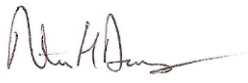
A multicenter comparative diagnostic accuracy study

Diagnostic tools to establish the presence and severity of peripheral arterial disease in people with diabetes **(DM PAD)**

Statistical Analysis Plan

CONFIDENTIAL

Version No	V1.0
Date Finalised	14 November 2023
Author(s)	Roz Pollock (until unblinded on 01 Feb 2023), Sharon Tuck (from Feb 2023)
CI Name	Mr Usman Jaffer, St Marys Hospital
CI Email address	usmanjaffer.pa@outlook.com; a.h.davies@imperial.ac.uk

Signatures	
Trial Statistician: Prof. John Norrie 	**Date: 2023-11-14**
Chief Investigator: Mr. Usman Jaffer 	**Date: 2023-11-14**
Designee Chief Investigator: Prof. Alun Davies 	**Date: 2023-11-14**

Document Control		
Version No	**Date**	**Summary of Revisions**
1.0	14Nov2023	Initial Creation

Table of Contents

List of Abbreviations

Abbreviation	Full name
ABPI	Ankle brachial pressure index
AE	Adverse Event
BMI	Body Mass Index
CI	Chief Investigator
CTA	Computed tomography angiography
DFU	Diabetic foot ulcer
DP	Dorsal Pedis
ECTU	Edinburgh Clinical Trials Unit
eGFR	Estimated glomerular filtration rate
EQ-5D-5L	European Quality of Life-5 dimensions 5 Level
IQR	Inter-quartile range
MRA	Magnetic resonance angiograph
PAD-scan	Podiatry ankle duplex scan
PT	Posterior Tibial
QALY	Quality Adjusted Life Years
ROC	Receiver Operating Characteristics
SAE	Serious Adverse Event
SAP	Statistical Analysis Plan
SOP	Standard Operating Procedure
TBPI	Toe-brachial pressure index
TcPO2	Transcutaneous pressure of oxygen
TrEAD	TEsting for Arterial disease in Diabetes study
WIfI	Wound, ischaemia, and foot infection

1. Introduction

This is a prospective multicentre diagnostic accuracy study comparing the diagnostic performance, powered to compare five index tests (the sixth exploratory test in four sites) in the diagnosis of peripheral arterial disease (PAD) in patients with diabetes against a reference test (CTA or MRA). The tests shall be performed in the following order:

1. Audible handheld Doppler
   2a. Visual handheld Doppler
   2b. PAD-scan (in 4 selected centres)[1]
3. Toe-brachial pressure index (TBPI)
4. Ankle-brachial pressure index (ABPI)
5. Exercise ABPI i.e. ABPI performed following repetitive heal raising.

Participants will firstly be assessed via remote screening. If considered eligible, they will be invited in for a face-to-face appointment (i.e. Visit 1) alongside a routine/planned visit. They will then be invited for a reference scan (i.e. Visit 2) within 6 weeks of Visit 1. Finally a subset of patients will be invited to a 1 year follow up. The target sample size is 584 participants from 18 sites.

This document has been compiled according to the Edinburgh Clinical Trials Unit (ECTU) standard operating procedure (SOP) "Statistical Analysis Plans" ECTU_ST_04 and has been written based on information contained in the study protocol version 5.0 25 July 2022.

2. Statistical Methods section from the protocol

The five individual tests (and the sixth exploratory test in four sites) will be compared against the reference test (CTA/MRA), calculating standard diagnostic accuracy metrics of sensitivity, specificity, predictive values, likelihood ratio, and diagnostic odds ratio (using the bivariate model approach implemented in R).

95% confidence intervals calculated at 99% to adjust for the five comparisons will be presented.

The robustness of the findings to any observed patterns of missing data will be assessed, which are expected to differ by test. A multiple imputation approach will be used assuming the data are missing at random. In addition, and probably more consistent with the likely missing data generating mechanisms, sensitivity type analyses assuming the data are missing not at random (i.e. informatively missing) will be explored. This would attempt to identify different types of missing data by an underlying reason or reasons, and then imputing values that capture plausible measurements for those missing data. The (gamma) γ-adjustment approach given by van Buuren will be followed (Flexible Imputation of Missing Data, Chapman and Hall, 2018, section 3.8ff), and also the recommendations of Molenburghs & Kenward (Missing Data in Clinical Studies, Wiley, 2007; Section 19ff on sensitivity analyses). These approaches would allow the set of reasons for missing values to vary across the tests. The purpose is to stress the calculated findings to test their robustness to the observed patterns of missing data.

[1] In four selected centres, participants will be 1:1 randomised either to receive the visual handheld Dopper assessment first or the PAD-scan first. These participants will receive six tests in total (excluding reference tests) as oppose to five.

The subgroups of disease severity (both clinically and radiologically defined as detailed below) will be explored and those with/without neuropathy or DFU. Any further subgroup analysis (e.g. if suggested later by new data external to the study) will be labelled exploratory. Pre-specified subgroup analyses will be unlikely to be adequately powered. Clinical severity will be graded according to the severity of symptoms (from least to most severe; asymptomatic, intermittent claudication, rest pain and tissue loss). Severity will be measured radiologically using the ANGIO-score as outlined in section 5.7. Both will be analysed as pre-specified subgroup analyses in the Statistical Analysis Plan.

Combinations of tests will be explored to see if using more than one test has incremental diagnostic value. The combinations of tests that were clinically felt to potentially offer an improvement over individual tests and acknowledging the paired data, use the approach of Pepe and Thomson (Biostatistics, 2000; 1, 2; 123-140 'Combining diagnostic test results to increase accuracy'), which looks at linear combinations of the underlying tests. Post-hoc checks will be made if there were combinations that were not pre-specified that performed even better, as hypotheses for subsequent evaluation.

It is important to quantify the ability of each of the 5 index tests to measure consistently the same measurement of interest on the same leg of the same subject using the same test kit in the same location and the same environmental conditions, within a short period of time. This quantification of the intra-rater repeatability (or reproducibility) will be undertaken using the test-retest approach (35,36). The inter-rater reliability (the agreement between two or more clinicians measuring the same subject, again as under the conditions above) using appropriate methodology (35,36) will be quantified. For the inter- and intra-rater repeatability, we will aim for a sample size of 100 per pair of index tests.

These reliability studies will be performed at the start of the study and analysed as soon as the data are mature. If an index test has unacceptable intra-rater repeatability, or unacceptable inter-rater reliability, it could be dropped from further consideration, following discussions with the independent TSC. Unacceptable intra- and inter-rater reliability will be assessed in two ways— first, in an absolute sense, by looking at the kappa statistics and using the published guidance as to what an acceptable magnitude is (Fleiss, J.L. (1981). Statistical methods for rates and proportions (2nd ed.). New York: John Wiley) with a kappa of <0.4 considered unacceptable. This is not unanimity over interpreting the magnitude of kappa statistics, so our second approach will compare the kappa statistics across the tests, and label unacceptable any tests that are substantially worse than the other tests.

Inter- and intra-rater reliability will also be assessed for the reporting of reference tests using the methods outlined above. Reference tests will not be repeated due to feasibility and ethical considerations.

Index test diagnostic thresholds

The performance of the index tests based on prespecified diagnostic thresholds for PAD will be evaluated. These thresholds have been selected as they demonstrated optimal diagnostic performance in the TrEAD study or are commonly used in clinical practice. However, other thresholds have been described in the literature and there is no consensus as to which are best. Therefore, different 'exploratory' thresholds will be evaluated as part of our secondary analyses. Tests generating continuous results (ABPI, TBPI and exercise ABPI) will be evaluated for performance based on optimised thresholds derived from Receiver Operating Characteristics (ROC) analysis. A 'net benefit' approach will be used (as a sensitivity type analysis over a range of plausible thresholds) following ideas for assessing the clinical utility of prognostic models summarised in Riley R et al

(Prognostic Research in Health Care; 2018; Oxford; section 7.4.3 page 168-170). From this, it should be possible to integrate cost-effectiveness parameters into assessing the best threshold.
Diagnostic thresholds:

- *o Visual waveform assessment- monophasic or biphasic waveforms with adverse features.*
- *o Audible waveform assessment- monophasic waveform*
- *o ABPI- ≤0.9 in either vessel*
- *o TBPI-<0.75 in either vessel*
- *o Exercise ABPI (31)- Post exercise ABPI ≤0.9 in either vessel.*

3. Overall Statistical Principles

3.1 SAP objectives

The objective of this SAP is to describe the statistical analyses contributing to the final report and publication(s) of the DM PAD study. All analyses detailed in the study protocol are addressed with the exception of the internal pilot of the study which will be documented separately.

3.2 Analysis population

The analysis datasets are detailed below. All analyses will be performed on the intention to treat (ITT) population unless otherwise specified.

The trial population will include all participants who have been recruited into this trial, and who did not withdraw consent for any of their data to be stored in the trial database.

3.3 General

In general terms, categorical data will be presented using counts and percentages, whilst continuous variables will be presented using the mean, median, standard deviation (SD), minimum, maximum, Q1, Q3, inter-quartile range (IQR) and number of patients with an observation (n).

The comparison group here is diagnostic technique, so where applicable results will be split by diagnostic tests.

A further reporting metric is the number of participants who were included in the analysis, which will be supported for each of the primary (section 4.4) and secondary (section 4.5) analyses by the preparation of a STARD-style flow chart showing the 'pipeline leakage' from participants who were enrolled, to participants included in the analysis.

Where there is missing data for an outcome variable, in the first instance, those individual cases will be removed from any statistical analysis relating to that outcome variable (i.e. complete case analysis), unless otherwise specified. In tabulations, numbers of missing observations will be provided, but percentages will not include them.

All analysis and data manipulation will be carried out using SAS [1] unless otherwise stated.

Qualitative or health economic analysis will not be covered by this SAP. However, data from the statistics report may be used for health economics analyses.

3.4 Determining PAD diagnosis from Index test results

For each index test, a participant is defined as having a positive result in PAD if they have recorded the following criteria ***in either vessel (i.e. Dorsalis pedis OR Posterior tibial):***

1. Audible waveform assessment: monophasic waveform.

2a. Visual waveform assessment: monophasic ***OR*** biphasic waveforms with adverse features.

2b. Podiatry ankle duplex (PAD) scan: monophasic or biphasic waveforms with adverse features.

3. ABPI: ABPI (DP)≤0.9 ***OR*** ABPI (PT)≤0.9 in the limb being assessed.

4. TBPI: TBPI<0.75 in the limb being assessed.

5. Exercise ABPI: Exercise ABPI (DP) ≤0.9 ***OR*** Exercise ABPI (PT) ≤0.9 in the limb being assessed.

For tests that have measurements for the right and left limbs, we are only interested in the limb that was assessed, i.e. in response to the question in section 'Index Tests (Repeat 1) 'Which limb is being assessed?'.

Each criteria assessment above, except TBPI, has two vessels, a 'Dorsal pedis artery (DP)' and 'Posterior tibial artery (PD)'. If both vessels are completed then a PAD diagnosis is determined as follows:

- DP(+ve) + PT(+ve) = PAD(**+ve**)
- DP(+ve) + PT(-ve) = PAD(**+ve**)
- DP(-ve) + PT(+ve) = PAD(**+ve**)
- DP(-ve) + PT(-ve) = PAD(**-ve**)

There may be a case where either the DP or PT vessel has been recorded, not both. When this is the case a comment as to why this vessel is missing should be recorded:

- DP(+ve) + PT(missing) = PAD(**+ve**)
- DP(missing) + PT(+ve) = PAD(**+ve**)
- DP(-ve) + PT(missing) = PAD(**inconclusive)**
- DP(missing) + PT(-ve) = PAD(**inconclusive)**

When a diagnosis of PAD is inconclusive, these cases will be recorded and reported but shall not be used in the final analysis.

3.5 Determining PAD diagnosis from Reference test results

In the CRF *'DM-PAD_CRFs_CTA MRA Core Lab Reporting_Version 1.0_02082021',* there is a section on Right and Left Leg ANGIO scores. These scores are to be used to determine whether a participant has a diagnosis of PAD or not. One score should be selected for each artery.

The database captures if the participant has overall PAD however, for the analysis we need to determine if a participant has PAD in the left or right leg, which ever limb was being assessed (defined from the 'Repeat Index Tests' CRF: 'Which limb is being assessed?').

For each participant, for each leg, the total of all the scores shall be calculated which determines the following diagnoses:

- No PAD: Total score of 0
- PAD: Total score ≥ 1

Only the result on the limb being assessed is to be used in the analysis.

4. Final Analysis

This analysis plan describes the end of trial statistical analyses to be performed on DM PAD by the ECTU statistics team.

There are no statements in the report declaring what the optimal cut-point is from the ROC analysis. The optimal cut-point is that with the highest sensitivity and highest specificity, if the "cost" of incorrectly identifying a participant without PAD is the same as the "cost" of incorrectly identifying a participant with PAD. However the cut-point with the highest sensitivity cut-point is often not the same as the cut-point with the highest specificity so the final decision on the optimal cut-point needs to be a clinical decision.

4.1 Recruitment and retention

The date of first and last participants randomised, the number of participants randomised (overall and by site) will be reported.

Using the screening population, a tabulation/graphic will be produced that contains information with which to construct a STARD-style flow chart], to show the flow of participants through the different phases of the trial, as follows:

1. Number of participants assessed for eligibility
2. Number excluded, including
    a. Number not meeting inclusion criteria
    b. Declined to participate
    c. Other reasons
3. Number of eligible participants who consented – reasons for non-inclusion prior to randomisation will be categorised.
4. Number of participants recruited
5. Number of participants who withdrew consent – reasons will be categorised.
6. Number of participants to complete 1 year follow-up

The number and percentage of deaths and withdrawals will be reported.

A summary of number of participants recruited per month per site shall also be presented.

4.2 Baseline data

No formal statistical testing will be performed. Using the ITT population, the following baseline assessment results shall be summarised overall:

- age in years,
- gender (male, female, other),
- ethnicity,
- physical or mental health condition (Y/N),
- working status (fulltime, part time, unemployed, retired or student),
- weight (kg),
- height (cm),
- BMI (kg/m^2),
- smoking status (Y/N former),
- alcohol consumption (none/neglible, average units per week),
- physical activity level (low, moderate or vigorous),
- type of diabetes (I or II),
- concomitant diseases (malignancy, hypertension, stroke, heart attack, high cholesterol, angina, retinopathy, chronic kidney disease),

- foot history (previous diabetic foot ulcer and/ or minor/ major amputation),
- PAD symptoms,
- medications list (currently taking antiplatelets, anticoagulants, inflammatories, lipid modification, antihypertensives, other),
- foot examination (infection, ulceration),
- pedal pulse examination (for right and left leg – dorsal pedis, posterior tibial, neuropathy),
- wifi classification (I, II or III),
- index test results (Audible handheld Doppler, Visual handheld Doppler, PAD-scan (only in 4 centres), Ankle-brachial pressure index (ABPI), Toe-brachial pressure index (TBPI) & Exercise ABPI i.e. ABPI performed following repetitive heal raising).

4.3 Graphical analysis

We will produce boxplots, i.e. a graphical summary of the distribution including mean, median, first and third quartile, minimum and maximum values, to compare each test score according to the reference standard. Descriptive statistics will also be provided. This will be summarised overall and by site where feasible.

4.4 Primary outcome

The primary outcome will be the sensitivity of each diagnostic test compared against the reference test at diagnosing PAD. Sensitivity refers to the ability in correctly identifying participants who have PAD.

Tests generating continuous outcomes (ABPI, TBPI and exercise ABPI) shall also be evaluated for performance based on optimised thresholds derived from Receiver Operating Characteristics (ROC) analysis. This will involve calculating, for a range of cut-points, the sensitivity of each continuous outcome diagnostic test with respect to the reference test for assessing PAD. This will determine how accurate each test is in identifying patients who will develop PAD. It is desirable to have both high sensitivity and high specificity (see Secondary outcomes).

Table 1

		CMT/MRA Reference Test		
		PAD	**Non PAD**	**Total**
Diagnostic test	**Positive**	A (True Positives; TP)	B (False Positives; FP)	$T_{Test\ Positive}$
	Negative	C (False Negatives; FN)	D (True Negatives; TN)	$T_{Test\ Negative}$
		$T_{Disease}$	$T_{Non\ Disease}$	Total

4.5 Secondary analyses

4.5.1 ROC analysis

The following analyses will be obtained through the same ROC analysis performed in the primary outcome analysis for each diagnostic test:

a) The specificity of each diagnostic test compared against the reference test. Specificity refers to the ability in correctly identifying participants who do not have PAD. Tests generating continuous outcomes (ABPI, TBPI and exercise ABPI) shall also be evaluated for performance based on optimised thresholds derived from Receiver Operating Characteristics (ROC) analysis. This will involve calculating, for a range of cut-points, the specificity of each continuous outcome diagnostic test with respect to the reference test for assessing PAD.

b) The positive likelihood ratio (i.e. the ratio between the probability of a positive test result give the presence of PAD and the probability of a positive result given the absence of PAD; sensitivity/(1-specificity)) and the negative likelihood ratio (i.e. the ratio between the probability of a negative test result given the presence of PAD and the probability of a negative result given the absence of PAD, (1-sensitivity)/specificity) shall be calculated.

c) The positive predicted value (i.e. the probability of the presence of PAD when the test is positive; a/(a+b)) and the negative predicted value (i.e. the probability of the absence of PAD when the test is negative; d/(c+d)) shall be presented.

d) The diagnostic odds ratio is along with the associated 95% CI (calculated at 99% to adjust for the five or six comparisons) will be presented.

e) Inter and intra-rater reliability shall also be assessed for each diagnostic test and for the reporting of the reference tests (as described in the Section 2)

4.5.2 Combinations of tests

The protocol says that an approach described my Pepe and Thompson [2] was to be used which examines linear combinations of the underlying tests was to be used. However, the paper specifies that this method is extremely complex when we wish to consider including more than two tests in the model. So that final models are comparable, a logistic regression approach shall be used instead.

Different combinations of tests will be explored to see if using more than one index test proves to have incremental diagnostic value. A logistic regression analysis shall be performed on each combination below and presented in a table including the number and percentage of observations in each test against PAD diagnosis, odds ratios, 95% confidence intervals, and the overall C statistic for that model.

There are three different technologies when looking at the Index Tests:

- Doppler – Audible; Visual
- Brachial Pressure Index (BPI) – Toe (TBPI); Ankle (ABPI)
- Exercise

The following combinations shall be explored:

1. Audible Doppler; TBPI
2. Audible Doppler; ABPI
3. Audible Doppler; Exercise
4. Visual Doppler; TBPI
5. Visual Doppler; ABPI
6. Visual Doppler; Exercise
7. TBPI; ABPI
8. TBPI; Exercise
9. Audible Doppler; TBPI; ABPI
10. Audible Doppler; TBPI; Exercise
11. Visual Doppler; TBPI; ABPI
12. Visual Doppler; TBPI; Exercise

The results of these analyses might benefit from being presented in a forest plot to clearly show the differences between the different combinations. This may be performed after initial analysis.

Note that this analysis is exploratory and as such there is no need to consider multiple comparisons.

4.6 Sensitivity analysis

Missing data

A sensitivity analysis may be performed in which missing primary outcome data are imputed. The robustness of the findings to any observed patterns of missing data will be assessed, which are expected to differ by test. The imputation method will be determined at time of analysis, taking into consideration the choice of assumptions such as missing at random (MAR) or missing not at random (MNAR) depending on reasons for loss to follow-up[3].

4.7 Subgroup analyses

Descriptive statistics of the following subgroups shall be presented summarised by PAD vs. without PAD and overall:

- Disease severity
    - graded clinically according to the severity of symptoms (from least to worst severe; asymptomatic, intermittent claudication, rest pain and tissue loss).
    - Measured radiologically using the ANGIO-score (categorised as mild (≤4); moderate (5-9); severe (≥10)).
- Presence vs. absence of neuropathy at presentation.
- Presence vs. absence of diabetic foot ulcer (DFU) at presentation.
- Age groups (i.e. ≤59 years; 60 – 69 years; 70 – 79 years; ≥80 years).

Note that pre-specified subgroup analyses will be unlikely to be adequately powered and so no formal statistical analysis will be performed here and table shall be used for descriptive purposes.

5. Validation and QC

Data collection and quality checks will be carried in accordance to SOP ECTU_DM_01. A second statistician will perform separate programming and checking of primary outcome results and conclusions. The statistical report will also be read and sense-checked.

6. Data sharing

A file, or set of files, containing the final data will be prepared, along with a data dictionary. These will be made available to the Chief Investigator at the end of the analysis phase.

7. References

[1] SAS® Institute Inc. SAS for Windows. SAS Institute Inc.: Cary, NC, U.S.A

[2] *Pepe and Thomson (Biostatistics, 2000; 1, 2; 123-140 'Combining diagnostic test results to increase accuracy')*

[3] Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT et al. The prevention and treatment of missing data in clinical trials. N Engl J Med; 367:1355-60.