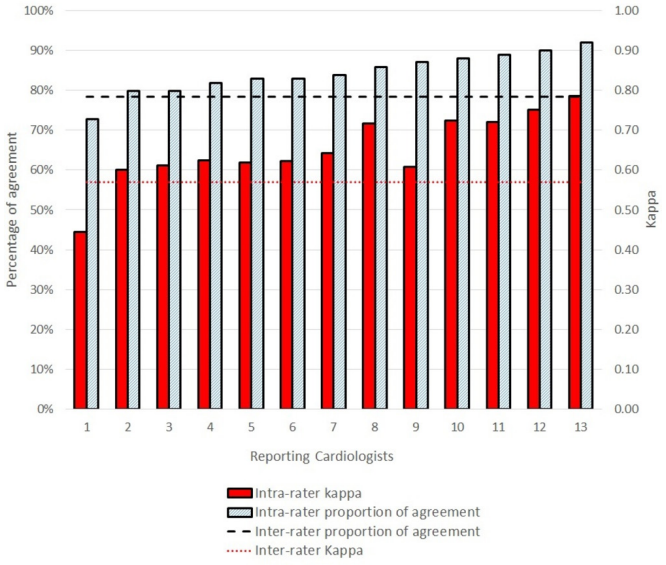

Figure 6 Presence of two or more morphological features of rheumatic heart disease of the mitral valve: inter-rater and intra-rater reliability and agreement.

Table 2 Inter-rater and intra-rater reproducibility of the WHF criteria

Echo features	Inter-rater reproducibility		Intra-rater reproducibility	
	Reliability Kappa mean (95% CI)	Proportion of agreement (95% CI)	Reliability Kappa median (IQR)	Proportion of agreement Median (IQR)
RHD				
No vs borderline vs definite RHD	0.49 (0.45 to 0.54)	66.29% (63.14% to 69.44%)	0.68 (0.60–0.72)	74.75% (68.69%–80.81%)
Any RHD (definite or borderline)	0.60 (0.55 to 0.65)	80.04% (77.41% to 82.67%)	0.65 (0.62–0.76)	86.87% (84.85%–89.90%)
Definite RHD	0.65 (0.59 to 0.70)	82.27 % (79.54% to 84.99%)	0.69 (0.61–0.72)	86.87% (81.82%–88.89%)
Components of criteria				
Pathological MR	0.79 (0.75 to 0.84)	89.62% (87.39% to 91.84%)	0.92 (0.88–0.94)	95.96% (93.94%–96.97%)
Pathological AR	0.86 (0.83 to 0.90)	93.19% (91.28% to 95.10%)	0.88 (0.85–0.93)	95.96% (94.95%–97.98%)
≥2 morphological features of MV	0.57 (0.51 to 0.62)	78.30 % (75.49% to 81.11%)	0.62 (0.61–0.72)	83.84% (81.82%–87.88%)
Mitral valve morphology				
Thickened AMVL	0.75 (0.70 to 0.80)	87.27% (84.83% to 89.72%)	0.79 (0.78–0.85)	89.9% (88.89%–92.93%)
Chordal thickening	0.54 (0.49 to 0.59)	76.78% (74.21% to 79.33%)	0.37 (0.25–0.48)	81.82% (77.78%–88.89%)
Restrictive motion	0.55 (0.49 to 0.60)	77.36% (74.64% to 80.01%)	0.56 (0.33–0.62)	83.84% (76.77%–86.87%)
Excessive leaflet motion	0.63 (0.58 to 0.68)	81.46% (79.01% to 83.90%)	0.41 (0.22–0.48)	88.89% (87.88%–92.93%)
Aortic valve				
Thickened	0.67 (0.62 to 0.72)	83.29% (80.85% to 85.72%)	0.39 (0.24–0.51)	90.91% (86.87%–94.95%)
Coaptation defect	0.91 (0.88 to 0.94)	95.43% (93.92% to 96.94%)	0.56 (0.45–65)	97.98% (95.96%–98.99%)
Restrictive motion	0.97 (0.96 to 0.98)	98.39% (97.73% to 99.01%)	0.66 (Not Calculable)	98.99% (97.98%–100%)
Prolapse	0.93 (0.91 to 0.96)	96.70% (95.36% to 98.04%)	0.57 (0.45–1.0)	98.99% (97.98%–100%)
Congenital heart disease				
MV prolapse disease	0.92 (0.89 to 0,95)	95.92% (94.47% to 97.38%)	0.66 (0.65–0.85)	97.98% (95.96%–98.99%)
Bicuspid AV	0.95 (0.93 to 0.98)	97.66% (96.41% to 98.91%)	0.66 (0.65–0.89)	98.99% (97.98%–98.99%)

AMVL, anterior mitral valve leaflet; AR, aortic regurgitation; AV, aortic valve; MR, mitral regurgitation; MV, mitral valve; NC, not calculable; RHD, rheumatic heart disease; WHF, World Heart Federation.

The intra-rater reproducibility (reliability and agreement) parameters in categorising echocardiograms as no RHD, borderline and definite RHD were as follows: the median linearly weighted Cohen κ was 0.68 (IQR 0.60–0.72) and total proportion of agreement was 74.75% (IQR 68.69%–80.81%). Median results are detailed in table 2 and individual results of reporting cardiologist depicted infigures 2–4.

Secondary endpoints

The inter-rater reliability of identifying isolated pathological mitral and aortic regurgitation was 'good' and 'almost perfect', κ 0.79 (95% CI 0.75 to 0.84) and κ 0.86 (95% CI 0.83 to 0.90), repectively see table 2 and figures 5 and 6. The inter-rater reliability of detecting ≥2 morphological features of RHD of the MV was 'substantial' with a κ of 0.57 (95% CI 0.51 to 0.62), with a proportion of agreement of 78.3% (95% CI 75.49% to 81.11%), see table 2 and figure 7. The most reliably detected morphological feature of the MV was the objective measure of thickening of the AMVL with an inter-rater κ of 0.75 (95% CI 0.7 to 0.8). The least reliable morphological feature of the MV was chordal thickening with an inter-rater κ of 0.54 (95% CI 0.49 to 0.59). The most reliably detected morphological feature of the AV was restricted leaflet motion with an inter-rater κ of 0.97 (95% CI 0.96 to 0.98). The least reliably detected morphological feature of the AV was the subjective measure of thickening with an inter-rater κ of 0.67 (95% CI 0.62 to 0.72). Further details are provided in table 2 and individual results detailed in figures 2–7.

DISCUSSION

This study demonstrates that WHF echocardiographic criteria enable reliable categorisation of screening echocardiograms as no RHD, borderline and definite RHD. The inter-rater and intra-rater reliability were substantial with a κ of 0.49 and 0.68, respectively. This level of reliability is comparable with that of other screening tests such as mammography for breast cancer screening κ 0.53–0.77[24 25] and surpasses the reliability associated with other tests like the cytological assessment of the screening Papanicolaou (Pap) smears testing for cervical cancer (κ 0.46).[26] Reliability improved when catagorisation of echocardiograms

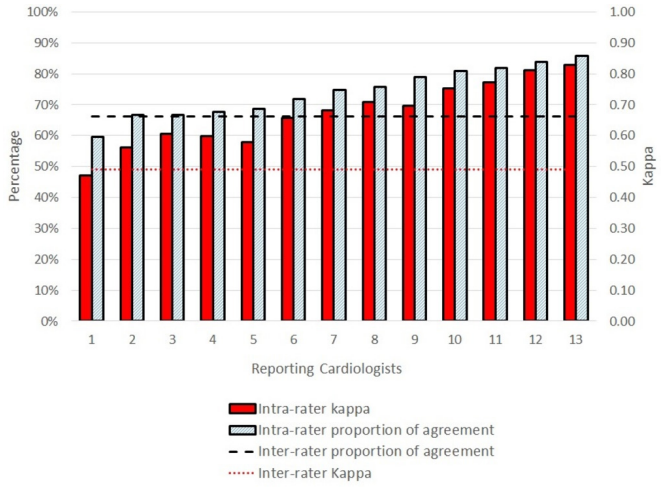

Figure 7 Categorising echocardiograms as 'no RHD', 'borderline RHD' and 'definite RHD': inter-rater and intra-rater reliability and agreement. RHD, rheumatic heart disease.

was dichotomised—"is there any RHD (borderline or definite)?" or "is there definite RHD?" with respective inter-rater κ values of 0.6 and 0.65, respectively.

Similarly, there was a good level of absolute agreement in deciding if definite RHD was present, with a total proportion of agreement being 82.27%. A test that is associated with a high level of absolute proportion of agreement is deemed to be a suitable tool to detect change over time,[23] indicating that the WHF criteria should be a suitable tool to monitor disease progression or resolution.

There was almost perfect inter-rater and intra-rater agreement detecting pathological mitral regurgitation with κ of 0.79 and 0.92, respectively. This is substantially superior to agreement over the presence of severe MR regardless of methodology used[17 18] and is likely the result of having very strict definitions where all four criteria must be met for regurgitation to be considered pathological (table 1).[7] Therefore, physiological mitral regurgitation, which occurs in up to 18% of healthy children, can be very reliably differentiated from pathological mitral regurgitation that occurs in less than 0.5% of low-risk and up to 3% of children in high-risk populations for RHD.[10]

The reproducibility of identifying two or more morphological features of RHD of the MV (borderline category A) in a given echocardiogram was substantial with a κ 0.57 and an absolute proportion of agreement of 78.3%. The most reliable detected morphological features of the MV were AMVL thickening and excessive leaflet motion, while for the AV, it was restricted leaflet motion and AV prolapse.

Cohen's kappa that was used to analyse intra-rater agreement is a relative measure of agreement (actual agreement minus expected agreement by chance). When disease distribution is skewed and prevalence is either very high or very low, then the expected level of agreement by chance rises and the actual kappa value lowers. Hence, kappa value is a relative measure of agreement and is influenced by disease prevalence. For inter-rater agreement, Fleiss' free-marginal multi-rater kappa was used which better compensates for skewed distribution. By necessity, the different kappa statistics were used for multi-rater inter-observer agreement and bi-rater intra-observer agreement, and the results varied for echocardiographic features that were rare and this highlights some of the limitations of kappa statistics.

The total proportion of agreement over the presence of thickening of the MV and AV were similar: 87.27% and 83.29%, respectively. Similarly, inter-rater kappa values were 0.75 and 0.67, respectively. This is despite the fact the AMVL thickening had an objective measure (of >3 mm) while AV thickening was a subjective observation. Webb and colleagues found similarly high inter-observer agreement in relation to MV thickness measurements with an inter-class correlation coefficient of 0.85.[27] They applied the same strict methodology as described in the WHF diagnostic guidelines.[7]

The absolute proportion of agreement in identifying individual morphological features of RHD was high for all features and ranged from 76.78% for chordal thickening to 98.39% for restricted motion of the AV.

To implement active surveillance for RHD on a global scale, as recommended by WHO some decades ago,[28] would require considerable increase in human resources. Task shifting, through echocardiography performed by health workers, could provide part of the solution to make active case finding a reality in resource-poor settings. Concerns have been raised that the use of WHF criteria may be too complex for population-based screening and simplified criteria might be more practical in the field.[14 16] As a result, the WHF criteria have already been modified by some researchers to allow for the use of hand-held echocardiography machines without CW capabilities and for health worker–led echocardiographic screening.[15 29] Those criteria have focused on detecting mitral and/or aortic regurgitation and have ignored morphological features of RHD.

Our study supports the use of simplified criteria in the field. The most reliable component of the WHF criteria was the diagnosis of pathological mitral and aortic regurgitation, and hence it is appropriate to focus on these features when large-scale screening is being considered. The current study supports the use of the WHF guidelines for the final diagnosis of RHD for those individuals detected as positive for RHD by simplified screening protocols.

Regardless of skill level and whether the full WHF criteria or modified criteria are used for simplicity, rigorous training protocols and evaluation of competency prior to engaging in performing or reporting on screening echocardiograms for RHD should be mandatory.[30]

There are many unknowns that remain about echocardiographic screening for RHD. Perhaps the most important of these is the natural history of echocardiography-detected RHD. This study demonstrated that the WHF criteria could be useful in detecting change over time and therefore it could be an appropriate tool to use to evaluate the impact of secondary prophylaxis on disease progression of borderline RHD. A randomised control trial is currently under way to determine the absolute benefit of secondary prophylaxis in the setting of subclinical mild, definite and borderline RHD (The GOAL trial, Clinicaltrials.gov: NCT03346525).

The WHF echocardiographic criteria have shown good discriminating capacity and hence would be a suitable tool for population-based screening, active case finding and for diagnosis of RHD in the clinical setting. Having a reliable diagnostic method also permits the monitoring of epidemiological patterns and could aid the evaluation of interventions that are designed to reduce RHD burden, for example, sore throat programmes, Group A streptococcal vaccine trials or echocardiographic screening programmes.

Limitations

This study was limited to interpretation of echocardiograms by cardiologists experienced in RHD. It is recognised that acquisition of high-quality images is fundamental to accurate diagnoses. In this study, all images were obtained by highly qualified echocardiographers in Australia and New Zealand, which may not be the case in screening studies in many resource-limited settings. Echocardiograms were obtained from screening studies from Australia and New Zealand only and may not be representative of demographics or disease pattern elsewhere. The 2012 WHF echocardiographic definitions for RHD are considered to be the current gold standard and were based on the best available echocardiographic, pathological and postmortem evidence of RHD.[7] The current study represents the definitive validation of their reliability and agreement. Randomised controlled trials or carefully designed longitudinal studies are needed to ascertain risk of disease progression and the benefit of secondary prophylaxis for borderline RHD. Finally, the provision of still-frame images in our study may have inadvertently increased agreement.

CONCLUSION

This study demonstrates that application of the WHF echocardiographic criteria by specialist cardiologists enables reliable categorisation of screening echocardiograms as no RHD, borderline RHD and definite RHD. Pathological regurgitation is reliably differentiated from physiological regurgitation by experienced cardiologists. Agreement over the presence of morphological features of RHD was substantial, but the reliability was lower due to low prevalence of individual features. This study has demonstrated that the WHF criteria are useful tools for screening for RHD and for monitoring disease progression and resolution. They can also be used for clinical evaluation of new cases of MV and AV disease. Longitudinal studies are needed to evaluate the clinical significance of echocardiography-detected mild borderline and definite RHD.

Author affiliations
[1]Menzies School of Health Research, Casuarina, Northern Territory, Australia
[2]Green Lane Cardiovascular Services, Auckland City Hospital, Auckland, New Zealand
[3]Telethon Kids Institute, University of Western Australia, Subiaco, Western Australia, Australia
[4]Paediatric and Congenital Cardiac Services, Starship Children's Hospital, Auckland, New Zealand
[5]Maputo HeartInstitute, Maputo, Mozambique
[6]Amrita Institute of Medical Sciences and Research Centre, Kochi, India
[7]Paediatrics and Child Health, Stellenbosch University, Cape Town, South Africa
[8]Department of Paediatrics and Child Health, Cape Town, South Africa
[9]Hop Europeen Georges Pompidou, Paris, France
[10]INSERM U970, Paris Cardiovascular Research Center PARCC, Paris, France
[11]Inst Coracao, New York City, New York, USA
[12]Federal University of Minas Gerais, Belo Horizonte, Brazil
[13]Cardiology, Project Health for León, Raleigh, North Carolina, USA
[14]All India Institute of Medical Sciences, New Delhi, India
[15]Pediatric Cardiology, Children's National Health System, Washington, District of Columbia, USA
[16]Cardiology, Samoa National Hospital, Apia, Samoa
[17]Pediatric Cardiology, Sri Jayadeva Institute of Cardiovascular Sciences and Research, Bangalore, , India
[18]Cardiology, Women's and Children's Hospital, Adelaide, South Australia, Australia
[19]Groote Schuur Hospital and University of Cape Town, Cape Town, South Africa
[20]Paediatric and Congenital Cardiology, Starship Children's Hospital, Auckland, New Zealand
[21]University of Auckland, Auckland, New Zealand

Acknowledgements BR received a scholarship from Heart Foundation of New Zealand and from the Lowitja Institute of Australia.

Contributors JC, NJW, TLG. KS and BR made substantial contributions to the conception and design of the work. BR, JC, JWS, BF, KK, JL, EM, MM, AOM, CM, JP, AS, JS, SV, IBV, GRW, LZ, KS, TLG and NJW made substantial contributions to the acquisition, analysis or interpretation of data for the work. BR prepared draft of manuscript. All authors made substantial contribution to the work or revising it critically for important intellectual content and final approval of the version to be published.

Funding Funding was received from the Green Lane Research and Education Fund, Auckland, New Zealand for the development of the study website.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval Ethics approvals were obtained for the study from the Northern X Regional Ethics Committee of the Ministry of Health of New Zealand and from the Human Research Ethics Committee of the Northern Territory Department of Health and Community Services of Australia. Both Ethics Committees waived individual patient consent.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as online supplementary information.

REFERENCES

1. Watkins DA, Johnson CO, Colquhoun SM, *et al*. Global, regional, and national burden of rheumatic heart disease, 1990–2015. *N Engl J Med Overseas Ed* 2017;377:713–22.
2. Zühlke L, Engel ME, Karthikeyan G, *et al*. Characteristics, complications, and gaps in evidence-based interventions in rheumatic heart disease: the global rheumatic heart disease registry (the remedy study). *Eur Heart J* 2015;36:1115–22.
3. Kassem AS, el-Walili TM, Zaher SR, *et al*. Reversibility of mitral regurgitation following rheumatic fever: clinical profile and echocardiographic evaluation. *Indian J Pediatr* 1995;62:717–23.
4. Tompkins DG, Boxerbaum B, Liebman J. Long-term prognosis of rheumatic fever patients receiving regular intramuscular benzathine penicillin. *Circulation* 1972;45:543–51.
5. Roberts KV, Brown ADH, Maguire GP, *et al*. Utility of auscultatory screening for detecting rheumatic heart disease in high-risk children in Australia's Northern Territory. *Med J Aust* 2013;199:196–9.
6. Marijon E, Ou P, Celermajer DS, *et al*. Prevalence of rheumatic heart disease detected by echocardiographic screening. *N Engl J Med* 2007;357:470–6.
7. Reményi B, Wilson N, Steer A, *et al*. World Heart Federation criteria for echocardiographic diagnosis of rheumatic heart disease—an evidence-based guideline. *Nat Rev Cardiol* 2012;9:297–309.
8. Gewitz MH*et al*. Revision of the Jones criteria for the diagnosis of acute rheumatic fever in the era of Doppler echocardiography: a scientific statement from the American Heart Association. *Circulation* 2015;131:1806–18.
9. Saxena A. Echocardiographic diagnosis of chronic rheumatic valvular lesions. *Global Heart* 2013;8:203–12.
10. Roberts K, Maguire G, Brown A, *et al*. Echocardiographic screening for rheumatic heart disease in high and low risk Australian children. *Circulation* 2014;129:1953–61.
11. Roberts KV, Maguire GP, Brown A, *et al*. Rheumatic heart disease in indigenous children in northern Australia: differences in prevalence and the challenges of screening. *Med J Aust* 2015;203.
12. Webb RH, Gentles TL, Stirling JW, *et al*. Valvular regurgitation using portable echocardiography in a healthy student population: implications for rheumatic heart disease screening. *J Am Soc Echocardiogr* 2015;28:981–8.
13. Clark BC, Krishnan A, McCarter R, *et al*. Using a low-risk population to estimate the specificity of the World Heart Federation criteria for the diagnosis of rheumatic heart disease. *J Am Soc Echocardiogr* 2016;29.
14. Lu JC, Sable C, Ensing GJ, *et al*. Simplified rheumatic heart disease screening criteria for handheld echocardiography. *J Am Soc Echocardiogr* 2015;28:463–9.
15. Engelman D, Kado JH, Reményi B, *et al*. Focused cardiac ultrasound screening for rheumatic heart disease by briefly trained health workers: a study of diagnostic accuracy. *The Lancet Global Health* 2016;4:e386–94.
16. Nascimento BR, Nunes MCP, Lopes ELV, *et al*. Rheumatic heart disease echocardiographic screening: approaching practical and affordable solutions. *Heart* 2016;102:658–64.
17. Grayburn PA, Bhella P. Grading severity of mitral regurgitation by echocardiography: science or art? *JACC Cardiovasc Imaging* 2010;3:244–6.
18. Biner S, Rafique A, Rafii F, *et al*. Reproducibility of proximal isovelocity surface area, vena contracta, and regurgitant jet area for assessment of mitral regurgitation severity. *JACC: Cardiovascular Imaging* 2010;3:235–43.
19. Kottner J, Audigé L, Brorson S, *et al*. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *Journal of Clinical Epidemiology* 2011;64:96–106.
20. Webb RH, Wilson NJ, Lennon DR, *et al*. Optimising echocardiographic screening for rheumatic heart disease in New Zealand: not all valve disease is rheumatic. *Cardiol Young* 2011;21:436–43.

21. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–74.
22. Randolph JJ. Free-marginal multirater kappa (multirater K [free]): an alternative to Fleiss' fixed-marginal multirater kappa. *ERIC* 2005.
23. de Vet HCW, Mokkink LB, Terwee CB, *et al*. Clinicians are right not to like Cohen's. *BMJ* 2013;346:f2125.
24. Ooms EA, Zonderland HM, Eijkemans MJC, *et al*. Mammography: interobserver variability in breast density assessment. *The Breast* 2007;16:568–76.
25. Redondo A*et al*. Inter- and intraradiologist variability in the BI-RADS assessment and breast density categories for screening mammograms. *Br J Radiol* 2014.
26. Stoler MH, Schiffman M. Interobserver reproducibility of cervical cytologic and histologic interpretations realistic estimates from the ASCUS-LSIL triage study. *JAMA* 2001;285:1500–5.
27. Webb RH, Culliford-Semmens N, Sidhu K, *et al*. Normal echocardiographic mitral and aortic valve thickness in children. *Heart Asia* 2017;9:70–5.
28. WHO Technical Report Series. WHO expert consultation on rheumatic fever and rheumatic heart disease (2001: Geneva Switzerland), rheumatic fever and rheumatic heart disease: report of a WHO expert consultation. Geneva World Health Organization, WHO Technical Report Series; 2001.
29. Ploutz M, Lu JC, Scheel J, *et al*. Handheld echocardiographic screening for rheumatic heart disease by non-experts. *Heart* 2016;102:35–9.
30. Engelman D, Okello E, Beaton A, *et al*. Evaluation of computer-based training for health workers in echocardiography for RhD. *Global Heart* 2017;12:17–23.