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 Chuter1

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 Megafail 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-diabetes7. 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: 10g monofilament, 128Hz tuning fork, light touch/Ipswich touch test, temperature perception, vibration perception threshold (VPT), pinprick, 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
Pathophysiology/complications

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 Megafille 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:

\[
\text{variance} = \frac{p_0 \times (1-p_0)}{1-p_0} \times \frac{1}{n}
\]

- 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:

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

Results of the meta-analysis are presented as estimated Cohen’s kappa (95% CI) and total heterogeneity (I²) 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 (e.g., 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, 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, range (8–89), or unreported. Sex was reported in 11 studies, with more males (n=617, 59%) overall than females, while sex was unreported in six studies. Diabetes type was specified as type 1, type 2, both type 1 and type 2, or unreported. Diabetes duration was reported in years as a mean 3–54, range (0–63) or unreported. DPN diagnosis was reported in 65% of participants (range 3%–100%) across 12 studies, and prevalence was unreported in six. Nine studies assessed inter-rater...
<table>
<thead>
<tr>
<th>Reference</th>
<th>Number</th>
<th>Sex (M/F)</th>
<th>Age (years)</th>
<th>Diabetes type</th>
<th>Diabetes duration (years)</th>
<th>Neuropathy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arshad et al</td>
<td>105</td>
<td>44/61</td>
<td>50.90±11.53</td>
<td>Type 2</td>
<td>3 (0–23)</td>
<td>100</td>
</tr>
<tr>
<td>Bax et al</td>
<td>100</td>
<td>58/42</td>
<td>57±8</td>
<td>Type 1: 8, type 2: 92</td>
<td>12±7</td>
<td>100</td>
</tr>
<tr>
<td>Bril et al</td>
<td>42</td>
<td>NR</td>
<td>NR</td>
<td>Type 2</td>
<td>Type 2: 12</td>
<td>100</td>
</tr>
<tr>
<td>Chew et al</td>
<td>17</td>
<td>NR</td>
<td>NR</td>
<td>Type 2</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>Domínguez-Muñoz et al</td>
<td>90</td>
<td>56/34</td>
<td>65.64±8.65</td>
<td>Type 2</td>
<td>9.96±8.83</td>
<td>NR</td>
</tr>
<tr>
<td>Gentile et al</td>
<td>26</td>
<td>NR</td>
<td>NR</td>
<td>Type 2</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>Hirschfeld et al</td>
<td>88</td>
<td>46/42</td>
<td>NR</td>
<td>Type 1</td>
<td>54.34±18.08</td>
<td>NR</td>
</tr>
<tr>
<td>Lanting et al</td>
<td>50 (inter)</td>
<td>MF and VPT 44 (intra) MF and VPT 24 (Intra and Inten) TF</td>
<td>33/17</td>
<td>M: 72±11, F: 73±7, M: 72±7, F: 70±9</td>
<td>Type 2</td>
<td>54.34±18.08</td>
</tr>
<tr>
<td>Guy et al</td>
<td>20</td>
<td>Group A (DM+ulcer on one foot) Group B (DM nil ulcer)</td>
<td>mean=60 Group B (DM nil ulcer) mean=59</td>
<td>Group A (DM+ulcer on one foot) type 1: 5, type 2: 25 Group B (DM nil ulcer) type 1: 5, type 2: 25</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>Hirschfeld et al</td>
<td>88</td>
<td>46/42</td>
<td>NR</td>
<td>Type 1</td>
<td>54.34±18.08</td>
<td>NR</td>
</tr>
<tr>
<td>Lanting et al</td>
<td>50 (inter)</td>
<td>MF and VPT 44 (intra) MF and VPT 24 (Intra and Inten) TF</td>
<td>33/17</td>
<td>M: 72±11, F: 73±7, M: 72±7, F: 70±9</td>
<td>Type 2</td>
<td>54.34±18.08</td>
</tr>
<tr>
<td>Lasca et al</td>
<td>60</td>
<td>Group A (DM+ulcer on one foot) Group B (DM nil ulcer)</td>
<td>mean=60 Group B (DM nil ulcer) mean=59</td>
<td>Group A (DM+ulcer on one foot) type 1: 5, type 2: 25 Group B (DM nil ulcer) type 1: 5, type 2: 25</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>Louraki et al (2013)</td>
<td>118</td>
<td>58/60</td>
<td>13.5±3.4 (range: 8–20)</td>
<td>Type 1</td>
<td>5.7±3.5 (range: 2–16)</td>
<td>3</td>
</tr>
<tr>
<td>Maser et al</td>
<td>52</td>
<td>NR</td>
<td>Range: 25–34</td>
<td>Type 1 and 2</td>
<td>NR</td>
<td>100</td>
</tr>
<tr>
<td>Paisley et al</td>
<td>16</td>
<td>NR</td>
<td>NR</td>
<td>Type 1 and 2</td>
<td>NR</td>
<td>100</td>
</tr>
<tr>
<td>Smieja et al</td>
<td>200</td>
<td>142/58</td>
<td>63 (range: 18–89)</td>
<td>NR</td>
<td>Range 2 weeks–63 years</td>
<td>NR</td>
</tr>
<tr>
<td>Tentolouris et al</td>
<td>100</td>
<td>57/43</td>
<td>62.3±12.5</td>
<td>Type 1: 30, Type 2: 70</td>
<td>NR</td>
<td>37.5</td>
</tr>
<tr>
<td>van Deursen et al</td>
<td>15</td>
<td>9/6</td>
<td>62.1±8.4</td>
<td>Type 1: 26, Type 2: 59</td>
<td>Mean=12 (range: 1–26)</td>
<td>100</td>
</tr>
<tr>
<td>Young et al (1992)</td>
<td>85</td>
<td>60/15</td>
<td>Mean=61 (range: 21–82)</td>
<td>Type 1: 26, Type 2: 59</td>
<td>Mean=12 (range: 1–26)</td>
<td>100</td>
</tr>
</tbody>
</table>

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, and two studies examined both inter-rater and intra-rater reliability (table 1). Reliability was reported using Kappa statistic (κ) 24 26 30 34 35 37 and ICC. 25 30 Percentage agreement, Spearman’s ρ 27 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 33 36 38 39 to interns. 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 25 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, 26 30 34 37 39 one retested within 2 months, 25 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=0.86) 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).

Bioteiometer: 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 (κ=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 35 neurothesiometer 26 31 34 and Vibratron II 25 34 (table 3).

Bioteiometer: 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 34 and one reported very strong reliability (rho=0.91). 37

Neurothesiometer: one study reported weak to moderate agreement (κ=0.51) 38 and two studies reported excellent agreement (COV (%)=6–8.1). 31 34

Vibration Sensitivity Tester (Vibratron II): one study undertook meta-analyses of the inter-rater reliability of ankle reflexes (COV (%)=51–34) 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) 28 (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)).
Meta-analysis demonstrated the highest intra-rater reliability for VPT ($\kappa=0.63$ (95% CI 0.45 to 0.81)), followed by the 128 Hz tuning fork ($\kappa=0.54$ (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 ($\kappa=0.60$ (0.32 to 0.80)), pinprick ($\kappa=0.48$ (0.29 to 0.67)) and 128 Hz tuning fork ($\kappa=0.32$ (0.05 to 0.60)) (inter-rater reliability) and 128 Hz tuning fork ($\kappa=0.53$ (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 28 35 VPT,24–26 30–34 36 37 pinprick,28 33 39 ankle reflex 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>
<thead>
<tr>
<th>Reference</th>
<th>Four-site monofilament</th>
<th>128 Hz tuning fork</th>
<th>Vibration perception threshold</th>
<th>Pinprick</th>
<th>Reflexes</th>
<th>Proprioception</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arshad et al$^{23}$</td>
<td>$\kappa=0.33$ (95% CI 0.15 to 0.51)</td>
<td></td>
<td></td>
<td>$\kappa=0.58$ (95% CI 0.45 to 0.71)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bax et al$^{24}$</td>
<td>(Right) $\kappa=0.49$ (95% CI 0.30 to 0.69) (Left): $\kappa=0.51$ (95% CI 0.31 to 0.71)</td>
<td>Biothesiometer hallux (right): $\kappa=0.65$ (95% CI 0.46 to 0.84) Hallux (left): $\kappa=0.58$ (95% CI 0.40 to 0.70)</td>
<td>$\kappa=0.58$ (95% CI 0.48 to 0.72)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chew et al$^{35}$</td>
<td>$\kappa=0.07$ (left) $\kappa=0.007$ (right)</td>
<td></td>
<td></td>
<td>$\kappa=0.6$ (95% CI 0.19 to 1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hirschfeld et al$^{38}$</td>
<td>$\kappa=0.61$ (95% CI 0.45 to 0.77) $p&lt;0.01$</td>
<td>$\kappa=0.68$ (95% CI 0.41 to 0.95), $p&lt;0.01$</td>
<td>Neurothesiometer $\kappa=0.61$ (95% CI 0.45 to 0.77), $p&lt;0.01$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lanting et al$^{26}$</td>
<td>$\kappa=0.68$ (95% CI 0.41 to 0.95), $p&lt;0.01$</td>
<td>$\kappa=0.68$ (95% CI 0.41 to 0.95), $p&lt;0.01$</td>
<td>Neurothesiometer $\kappa=0.61$ (95% CI 0.45 to 0.77), $p&lt;0.01$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lasca et al$^{27}$</td>
<td>$\kappa=0.86$ $p&lt;0.001$</td>
<td>$\kappa=0.51$</td>
<td></td>
<td>$\kappa=0.35$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maser et al$^{39}$</td>
<td>$\kappa=0.26$ (% agreement: 75%)</td>
<td>$\kappa=0.31$ (95% CI 0.18 to 0.45)</td>
<td>$\kappa=0.36$ (95% CI 0.21 to 0.51) $\kappa=0.59$ (95% CI 0.47 to 0.71) $\kappa=0.28$ (95% CI 0.09 to 0.48)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paisley et al (2001)$^{33}$</td>
<td>$\kappa=0.61$ (95% CI 0.45 to 0.77)</td>
<td>$\kappa=0.68$ (95% CI 0.52 to 0.727) $\kappa=0.678$ (95% CI 0.576 to 0.780) $\kappa=0.615$ (95% CI 0.508 to 0.722)</td>
<td>$\kappa=0.61$ (95% CI 0.45 to 0.77) $\kappa=0.678$ (95% CI 0.576 to 0.780) $\kappa=0.615$ (95% CI 0.508 to 0.722)</td>
<td>$\kappa=0.61$ (95% CI 0.45 to 0.77) $\kappa=0.678$ (95% CI 0.576 to 0.780) $\kappa=0.615$ (95% CI 0.508 to 0.722)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smieja et al$^{28}$</td>
<td>$\kappa=0.624$ (95% CI 0.524 to 0.727) $\kappa=0.678$ (95% CI 0.576 to 0.780) $\kappa=0.615$ (95% CI 0.508 to 0.722)</td>
<td>$\kappa=0.36$ (95% CI 0.21 to 0.51) $\kappa=0.59$ (95% CI 0.47 to 0.71) $\kappa=0.28$ (95% CI 0.09 to 0.48)</td>
<td>$\kappa=0.61$ (95% CI 0.45 to 0.77) $\kappa=0.678$ (95% CI 0.576 to 0.780) $\kappa=0.615$ (95% CI 0.508 to 0.722)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tentolouris et al$^{29}$</td>
<td>$\kappa=0.61$ (95% CI 0.45 to 0.77)</td>
<td>$\kappa=0.61$ (95% CI 0.45 to 0.77) $\kappa=0.615$ (95% CI 0.508 to 0.722)</td>
<td>$\kappa=0.61$ (95% CI 0.45 to 0.77) $\kappa=0.615$ (95% CI 0.508 to 0.722)</td>
<td>$\kappa=0.61$ (95% CI 0.45 to 0.77) $\kappa=0.615$ (95% CI 0.508 to 0.722)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>van Deursen et al$^{36}$</td>
<td>$\kappa=0.624$ (95% CI 0.524 to 0.727)</td>
<td>$\kappa=0.615$ (95% CI 0.508 to 0.722)</td>
<td>$\kappa=0.61$ (95% CI 0.45 to 0.77) $\kappa=0.615$ (95% CI 0.508 to 0.722)</td>
<td>$\kappa=0.61$ (95% CI 0.45 to 0.77) $\kappa=0.615$ (95% CI 0.508 to 0.722)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Inter-rater reliability of peripheral neurological tests, reported as Cohen’s kappa (K), intraclass correlation coefficient (ICC) or per cent agreement
(eg, the intra-rater reliability of 128 Hz tuning fork ($\kappa = 0.32$ (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 DPN23 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.

### Table 3

<table>
<thead>
<tr>
<th>Reference</th>
<th>Four-site monofilament</th>
<th>128 Hz tuning fork</th>
<th>Vibration perception threshold</th>
<th>Reflexes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bax et al24</td>
<td>(Right): rater A: $\kappa = 0.52$, rater B: 0.66 (95% CI 0.45 to 0.73) (Left): rater A: $\kappa = 0.54$, rater B: 0.56 (95% CI 0.41 to 0.69)</td>
<td>Biothesiometer hallux (right): rater A: $\kappa = 0.51$, rater B: 0.57 (95% CI 0.40 to 0.68) Hallux (left): rater A: $\kappa = 0.64$, rater B: 0.51 (95% CI 0.44 to 0.71)</td>
<td>Rater A: $\kappa = 0.55$, Rater B: $\kappa = 0.55$ (95% CI 0.41 to 0.69)</td>
<td></td>
</tr>
<tr>
<td>Bril et al34</td>
<td></td>
<td>Vibratron II COV (%)=31–34 Neurothesiometer COV (%)=6–8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dominguez-Muñoz et al25</td>
<td></td>
<td>Vibratron II ICC=0.958 (95% CI 0.94 to 0.98)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gentle et al36</td>
<td></td>
<td>Biothesiometer hallux: COV (%)=16.5 ± 5.8 (4–21) Malleous: COV (%)=18.6 ± 9.5 (4.4–28)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Guy et al37</td>
<td></td>
<td>Biothesiometer $r=0.91$ (95% CI) p&lt;0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lanting et al26</td>
<td>$\kappa=0.34$ (95% CI 0.06 to 0.83) p&lt;0.01 to $\kappa=0.67$ (95% CI 0.45 to 0.89) p&lt;0.01</td>
<td>$\kappa=0.50$ (95% CI 0.17 to 0.83) p&lt;0.02 to $\kappa=0.57$ (95% CI 0.24 to 0.9) p&lt;0.01</td>
<td>Neurothesiometer $\kappa=0.52$ (95% CI 0.21 to 0.82) p&lt;0.01 to $\kappa=0.78$ (95% CI 0.58 to 0.96) p&lt;0.02</td>
<td></td>
</tr>
<tr>
<td>Louraki et al (2013)</td>
<td></td>
<td>Biothesiometer hallux (left): $\kappa=0.69$ (±0.05) Hallux (right): $\kappa=0.64$ (±0.05) Tibia (left): $\kappa=0.70$ Tibia (right): $\kappa=0.64$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ICC, intraclass correlation coefficient.
Pathophysiology/complications

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 and also by inconsistent comprehension of individual test instructions by participants. 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. 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 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.
of recommendations from current guidelines regarding test technique and test sites. The included literature is limited by use of small sample sizes, lack of blinding of assessors to previous results 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, conventional and graduated tuning forks, two-point discrimination, temperature sensation and the Michigan Neuropathy Screening Instrument. 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, terminologies, 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/0000-0002-5198-8943
Sarah Casey http://orcid.org/0000-0003-4024-1012
Vivienne Chuter http://orcid.org/0000-0003-4793-5340

REFERENCES
35 Chew BLA, Williams DB, Attia J. The diagnostic value of clinical examination for identifying patients with large and small fibro
Pathophysiology/complications


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


