Open access Review

BMJ Open Diabetes Research & Care

Reliability of recommended non-invasive chairside screening tests for diabetes-related peripheral neuropathy: a systematic review with meta-analyses

Ally McIlhatton,[1] Sean Lanting,[1] David Lambkin,[2] Lucy Leigh,[2] Sarah Casey,[1] Vivienne Chuter[1]

To cite: McIlhatton A, Lanting S, Lambkin D, *et al*. Reliability of recommended non-invasive chairside screening tests for diabetes-related peripheral neuropathy: a systematic review with meta-analyses. *BMJ Open Diab Res Care* 2021;**9**:e002528. doi:10.1136/bmjdrc-2021-002528

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjdrc-2021-002528).

Received 5 August 2021
Accepted 14 November 2021

© Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

[1]Discipline of Podiatry, The University of Newcastle Faculty of Health and Medicine, Ourimbah, New South Wales, Australia
[2]Hunter Medical Research Institute, The University of Newcastle, New Lambton, New South Wales, Australia

Correspondence to
Dr Sean Lanting;
sean.lanting@newcastle.edu.au

ABSTRACT

The objective is to determine, by systematic review, the reliability of testing methods for diagnosis of diabetes-related peripheral neuropathy (DPN) as recommended by the most recent guidelines from the International Diabetes Foundation, International Working Group on the Diabetic Foot and American Diabetes Association. Electronic searches of Cochrane Library, EBSCO Megafile Ultimate and EMBASE were performed to May 2021. Articles were included if they reported on the reliability of recommended chairside tests in diabetes cohorts. Quality appraisal was performed using a Quality Appraisal of Reliability Studies checklist and where possible, meta-analyses, with reliability reported as estimated Cohen's kappa (95% CI). Seventeen studies were eligible for inclusion. Pooled analysis found acceptable inter-rater reliability of vibration perception threshold (VPT) (κ=0.61 (0.50 to 0.73)) and ankle reflex testing (κ=0.60 (0.55 to 0.64)), but weak inter-rater reliability for pinprick (κ=0.45 (0.22 to 0.69)) and 128 Hz tuning fork (κ=0.42 (0.15 to 0.70)), though intra-rater reliability of the 128 Hz tuning fork was moderate (κ=0.54 (0.37 to 0.73)). Inter-rater reliability of the four-site monofilament was acceptable (κ=0.61 (0.45 to 0.77)). These results support the clinical use of VPT, ankle reflexes and four-site monofilament for screening and ongoing monitoring of DPN as recommended by the latest guidelines. The reliability of temperature perception, pinprick, proprioception, three-site monofilament and Ipswich touch test when performed in people with diabetes remains unclear.

INTRODUCTION

Globally, diabetes is reported to affect almost 500 million people.[1] Diabetes-related peripheral neuropathy (DPN) is a common complication of diabetes that results in sensory loss in the extremities, which can lead to impaired balance and gait,[2] as well as the formation of pressure ulcers and subsequent infection.[3] DPN is implicated in 50%–75% of all non-traumatic amputations.[4] DPN is estimated to be present in up to 50% of those with a diabetes duration of over 10 years,[5 6] is present in 10%–30% of people at time of diabetes diagnosis and has also been noted in pre-diabetes.[7]

Non-invasive chairside tests are recommended for diagnosis of DPN and used for ongoing monitoring to map disease progression. Early diagnosis is vital to implement strategies to reduce the risk of limb-threatening sequelae. A multidisciplinary approach in combination with patient education, compliance and routine foot care have demonstrated prophylactic capacity for reducing DPN progression and severity, as well as ulcer risk,[8 9] and intensive glucose control has been shown to reduce incidence of DPN.[10 11]

Various international guidelines exist providing direction as to which chairside tests should be performed for routine screening and monitoring of DPN. These guidelines differ in recommendations of test type and test protocol. The International Diabetes Federation (IDF),[12] International Working Group on the Diabetic Foot (IWGDF)[13] and American Diabetes Association (ADA)[14] represent three major international organizations that develop some of the most widely used guidelines for diabetes-related foot assessment, diagnosis and management. Collectively, these groups recommend variations of the following chairside tests: 10 g monofilament, 128 Hz tuning fork, light touch/Ipswich touch test, temperature perception, vibration perception threshold (VPT), pinprick,[12–14] proprioception and ankle reflexes.[12–14]

Due to the ongoing nature of testing required to facilitate early diagnosis and monitor progression of DPN, it is imperative that the recommended screening tests demonstrate acceptable reliability.[15] However, remarkably, despite the widespread use of recommended chairside testing there has been no comprehensive investigation of their reliability. Therefore, the aim of this research was to, by systematic review of available evidence, evaluate the reliability of screening tests for DPN in the lower limb of adults with diabetes, as per protocols recommended by the most recent guidelines from the IDF, IWGDF and ADA. We hypothesize that all recommended tests will demonstrate acceptable reliability.

RESEARCH DESIGN AND METHODS

Search strategy

This review was registered in the International Prospective Register of Systematic Reviews (PROSPERO ID: CRD42020186383), and reporting is consistent with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses statement. In order to identify studies that have investigated the reliability of non-invasive neurological tests in people with diabetes, an electronic search was performed independently by two authors (AM and SL) until May 2021 using the biomedical databases: Cochrane Library, EBSCO Megafile Ultimate and EMBASE. Search terms used in various combinations and with database relevant truncations were: reliability, consistency, accuracy, reproducibility, repeatability, agreement, precision, monofilament, neuropen, neurotip, tuning fork, vibration, VPT, neurothesiometer, biothesiometer, maxivibrometer, tip-therm, Ipswitch touch test, IpTT, reflex, perception, sensation, nociception, neuropathy and DPN. Abstracts were managed using Endnote X9 software.

Two authors (AM and SL) screened retrieved articles at title and abstract level and final determination of article suitability for inclusion following full-text review was performed in consultation with a third reviewer (VC). Lastly, reference lists of included articles were manually screened for any additional relevant research.

Inclusion and exclusion criteria

Inclusion criteria were: original peer-reviewed research articles or conference abstracts reporting reliability (inter-rater or intra-rater) of any of the non-invasive screening tests for DPN as recommended by either the IDF, IWGDF or ADA in a population with diabetes. Articles were also eligible if a subset of participants had diabetes, provided these data were reported separately or available from the authors. Articles investigating reliability of questionnaires, combination tests such as the Michigan Neuropathy Screening Instrument, tests performed in participant upper limbs, and non-English language texts were excluded. In addition, articles were excluded where time to retest made it likely that results may be affected by disease progression, for example, >1 year.

Statistical analysis

Data were extracted (AM) and cross-checked (SL) using a customized data extraction form that included study and participant characteristics, statistical analyses and reliability results. Where Kappa values or percentage agreement were provided, interpretation of reliability outcomes was in accordance with McHugh.[16] This is reported as none (0–0.20 or 0%–4%), minimal (0.21–0.39 or 4%–15%), weak (0.40–0.59 or 15%–35%), moderate (0.60–0.79 or 35%–63%), strong (0.80–0.90 or 64%–81%) and almost perfect (>0.90 or 82%–100%).[16] In addition to these interpretations, any kappa values >0.60 were considered acceptable, as per the conservative thresholds suggested by McHugh for health research and practice.[16] Coefficient of variation (COV) as the ratio of the SD to the mean was considered to indicate a higher reliability the lower the percentage score.[17] Intraclass correlation coefficients (ICCs) were interpreted in accordance with Portney and Watkins, that is, good (>0.75), moderate (0.5–0.75) and poor (<0.5) reliability.[18] Spearman's rho was interpreted in accordance with Prion and Haerling as negligible (0.00–0.20), weak (0.21–0.40), moderate (0.41–0.60), strong (0.61–0.80) and very strong (0.81–1.00).[19]

Adequate data were available to perform meta-analyses on four different neuropathy tests: ankle reflex, pinprick, 128 Hz tuning fork and VPT. The ankle reflex and pinprick tests were assessed only for their inter-rater reliability. The 128 Hz tuning fork and VPT tests were assessed for both their inter-rater and intra-rater reliability.

An alpha level of 0.05 was specified for all tests and confidence intervals. The data were analysed in R V.4.1.0. Data for each study was presented as Cohen's kappa, with their corresponding variances (of sample distribution) being calculated from additional study results, in order of preference, as below:

- If the SE was reported, the variance was calculated by squaring it.
- If the percentage agreement and number of observations was reported, the variance was calculated by the below formula,[20] where 'p0' is the percentage agreement, 'k' is Cohen's kappa and 'n' is the number of observations:

$$variance = \frac{p_0 \times (1 - p_0)}{(1 - \frac{p_0 - k}{1 - k})^2 \times n}$$

- If the CI was reported, the variance was calculated by the below formula, where 'k_u' and 'k_l' are the upper and lower 95% confidence limits of kappa, respectively:

$$variance = \left(\frac{k_u - k_l}{2 \times 1.96}\right)^2$$

Results of the meta-analysis are presented as estimated Cohen's kappa (95% CI) and total heterogeneity (I^2) with accompanying forest plots. The trim-and-fill method was used to detect and adjust for publication bias.

Meta-analyses were assessed using the R package 'metafor'.[21] For the inter-rater reliability analyses, if at least three papers were available, a random-effects model was specified, using the DerSimonian-Laird method and Knapp-Hartung adjustment. If only two papers were available, a fixed-effects model was specified, using the fixed-effects method. For the intra-rater reliability analyses, if at least three raters were available, a random-effects model was specified, using the DerSimonian-Laird method and Knapp-Hartung adjustment. If only two raters were available, a fixed-effects model was specified, using the fixed-effects method.

When data were collected more than once on the same participant, the mean kappa and variance was used in the meta-analysis. This occurred if either the same participant was measured in more than one location by the same rater (eg, left toe and right toe) or the same participant was measured by more than one rater.

Assessment of methodological quality

Methodological quality and risk of bias of included articles was performed independently by two reviewers (AM and SL) using the Quality Appraisal of Reliability (QAREL) Checklist and qualitative methodological assessment,[22] with disagreements arbitrated by a third reviewer (SC). Where data were incomplete or methodology unclear, relevant authors were contacted for clarification.

RESULTS

A total of 2431 articles were retrieved from the database search. Seventy-nine articles were identified as suitable for full-text review, of which 17 satisfied eligibility criteria for inclusion (figure 1). Seven articles were included in respective meta-analyses for individual test methods.

Characteristics and overview of included studies

The 17 included studies in this review reported a total of 1248 participants (table 1). Age (years) was reported as a mean 50–73,[23–31] range (8–89),[28 32] or unreported.[33–38] Sex was reported in 11 studies, with more males (n=617, 59%) overall than females, while sex was unreported in six studies.[33–37 39] Diabetes type was specified as type 1,[32 38] type 2,[23 25 26 39] both type 1 and type 2[24 27–29 31 33 34 36 37] or unreported.[30 35] Diabetes duration was reported in years as a mean 3–54,[23–25 27 31 32 38] range (0–63)[23 28 31 32] or unreported.[26 29 30 33–37 39] DPN diagnosis was reported in 65% of participants (range 3%–100%) across 12 studies,[23 24 26 27 30 32 34 36 37 39] and prevalence was unreported in six.[25 28 31 33 35 38] Nine studies assessed inter-rater

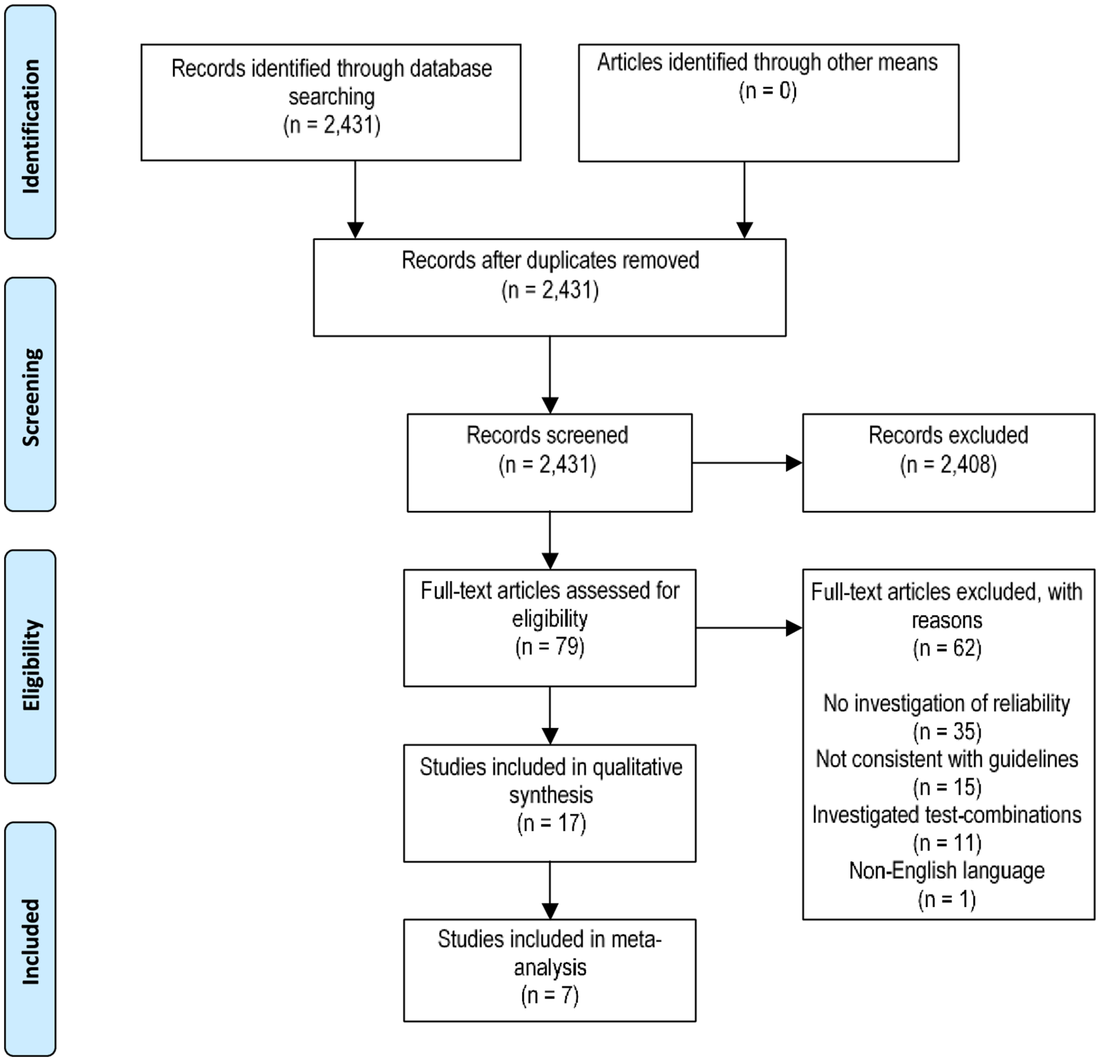
Identification
Screening
Eligibility
Included
Records identified through database searching
(n = 2,431)
Articles identified through other means
(n = 0)
Records after duplicates removed
(n = 2,431)
Records screened
(n = 2,431)
Records excluded
(n = 2,408)
Full-text articles assessed for eligibility
(n = 79)
Full-text articles excluded, with reasons
(n = 62)
No investigation of reliability
(n = 35)
Not consistent with guidelines
(n = 15)
Investigated test-combinations
(n = 11)
Non-English language
(n = 1)
Studies included in qualitative synthesis
(n = 17)
Studies included in meta-analysis
(n = 7)

Figure 1 PRISMA flow chart of search strategy. PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Table 1 Participant characteristics

Reference	Number	Sex (M/F)	Age (years)	Diabetes type	Diabetes duration (years)	Neuropathy (%)
Arshad *et al*[23]	105	44/61	50.90±11.53	Type 2	3 (0–23)	100
Bax *et al*[24]	100	58/42	57±8	Type 1: 8, type 2: 92	12±7	100
Bril *et al*[34]	42	NR	NR	NR	NR	100
Chew *et al*[35]	17	NR	NR	NR	NR	NR
Domínguez-Muñoz *et al*[25]	90	56/34	65.64±8.65	Type 2	9.96±8.83	NR
Gentile *et al*[36]	26	NR	NR	NR	NR	58
Guy *et al*[37]	20	Group 1 (DPN and ulcer): 5/5 Group 2 (Charcot joints): 4/8 Group 3 (autonomic neuropathy): 5/5 Group 4 (painful neuropathy): 6/9	Group 1 (DPN and ulcer): 42.5 (35–52) Group 2 (Charcot Joints): 36.9 (26–55) Group 3 (autonomic neuropathy): 31.7 (24–43) Group 4 (painful neuropathy): 35.9 (20–52)	Group 1 (DPN and ulcer): type 1: 7, type 2: 3 Group 2 (Charcot joints): type 1: 12 Group 3 (autonomic neuropathy): type 1: 10 group 4 (painful neuropathy): type 1: 14, type 2: 1	Group 1 (DPN and ulcer): 16.4 (7–30) Group 2 (Charcot joints): 21.7 (12–37) Group 3 (autonomic neuropathy): 15.5 (8–26) Group 4 (painful neuropathy): 14.7 (3–42)	100
Hirschfeld *et al*[38]	88	46/42	NR	Type 1	54.34±18.08	NR
Lanting *et al*[26]	50 (inter) MF and VPT 44 (intra) MF and VPT 24 (Intra and Inter) TF	33/17 15/9	M: 72±11, F: 73±7, M: 72±7, F: 70±9	Type 2	NR	17 4
Lasca *et al*[27]	60	Group A (DM+ulcer on one foot) 22/8 Group B (DM nil ulcer) 17/13	Group A (DM+ulcer on one foot) mean=60 Group B (DM nil ulcer) mean=59	Group A (DM+ulcer on one foot) type 1: 5, type 2: 25 Group B (DM nil ulcer) type 1: 5, type 2: 25	Group A (DM+ulcer on one foot) 16.83 Group B (DM nil ulcer) 14.80	100
Louraki *et al* (2013)[32]	118	58/60	13.5±3.4 (range: 8–20)	Type 1	5.7±3.5 (range: 2–16)	3
Maser *et al*[39]	52	NR	Range: 25–34	NR	NR	100
Paisley *et al*[33]	16	NR	NR	Type 1 and 2	NR	NR
Smieja *et al*[28]	200	142/58	63 (range: 18–89)	NR	Range 2 weeks–63 years	NR
Tentolouris *et al*[29]	100	57/43	62.3±12.5	Type 1: 30, Type 2: 70	NR	37.5
van Deursen *et al*[30]	15	9/6	62.1±8.4	NR	NR	100
Young *et al* (1992)[31]	85	60/15	Mean=61 (range: 21–82)	Type 1: 26, Type 2: 59	Mean=12 (range: 1–26)	NR

Age and diabetes duration are reported as mean±SD unless otherwise stated.
DM, diabetes mellitus; DPN, diabetes-related peripheral neuropathy; MF, monofilament; NR, not reported; TF, tuning fork; VPT, vibration perception threshold.

reliability,[23,27–30,33,35,38,39] six studies assessed intra-rater reliability[25,31,32,34,36,37] and two studies examined both inter-rater and intra-rater reliability[24,26] (table 1). Reliability was reported using Kappa statistic,[23,24,26,28,29,32,33,35,38,39] COV,[31,34,36] percentage agreement,[39] Spearman's rho[37] and ICC.[25,30]

Raters and measurement methods

There was little consistency between skills and qualifications of raters with experience ranging from doctors and specialists or people with specialized training with these devices[23,24,26,29,30,35,36,38,39] to internists.[28] Seven studies did not report the experience level of the raters.[25,27,31–34,37] Testing environment varied with eight studies reported to be in tertiary settings,[23,27,29,32,34,35,37,39] three in a secondary setting,[24,33,36] three in an education or university setting[26,28,30] and three did not specify the testing environment.[25,31,38] Furthermore, time periods between subsequent retests varied as four studies retested on the same day,[24,27,32,38] four retested within 1 week,[25,26,31,36] four retested within 1 month,[30,34,37,39] one retested within 2 months[23] and two studies did not specify their retesting periods.[28,33]

Methodological quality

All studies evaluated a relevant sample of participants, applied tests appropriately and used appropriate statistical measures of agreement (online supplemental table 1). The overall quality of studies varied however, primarily due to inconsistency in reporting on blinding of raters and participants, randomization of raters or assessments, and general methodology. For example, two studies did not blind raters to their own prior outcomes,[28,32] the results of the reference standard[28,38] and a further two did not blind raters to clinical information, which was not a part of the testing procedure.[28,30] Three studies did not blind raters to additional cues that were not a part of the test.[26,31,34] Therefore, the results of these studies need to be interpreted within the context of these limitations.

Four-site 10 g monofilament reliability

One study assessed the reliability of the four-site monofilament that reported moderate inter-rater reliability (κ=0.61) (table 2) and varied intra-rater reliability ranging from minimal to moderate (κ=0.34–0.67),[26] table 3.

128 Hz tuning fork reliability

Eight studies assessed inter-rater reliability of 128 Hz tuning fork that demonstrated a largely varied reliability ranging from none to strong (κ=0–0.86)[23,24,26–29,38,39] (table 2). Two studies reported intra-rater reliability of the 128 Hz tuning fork as weak to moderate agreement (κ=0.41–0.66)[24,26] (table 3).

VPT reliability

Four studies assessed inter-rater reliability of VPT through various modalities including the biothesiometer,[24,30] neurothesiometer[26,33] and maxivibrometer[30] (table 2).

Biothesiometer: one study reported weak to moderate reliability (κ=0.58–0.65)[24] and one study reported good reliability (ICC: 0.927).[30]

Neurothesiometer: two studies reported weak to moderate reliability (k=0.51–0.61).[26,33]

Maxivibrometer: one study reported good reliability (ICC: 0.961–0.958).[30]

Eight studies assessed intra-rater reliability of VPT through various modalities including biothesiometer,[24,31,32,36] neurothesiometer[26,31,34] and Vibratron II[25,34] (table 3).

Biothesiometer: two studies reported weak to moderate agreement (κ=0.51–0.81),[24,32] two reported high levels of agreement (COV (%)=8.6–18.6)[31,36] and one reported very strong reliability (rho=0.91).[37]

Neurothesiometer: one study reported weak to moderate agreement (κ=0.51)[26] and two studies reported excellent agreement (COV (%)=6–8.1).[31,34]

Vibration Sensitivity Tester (Vibratron II): one study reported moderate intra-rater reliability (COV (%)=31–34)[34] and one study reported excellent reliability.[25]

Pinprick reliability

Three studies assessed pinprick inter-rater reliability reporting minimal to weak reliability (κ=0.35–0.48)[28,33,39] (table 2). Intra-rater reliability was not reported.

Ankle reflex reliability

Four studies assessed ankle reflex inter-rater reliability and reported weak to moderate reliability (κ=0.58–0.60)[23,24,28,35] (table 2). One study assessed intra-rater reliability of ankle reflexes reporting weak to moderate agreement (κ=0.51–0.64)[24] (table 3).

Proprioception reliability

One study assessed inter-rater reliability of proprioception and reported minimal reliability (κ=0.28)[28] (table 2). Intra-rater reliability has not been examined.

Other recommended tests

Our literature search did not identify any investigations into the reliability of light touch/Ipswich touch test, three-site monofilament or temperature perception as performed according to current guidelines.

Meta-analyses

There were sufficient data from included studies to undertake meta-analyses of the inter-rater reliability of ankle reflexes,[23,24,28,35] pinprick,[28,39] 128 Hz tuning fork[23,24,26,28,29,39] and VPT (figure 2)[24,26] as well as the intra-rater reliability of 128 Hz tuning fork[24,26] and VPT (figure 3).[24,26,32]

Meta-analysis demonstrated the highest inter-rater reliability – reported as estimated Cohen's kappa (95% CI) – for VPT (κ=0.61 (0.50 to 0.73)), followed by ankle reflexes (κ=0.60 (0.55 to 0.64)), pinprick (κ=0.45 (0.22 to 0.69)) and 128 Hz tuning fork (κ=0.42 (0.15 to 0.70)).

Table 2 Inter-rater reliability of peripheral neurological tests, reported as Cohen's kappa (Κ), intraclass correlation coefficient (ICC) or per cent agreement

Reference	Four-site monofilament	128 Hz tuning fork	Vibration perception threshold	Pinprick	Reflexes	Proprioception
Arshad *et al*[23]		κ=0.33 (95% CI 0.15 to 0.51)			κ=0.58 (95% CI 0.45 to 0.71)	
Bax *et al*[24]		(Right) κ=0.49 (95% CI 0.30 to 0.69) (Left): κ=0.51 (95% CI 0.31 to 0.71)	Biothesiometer hallux (right): κ=0.65 (95% CI 0.46 to 0.84) Hallux (left): κ=0.58 (95% CI 0.40 to 0.70)		κ=0.58 (95% CI 0.48 to 0.72)	
Chew *et al*[35]					κ=0.6 (95% CI 0.19 to 1)	
Hirschfeld *et al*[38]		κ=0.07 (left) κ=0.007 (right)				
Lanting *et al*[26]	κ=0.61 (95% CI 0.45 to 0.77) p<0.01	κ=0.68 (95% CI 0.41 to 0.95), p<0.01	Neurothesiometer κ=0.61 (95% CI 0.45 to 0.77), p<0.01			
Lasca *et al*[27]		κ=0.86 p<0.001				
Maser *et al*[39]		κ=0.26 (% agreement: 75%)		κ=0.48 (% agreement: 81%)		
Paisley *et al* (2001)[33]			Neurothesiometer κ=0.51	κ=0.35		
Smieja *et al*[28]		κ=0.31 (95% CI 0.18 to 0.45)		κ=0.36 (95% CI 0.21 to 0.51)	κ=0.59 (95% CI 0.47 to 0.71)	κ=0.28 (95% CI 0.09 to 0.48)
Tentolouris *et al*[29]		κ=0.624 (95% CI 0.524 to 0.727) κ=0.678 (95% CI 0.576 to 0.780) κ=0.615 (95% CI 0.508 to 0.722)				
van Deursen *et al*[30]			Maxivibrometer ICC: 0.96 Biothesiometer ICC: 0.93			

Meta-analysis demonstrated the highest intra-rater reliability for VPT (κ=0.63 (0.45 to 0.81)), followed by the 128 Hz tuning fork (κ=0.54 (0.37 to 0.73)).

The trim-and-fill method used to detect and adjust for publication bias resulted in adjusted estimated Cohen's kappa (95% CI) for ankle reflexes (κ=0.60 (0.32 to 0.80)), pinprick (κ=0.48 (0.29 to 0.67)) and 128 Hz tuning fork (κ=0.32 (0.05 to 0.60)) (inter-rater reliability) and 128 Hz tuning fork (κ=0.53 (0.35 to 0.72)) (intra-rater reliability).

CONCLUSIONS

Of the recommended tests, included articles investigated the reliability of the four-site monofilament,[26] 128 Hz tuning fork,[23 24 26–29 38 39] VPT,[24–26 30–34 36 37] pinprick,[28 33 39] ankle reflex[23 24 28 35] and proprioception.[28] The findings of this review are that the inter-rater and intra-rater reliability of recommended neurological tests are largely varied when performed in people with diabetes. Based on the limited data available, results of pooled analyses suggest that VPT and ankle reflexes demonstrate acceptable reliability, whereas the reliability of pinprick and 128 Hz tuning fork tests is questionable. Additionally, cohort studies suggest that the four-site monofilament also demonstrates acceptable reliability,[26] whereas reliability of proprioception may be inadequate.[28] These findings should be considered in the context of the results of the QAREL assessment and the variability in methodological reporting, in conjunction with the wide CIs for the adjusted pooled estimates for the reliability

Table 3 Intra-rater reliability of peripheral neurological tests, reported as Cohen's kappa (Κ), coefficient of variance (COV) or Spearman's (r)

Reference	Four-site monofilament	128 Hz tuning fork	Vibration perception threshold	Reflexes
Bax *et al*[24]		(Right): rater A: κ=0.52, rater B: κ=0.66 (95% CI 0.45 to 0.73) (Left): rater A: κ=0.54, rater B: κ=0.56 (95% CI 0.41 to 0.69)	Biothesiometer hallux (right): rater A: κ=0.51, rater B: κ=0.57 (95% CI 0.40 to 0.68) Hallux (left): rater A: κ=0.64, rater B: κ=0.51 (95% CI 0.44 to 0.71)	Rater A: κ=0.55 Rater B: κ 0.55 (95% CI 0.41 to 0.69)
Bril *et al*[34]			Vibratron II COV (%)=31–34 Neurothesiometer COV (%)=6–8	
Domínguez-Muñoz *et al*[25]			Vibratron II ICC=0.958 (95% CI 0.94 to 0.98)	
Gentile *et al*[36]			Biothesiometer hallux: COV (%)=16.5 ± 5.8 (4–21) Malleous: COV (%)=18.6 ± 9.5 (4.4–28)	
Guy *et al*[37]			Biothesiometer *r*=0.91 (95% CI) p<0.01	
Lanting *et al*[26]	κ=0.34 (95% CI 0.06 to 0.63) p<0.01 to κ=0.67 (95% CI 0.45 to 0.89) p<0.01	κ=0.50 (95% CI 17 to 0.83) p<0.02 to κ=0.57 (95% CI 0.24 to 0.9) p<0.01	Neurothesiometer κ=0.52 (95% CI 0.21 to 0.82) p<0.01 to κ=0.78 (95% CI 0.58 to 0.98) p<0.02	
Louraki *et al* (2013)			Biothesiometer hallux (left): κ=0.69 (±0.05) Hallux (right): κ=0.64 (±0.05) Tibia (left): κ=0.70 Tibia (right): κ=0.64	
Young *et al* (1992)			Biothesiometer: COV (%): 8.6 Neurothesiometer: COV (%): 8.1	

ICC, intraclass correlation coefficient.

(eg, the intra-rater reliability of 128 Hz tuning fork (κ=0.32 (0.05 to 0.60)) and the variability of results that indicate available evidence is low or moderate quality. Of note, although included in IDF, IWGDF and ADA guidelines, we did not identify any article reporting the reliability of the three-site monofilament, light touch, Ipswich Touch Test or temperature perception tests in people with diabetes. These results need to be considered in light of the established predictive capacity for the development of foot wounds as demonstrated by the 10 g monofilament and 128 Hz tuning fork.[40 41]

The findings of this systematic review highlight the need for more exhaustive investigation of reliability of recommended chairside tests for DPN. A number of these studies assessing reliability for DPN testing reported that 100% of their population cohorts had DPN[23 24 27 30 34 37 39] making the weak to moderate reliability reported for both inter-rater and intra-rater reliability concerning. Although not inferring diagnostic accuracy, studies of reliability are affected by disease prevalence.[42] Therefore, when conducted in a cohort all with the target disease, the results are likely to overstate the reproducibility of the measurement.[42] In the case of tests such as monofilament testing for which pooled estimates of diagnostic accuracy have shown low sensitivity of 0.53 and adequate specificity of 0.88, the likelihood of a false negative test result is high for any given test point.[43] This is consistent with our findings of weak to moderate test reliability even in populations consisting entirely of participants with DPN. As chairside DPN testing is both used for the diagnosis and ongoing monitoring of DPN the usefulness of a test that has limited capacity to rule out the presence of the target disease or to reproduce a positive result in those with the disease is questionable. Furthermore, given that the earliest nerve damage in DPN is likely to be to small fibers,[44] reliability of chairside small-fiber tests is under investigated. We identified three studies that included investigation into the reliability of pinprick. However, we did not identify any tests investigating the reliability of thermal perception, and our present review did not investigate question-based tests such as the Total Symptom Score.[12] In this context, the reliability of large-fiber tests such as monofilament and vibration perception need to be considered together

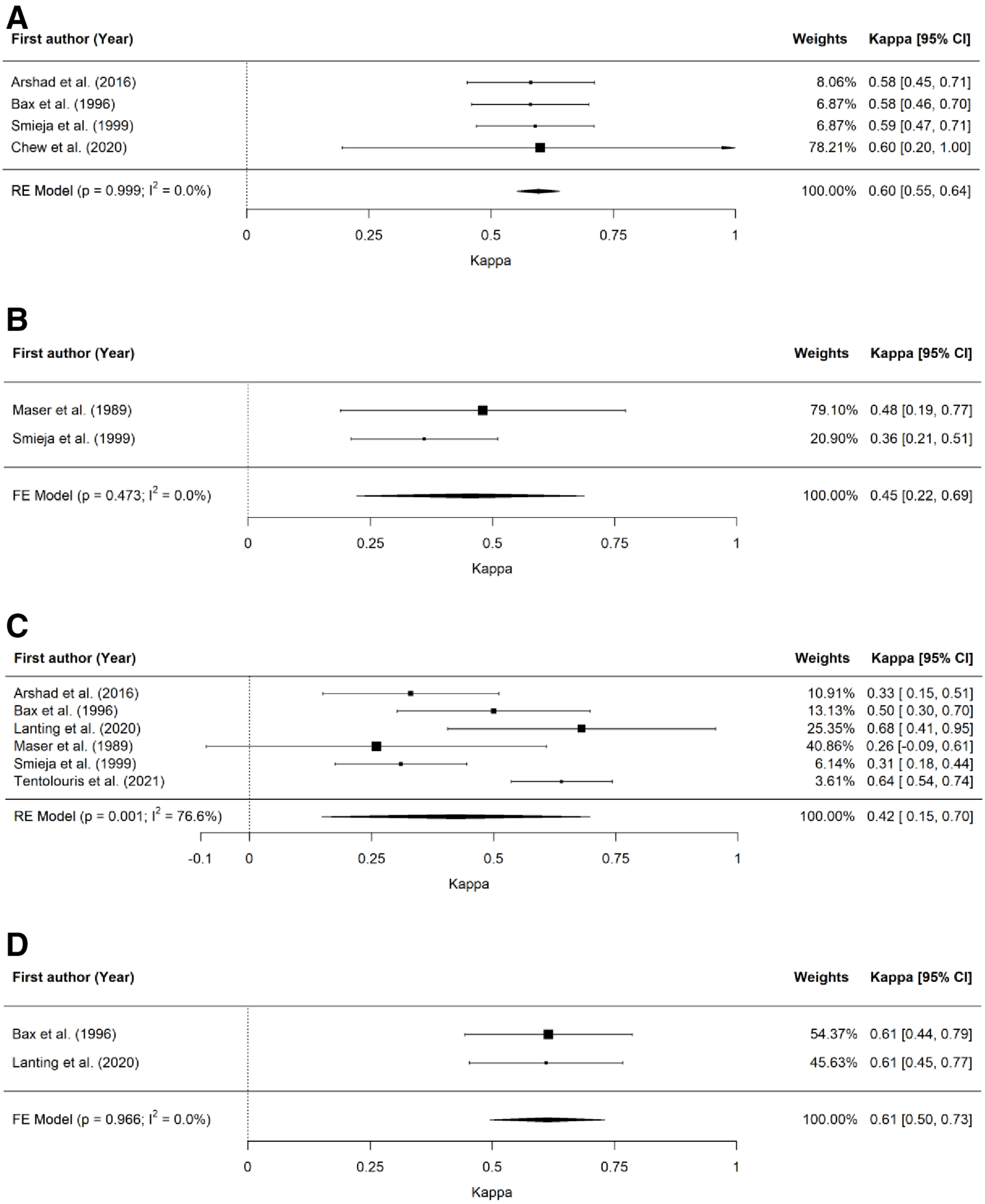

Figure 2 Forest plots for inter-rater reliability of screening tests for DPN. (A) Forest plot for inter-rater reliability of ankle reflex test. (B) Forest plot for inter-rater reliability of pinprick test. (C) Forest plot for inter-rater reliability of 128 Hz tuning fork test. (D) Forest plot for inter-rater reliability of vibration perception threshold test. DPN, diabetes-related peripheral neuropathy.

with their limited ability to detect early disease. Further research is thus warranted to determine the reliability of tests capable of detecting early disease.

Methodological differences between included studies is likely to have contributed to the range of results available in the literature. Reliability of various chairside tests was reportedly affected by limited training or variances in experience levels of clinicians[23 26 28 34 35 39] and also by inconsistent comprehension of individual test instructions by participants.[23 24 26 32 39] Tests such as the tuning fork, monofilament and pinprick all rely on application of controlled pressure by the clinician. As the rate of pressure is difficult to control for, especially between different raters, several studies identified this as possibly influencing test reliability.[23 24 26–28 38 39] These issues suggest that adequate clinician training should be undertaken, that the training is consistent with guidelines and that the instructions to patients should be clear, all of which may lead to improved reliability of chairside tests. Clinically, this can be improved through consideration

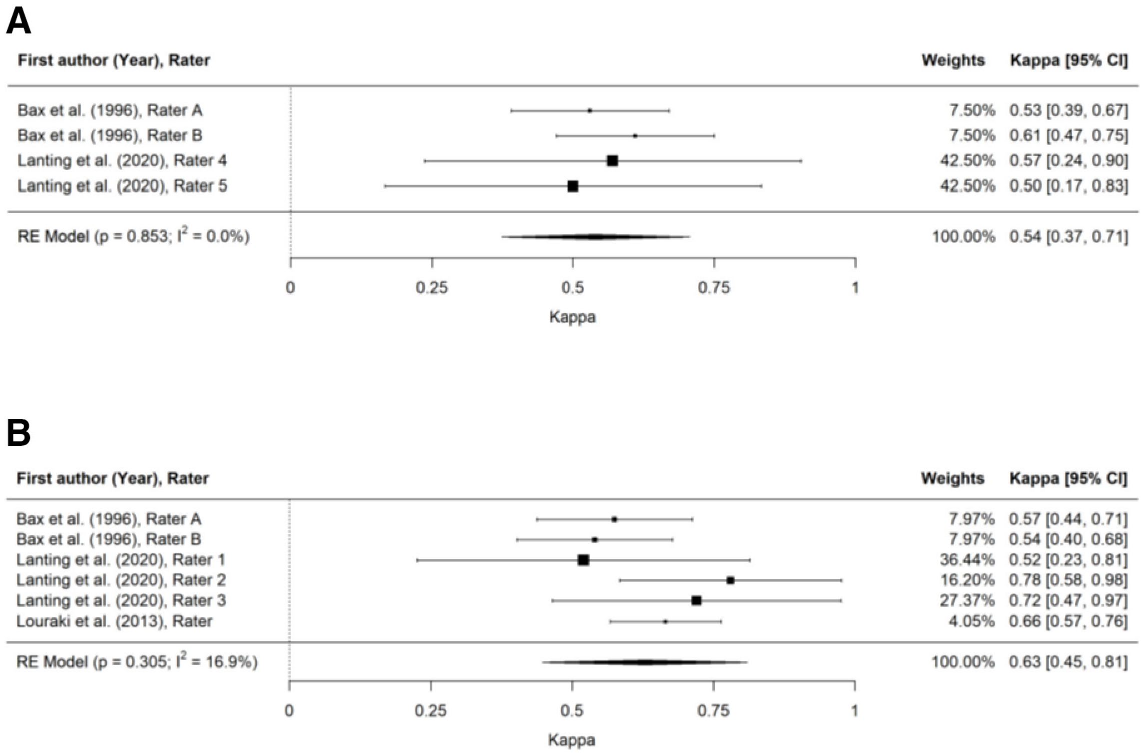

Figure 3 Forest plots for intra-rater reliability of screening tests for DPN. (A) Forest plot for intra-rater reliability of 128 Hz tuning fork test. (B) Forest plot for intra-rater reliability of vibration perception threshold test. DPN, diabetes-related peripheral neuropathy.

of recommendations from current guidelines regarding test technique and test sites.[12–14] The included literature is limited by use of small sample sizes,[26 34 35 37 39] lack of blinding of assessors to previous results[28 30] and heterogeneity of measures of statistical agreement used. Although the majority of studies used kappa values, some used COV, Spearman's rho, percentage agreement or ICCs, making comparison of available data across testing methods challenging.

This review has highlighted the need for further investigation of reliability of chairside DPN testing. Due to the range of reliability and varied reliability measures across all recommended neurological tests, it is suggested that there be more extensive research into the reliability of pinprick, proprioception and other recommended chairside DPN tests that have not been investigated. Furthermore, future research should be conducted in specific populations with diabetes and be conducted in populations where prevalence of DPN has been established through testing methods with high diagnostic accuracy. Given the additional impacts of age on neurological and cognitive function beyond those results from diabetes, there may be age-specific differences in reliability of chairside tests, and as such, investigations taking age into account are required. To this end, simplifying neurological testing will allow clinicians and patients to better communicate test instructions as well as reduce the variability between clinicians when performing the tests to improve overall reliability. Furthermore, increased clinical knowledge of reliability of neurological screening tests allows for more informed clinical decision making when selecting multiple tests (eg, monofilament and tuning fork) to aid in the diagnosis and monitoring of DPN.

Although the search strategy employed in this review was designed to be robust, there may be some evidence that was not captured, for example, unpublished data. It should also be acknowledged that the reliability of chairside tests included in this review are from three international consensus statements only. Other commonly used chairside neuropathy tests that warrant further investigation include the monofilament test using additional sites for all cause peripheral neuropathy,[45] conventional and graduated tuning forks,[46] two-point discrimination,[47] temperature sensation and the Michigan Neuropathy Screening Instrument.[48] Lastly, future studies investigating test reliability should ensure adequate reporting, sufficient detail for cohort characteristics, methodology and appropriate statistical tests, for example, kappa or intraclass correlation coefficients with relevant CIs.

The results of this systematic review found evidence of acceptable reliability for VPT using a biothesiometer, neurothesiometer or maxivibrometer, ankle reflexes and the four-site monofilament test. Due to the large range of reported reliability for the 128 Hz tuning fork, we are unable to appropriately comment on this testing method. These results support the clinical use of these identified tests for screening and ongoing monitoring of DPN as recommended by the latest guidelines by IDF, IWGDF and ADA, respectively. The reliability of temperature perception (IDF and ADA), pinprick, proprioception (ADA), three-site monofilament and Ipswich touch test (IWGDF) when performed in people with diabetes remains unclear and warrants investigation to determine their suitability for use for testing in this population.

Contributors AM, SL and VC conceived the review. AM and SL conducted the search, extracted data and performed quality appraisal. DL, LL, SL, SC and VC contributed to statistical analysis and interpretation. AM, SL and VC drafted the manuscript. All authors contributed to and approved the final version of the manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study does not involve human participants.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

ORCID iDs

Sean Lanting http://orcid.org/0000-0001-5596-7778
David Lambkin http://orcid.org/0000-0001-5729-1927
Lucy Leigh http://orcid.org/0000-0002-5198-8843
Sarah Casey http://orcid.org/0000-0003-4024-1012
Vivienne Chuter http://orcid.org/0000-0003-4793-5340

REFERENCES

1 Cho NH, Shaw JE, Karuranga S, *et al*. IDF diabetes atlas: global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Res Clin Pract* 2018;138:271–81.
2 Boulton AJM, Vinik AI, Arezzo JC, *et al*. Diabetic neuropathies: a statement by the American diabetes association. *Diabetes Care* 2005;28:956–62.
3 Mueller MJ, Minor SD, Sahrmann SA, *et al*. Differences in the gait characteristics of patients with diabetes and peripheral neuropathy compared with age-matched controls. *Phys Ther* 1994;74:299–308.
4 Armstrong DG, Lavery LA, Vela SA, *et al*. Choosing a practical screening instrument to identify patients at risk for diabetic foot ulceration. *Arch Intern Med* 1998;158:289–92.
5 Ang L, Jaiswal M, Martin C, *et al*. Glucose control and diabetic neuropathy: lessons from recent large clinical trials. *Curr Diab Rep* 2014;14:528.
6 Pop-Busui R, Lu J, Brooks MM, *et al*. Impact of glycemic control strategies on the progression of diabetic peripheral neuropathy in the bypass angioplasty revascularization investigation 2 diabetes (Bari 2D) cohort. *Diabetes Care* 2013;36:3208–15.
7 Young MJ, Boulton AJ, MacLeod AF, *et al*. A multicentre study of the prevalence of diabetic peripheral neuropathy in the United Kingdom Hospital clinic population. *Diabetologia* 1993;36:150–4.
8 Armstrong DG, Harkless LB. Outcomes of preventative care in a diabetic foot specialty clinic. *J Foot Ankle Surg* 1998;37:460–6.
9 Plank J, Haas W, Rakovac I, *et al*. Evaluation of the impact of chiropodist care in the secondary prevention of foot ulcerations in diabetic subjects. *Diabetes Care* 2003;26:1691–5.
10 Diabetes Control and Complications Trial Research Group, Nathan DM, Genuth S, *et al*. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *N Engl J Med* 1993;329:977–86.
11 Turner RC, Holman RR, Stratton I. Effect of intensive blood-glucose control with metformin on complications in overweight patients with type 2 diabetes (UKPDS 34). *Lancet* 1998;352:854–65.
12 Ibrahim A. IDF clinical practice recommendation on the diabetic foot: a guide for healthcare professionals. *Diabetes Res Clin Pract* 2017;127:285–7.
13 Schaper NC, Netten JJ, Apelqvist J, *et al*. Practical guidelines on the prevention and management of diabetic foot disease (IWGDF 2019 update). *Diabetes Metab Res Rev* 2020;36:e3266.
14 Pop-Busui R, Boulton AJM, Feldman EL, *et al*. Diabetic neuropathy: a position statement by the American diabetes association. *Diabetes Care* 2017;40:136–54.
15 Bril V, Tomioka S, Buchanan RA, *et al*. Reliability and validity of the modified Toronto clinical neuropathy score in diabetic sensorimotor polyneuropathy. *Diabet Med* 2009;26:240–6.
16 McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med* 2012;22:276–82.
17 Rodbard D. Clinical interpretation of indices of quality of glycemic control and glycemic variability. *Postgrad Med* 2011;123:107–18.
18 Portney LG, Watkins MP. *Foundations of clinical research: applications to practice*. Upper Saddle River, NJ: Pearson/Prentice Hall, 2009.
19 Prion S, Haerling KA. Making sense of methods and measurement: Spearman-rho ranked-order correlation coefficient. *Clin Simul Nurs* 2014;10:535–6.
20 Sun S. Meta-analysis of Cohen's kappa. *Health Serv Outcomes Res Method* 2011;11:145–63.
21 Viechtbauer W. Conducting Meta-Analyses in *R* with the meta*for* Package. *J Stat Softw* 2010;36:1–48.
22 Lucas NP, Macaskill P, Irwig L, *et al*. The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). *J Clin Epidemiol* 2010;63:854–61.
23 Arshad AR, Hussain K, Masood J. Interobserver agreement in eliciting signs of peripheral neuropathy in patients with diabetes mellitus type 2. *Rawal Medical J* 2016;41:171–4.
24 Bax G, Fagherazzi C, Piarulli F, *et al*. Reproducibility of Michigan neuropathy screening instrument (MNSI). A comparison with tests using the vibratory and thermal perception thresholds. *Diabetes Care* 1996;19:904–5.
25 Domínguez-Muñoz FJ, Adsuar JC, Villafaina S, *et al*. Test-Retest reliability of vibration perception threshold test in people with type 2 diabetes mellitus. *Int J Environ Res Public Health* 2020;17:1773.
26 Lanting SM, Spink MJ, Tehan PE, *et al*. Non-Invasive assessment of vibration perception and protective sensation in people with diabetes mellitus: inter- and intra-rater reliability. *J Foot Ankle Res* 2020;13:1–7.
27 Lasca M, Tăut AL, Vereşiu IA. Comparative evaluation of several simple screening tests for risk of neuropathic ulcerations of feet in patients with diabetes mellitus. *Rom J Diabetes Nutr Metab Dis* 2016;23:67–72.
28 Smieja M, Hunt DL, Edelman D, *et al*. Clinical examination for the detection of protective sensation in the feet of diabetic patients. *J Gen Intern Med* 1999;14:418–24.
29 Tentolouris A, Tentolouris N, Eleftheriadou I. The performance and interrater agreement of vibration perception for the diagnosis of loss of protective sensation in people with diabetes mellitus. *Int J Low Extrem Wounds* 2021;1534734621994058.
30 van Deursen RW, Sanchez MM, Derr JA, *et al*. Vibration perception threshold testing in patients with diabetic neuropathy: ceiling effects and reliability. *Diabet Med* 2001;18:469–75.
31 Young MJ, Every N, Boulton AJ. A comparison of the neurothesiometer and biothesiometer for measuring vibration perception in diabetic patients. *Diabetes Res Clin Pract* 1993;20:129–31.
32 Louraki M, Tsentidis C, Kallinikou D, *et al*. Reproducibility of vibration perception threshold values in children and adolescents with type 1 diabetes mellitus and associated factors. *Prim Care Diabetes* 2014;8:147–57.
33 Paisley AN, Abbott CA, van Schie CHM, *et al*. A comparison of the Neuropen against standard quantitative sensory-threshold measures for assessing peripheral nerve function. *Diabet Med* 2002;19:400–5.
34 Bril V, Kojic J, Ngo M, *et al*. Comparison of a neurothesiometer and vibration in measuring vibration perception thresholds and relationship to nerve conduction studies. *Diabetes Care* 1997;20:1360–2.
35 Chew BLA, Williams DB, Attia J. The diagnostic value of clinical examination for identifying patients with large and small fibre

neuropathy. *Intern Med J* 2020. doi:10.1111/imj.15079. [Epub ahead of print: 05 Oct 2020].

36 Gentile S, Turco S, Corigliano G, *et al*. Simplified diagnostic criteria for diabetic distal polyneuropathy. Preliminary data of a multicentre study in the Campania region. S.I.M.S.D.N. group. *Acta Diabetol* 1995;32:7–12. doi:10.1007/BF00581037

37 Guy RJ, Clark CA, Malcolm PN, *et al*. Evaluation of thermal and vibration sensation in diabetic neuropathy. *Diabetologia* 1985;28:131–7.

38 Hirschfeld G, von Glischinski M, Knop C, *et al*. Difficulties in screening for peripheral neuropathies in children with diabetes. *Diabet Med* 2015;32:786–9.

39 Maser RE, Nielsen VK, Bass EB, *et al*. Measuring diabetic neuropathy. assessment and comparison of clinical examination and quantitative sensory testing. *Diabetes Care* 1989;12:270–5.

40 Morshed GM, Mashahit MA, Shaheen HA. Simple screening tests for peripheral neuropathy as a prediction of diabetic foot ulceration. *FAQJ* 2011;4:2.

41 Young MJ, Breddy JL, Veves A, *et al*. The prediction of diabetic neuropathic foot ulceration using vibration perception thresholds. A prospective study. *Diabetes Care* 1994;17:557–60.

42 Gjørup T. Reliability of diagnostic tests. *Acta Obstet Gyn Scan* 1997;76:9–14.

43 Wang F, Zhang J, Yu J, *et al*. Diagnostic accuracy of monofilament tests for detecting diabetic peripheral neuropathy: a systematic review and meta-analysis. *J Diabetes Res* 2017;2017:1–12.

44 Malik RA, Veves A, Tesfaye S, *et al*. Small fibre neuropathy: role in the diagnosis of diabetic sensorimotor polyneuropathy. *Diabetes Metab Res Rev* 2011;27:678–84.

45 Mawdsley RH, Behm-Pugh AT, Campbell JD, *et al*. Reliability of measurements with Semmes-Weinstein Monofilaments in individuals with diabetes. *Phys Occup Ther Geriatr* 2004;22:19–36.

46 Thivolet C, el Farkh J, Petiot A, *et al*. Measuring vibration sensations with graduated tuning fork. simple and reliable means to detect diabetic patients at risk of neuropathic foot ulceration. *Diabetes Care* 1990;13:1077–80.

47 Ferreira MC, Rodrigues L, Fels K. New method for evaluation of cutaneous sensibility in diabetic feet: preliminary report. *Rev Hosp Clin Fac Med Sao Paulo* 2004;59:286–90.

48 Lunetta M, Le Moli R, Grasso G, *et al*. A simplified diagnostic test for ambulatory screening of peripheral diabetic neuropathy. *Diabetes Res Clin Pract* 1998;39:165–72.