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 Wilson2,21

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.

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
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 auscultation10 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 15 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 on to 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 Zealand20 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
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 guidelines21:

<table>
<thead>
<tr>
<th>Kappa Value</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 0</td>
<td>Poor agreement</td>
</tr>
<tr>
<td>0.01 – 0.20</td>
<td>Slight agreement</td>
</tr>
<tr>
<td>0.21 – 0.40</td>
<td>Fair agreement</td>
</tr>
<tr>
<td>0.41 – 0.60</td>
<td>Moderate agreement</td>
</tr>
<tr>
<td>0.61 – 0.80</td>
<td>Substantial agreement</td>
</tr>
<tr>
<td>0.81 – 1.00</td>
<td>Almost perfect agreement</td>
</tr>
</tbody>
</table>

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.

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 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 Australia10 and 11–13 years in New Zealand.20 In those studies, 79%
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.

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 $\kappa$ of 0.65 (95% CI 0.59 to 0.70) and $\kappa$ of 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.
TABLE 2. Inter-rater and intra-rater reproducibility of the WHF criteria

<table>
<thead>
<tr>
<th>Echo features</th>
<th>Inter-rater reproducibility</th>
<th>Intra-rater reproducibility</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reliability Kappa mean (95% CI)</td>
<td>Proportion of agreement (95% CI)</td>
</tr>
<tr>
<td>RHD</td>
<td>No vs borderline vs definite RHD</td>
<td>0.49 (0.45 to 0.54) 66.29% (63.14% to 69.44%)</td>
</tr>
<tr>
<td></td>
<td>Any RHD (definite or borderline)</td>
<td>0.60 (0.55 to 0.65) 80.04% (77.41% to 82.67%)</td>
</tr>
<tr>
<td></td>
<td>Definite RHD</td>
<td>0.65 (0.59 to 0.70) 82.27 % (79.54% to 84.99%)</td>
</tr>
<tr>
<td>Components of criteria</td>
<td>Pathological MR</td>
<td>0.79 (0.75 to 0.84) 89.62% (87.39% to 91.84%)</td>
</tr>
<tr>
<td></td>
<td>Pathological AR</td>
<td>0.86 (0.83 to 0.90) 93.19% (91.28% to 95.10%)</td>
</tr>
<tr>
<td>≥2 morphological features of MV</td>
<td>0.57 (0.51 to 0.62) 78.30% (75.49% to 81.11%)</td>
<td>0.62 (0.61–0.72) 83.84% (81.82%–87.88%)</td>
</tr>
<tr>
<td>Mitral valve morphology</td>
<td>Thickened AMVL</td>
<td>0.75 (0.70 to 0.80) 87.27% (84.83% to 89.72%)</td>
</tr>
<tr>
<td></td>
<td>Chordal thickening</td>
<td>0.54 (0.49 to 0.59) 76.78% (74.21% to 79.33%)</td>
</tr>
<tr>
<td></td>
<td>Restrictive motion</td>
<td>0.55 (0.49 to 0.60) 77.36% (74.64% to 80.01%)</td>
</tr>
<tr>
<td></td>
<td>Excessive leaflet motion</td>
<td>0.63 (0.58 to 0.68) 81.46% (79.01% to 83.90%)</td>
</tr>
<tr>
<td>Aortic valve</td>
<td>Thickened</td>
<td>0.67 (0.62 to 0.72) 83.29% (80.85% to 85.72%)</td>
</tr>
<tr>
<td></td>
<td>Coaptation defect</td>
<td>0.91 (0.88 to 0.94) 95.43% (93.92% to 96.94%)</td>
</tr>
<tr>
<td></td>
<td>Restrictive motion</td>
<td>0.97 (0.96 to 0.98) 98.39% (97.73% to 99.01%)</td>
</tr>
<tr>
<td></td>
<td>Prolapse</td>
<td>0.93 (0.91 to 0.96) 96.70% (95.36% to 98.04%)</td>
</tr>
<tr>
<td>Congenital heart disease</td>
<td>MV prolapse disease</td>
<td>0.92 (0.89 to 0.95) 95.92% (94.47% to 97.38%)</td>
</tr>
<tr>
<td></td>
<td>Bicuspid AV</td>
<td>0.95 (0.93 to 0.98) 97.66% (96.41% to 98.91%)</td>
</tr>
</tbody>
</table>

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 in figures 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), respectively see table 2 and figures 3 and 5. The inter-rater reliability of detecting ≥2 morphological features of RHD of the MV was ‘substantial’ with a κ of 0.57 (95% CI 0.50 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 subjective measure of thickening of the AMVL with an inter-rater κ of 0.75 (95% CI 0.7 to 0.8). The least reliably 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 the 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 0.53–0.7724 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 categorisation of echocardiograms

was dichotomised—"is there any RHD (borderline or definite)?" or "is there definite RHD?" with respective inter-rater $\kappa$ 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, 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 $\kappa$ of 0.79 and 0.92, respectively. This is substantially superior to agreement over the presence of severe MR regardless of methodology used 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). 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.

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 $\kappa$ of 0.57 and an absolute proportion of agreement of 78.3%. The most reliable detected individual morphological features of RHD was 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. They applied the same strict methodology as described in the WHF diagnostic guidelines.

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, 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. 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. 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.

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. 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 Heart Institute, 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 Leon, Raleigh, North Carolina, USA
14 Department of Paediatrics and Child Health, Cape Town, South Africa
15 Pediatric Cardiology, Children’s National Health System, Washington, District of Columbia, USA
16 Cardiology, Samo 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
22. Randolph JJ. Free-marginal multirater kappa (multirater K [free]): an alternative to Fleiss’ fixed-marginal multirater kappa. ERIC 2005.