Test-retest reliability, validation, and ‘minimal detectable change’ scores for frequently reported tests of objective physical function in patients with non-dialysis chronic kidney disease

Thomas J. Wilkinson, PhD 1; Soteris Xenophontos, MSc1; Douglas W. Gould, MSc1; Barbara P. Vogt, MSc 2; João L. Viana, PhD 3, 4; Alice C. Smith, PhD 1, 5; Emma L. Watson, PhD 1

1 Leicester Kidney Exercise Team, Department of Infection, Immunity and Inflammation, University of Leicester, United Kingdom; 2 Department of Clinical Medicine, Faculdade de Medicina de Botucatu, Univ Estadual Paulista, UNESP, Brazil; 3 School of Sport, Exercise and Health Sciences, Loughborough University, United Kingdom; 4 Research Center in Sports Sciences, Health Sciences and Human Development, CIDESD, University Institute of Maia, ISMAI, Portugal; 5 John Walls Renal Unit, Leicester General Hospital, University Hospitals of Leicester Trust, United Kingdom.

Disclosure of interest(s)

The authors report no conflicts of interest. The views expressed are those of the authors and not necessarily those of the National Health Service (NHS), the National Institute for Health Research (NIHR), or the Department of Health. At the time of writing this manuscript, Dr. EL Watson was supported by a Kidney Research UK Post-Doctoral Fellowship. Ms. BP Vogt was sponsored by the CAPES Foundation within the Ministry of Education, Brazil.
This analysis forms part of a larger body of work being completed by our group (ISRCTN registration number: 36489137).

**Corresponding author**

Dr. Thomas J Wilkinson, Leicester Kidney Exercise Team, Department of Infection, Immunity and Inflammation, University of Leicester, United Kingdom, LE5 4PW. Email: drthomaswilkinson@gmail.com.
ABSTRACT

Physical function is an important outcome in chronic kidney disease (CKD). We aimed to establish the reliability, validity, and the ‘minimal detectable change’ (MDC) of several common tests used in renal rehabilitation and research. In a repeated measures design, 41 patients with CKD not requiring dialysis (stage 3b to 5) were assessed at an interval of 6 weeks. The tests were: the incremental shuttle walk test (ISWT), ‘sit-to-stand’ (STS) test, estimated 1 repetition maximum for quadriceps strength (e1RM), and VO2peak by cardiopulmonary exercise testing (CPET). Reliability was assessed using intraclass correlation co-efficient (ICC) and Bland-Altman analysis, and absolute reliability by standard error of measurement and MDC. The ISWT, STS-60, e1RM, and CPET had ‘good’ to ‘excellent’ reliability (.973, .927, .927, and .866). STS-5 reliability was poor (.676). The MDC are: ISWT, 20 m; STS-5, 7.5 secs; STS-60, 4 reps; e1RM, 6.4 kg; VO2peak, 2.8 ml/kg/min. There was strong correlation between the ISWT and VO2peak (r = 0.73 and 0.74). Whilst there was poor correlation between the STS-5 and e1RM (r = 0.14 and 0.47), better correlation was seen between STS-5 and ISWT (r = 0.55 and 0.74). In conclusion, the ISWT, STS-60, e1RM, and CPET are reliable tests of function in CKD. The ISWT is a valid means of exercise capacity. The MDC can help researchers and rehabilitation professionals interpret changes following an intervention.

Keywords
Chronic Kidney Diseases; Muscle strength; Outcome Assessment; Rehabilitation; Renal Insufficiency; Walking
INTRODUCTION

Chronic kidney disease (CKD) is associated with adverse clinical outcome and reduced quality of life (Levey et al., 2005). CKD patients have reduced physical functioning (Hiraki et al., 2013; Kuo et al., 2015; Segura-Ortí, Gordon, Doyle, and Johansen, 2017), which is partly attributable to fatigue and muscle wasting characteristic of the condition (Wang and Mitch, 2014). Reductions in physical function begin early in the disease process (Hiraki et al., 2013), and are independently associated with mortality (Roshanravan et al., 2013). Consequently, physical function is an important target for research and rehabilitative intervention.

Physical function can be assessed using a range of field tests. The incremental shuttle walk test (ISWT) is a popular measure of exercise capacity, and has been used extensively in patient populations such as chronic obstructive pulmonary disease (COPD) (Singh, Jones, Evans, and Morgan, 2008), haemodialysis (HD) (Wilund et al., 2010), and non-dialysis CKD (Greenwood et al., 2012; Watson et al., 2015). Another well-established measure is the multivariate ‘sit-to-stand’ (STS) test. The STS-5 repetition test is an assessment of lower body muscle strength, dynamic balance (Mong, Teo, and Ng, 2010), and exercise capacity (Jones et al., 2013), whilst the STS-60 second test measures lower body muscle endurance (Jones et al., 2013; Mong et al., 2010; Puthoff and Saskowski, 2013; Rikli and Jones, 2013; Segura-Ortí and Martínez-Olmos, 2011). These tests reflect a common activity of daily living (i.e. getting up from a chair), and are widely used in clinical and CKD research (Greenwood et al., 2012; McIntyre et al., 2006; Segura-Ortí et al., 2017; Segura-Ortí and Martínez-Olmos, 2011).
Reliability indicates the degree to which scores of a test are free of measurement errors. Recognising the error inherent in outcome measures is imperative to the understanding of changes and interpretation of research or rehabilitative interventions. Reliability can be expressed in both relative and absolute terms. Relative (or test-retest) reliability can be measured using intraclass correlation coefficient (ICC). Absolute reliability refers to individual performance variation and measurement error, and is quantified as standard error of measurement (SEM) (Ries, Echternach, Nof, and Blodgett, 2009; Stratford, 2004).

Using the SEM, a more clinically useful means of interpreting reliability is the ‘minimal detectable change’ (MDC) (Haley and Fragala-Pinkham, 2006), defined as the smallest amount of reliable change in a measurement necessary to conclude that the difference is not attributable to error (Segura-Ortí and Martínez-Olmos, 2011). Change exceeding the MDC is considered ‘true’ change (Haley and Fragala-Pinkham, 2006). The MDC can be calculated at an individual (MDC\text{\text{indv}}) and group (MDC\text{\text{group}}) level. The MDC\text{\text{indv}} shows whether observed changes in the individual’s status are greater than variations of chance (Lee et al., 2013), whereas the MDC\text{\text{group}} is required to determine the relevance of changes across samples (Busija et al., 2008; De Vet, Bouter, Bezemer, and Beurskens, 2001).

Alongside reliability, validity is also an important construct of physical performance tests. Whilst the ‘gold standard’ measure of exercise capacity is cardiopulmonary exercise testing (CPET) of VO\text{\text{peak}}, it is often impractical in rehabilitative settings, particularly in vulnerable clinical patients. As such, the ISWT is frequently used as its surrogate measure (Holland et al., 2014; Singh et al., 1994). Although the ISWT has been validated against CPET (via cycle ergometer and treadmill modality) in other clinical populations (Arnardóttir et al., 2006;
Green et al., 2001; Holland et al., 2014; MacSween, Johnson, Armstrong, and Bonn, 2001; Moloney et al., 2003), it has not yet been validated in CKD.

Lower body strength is important in CKD as the muscle of the legs are typically atrophied (Wang and Mitch, 2014). Whilst the STS-5 may act as a surrogate measure of strength (McCarthy, Horvat, Holtsberg, and Wisenbaker, 2004; Mong et al., 2010; Rikli and Jones, 2013), calculating 1 rep maximum (1RM) strength using resistance machines may be more applicable in non-laboratory settings (i.e. a gymnasium) where dynamometry is not available (Gail and Künzell, 2014).

No previous estimates of MDC for physical function tests exist for non-dialysis CKD patients, and neither the ISWT nor STS-5 have not been validated in this group. Despite only ~5% of CKD patients progressing to end-stage renal disease (and requiring renal replacement therapy e.g., HD) (Dalrymple et al., 2011), research, particularly in regard to rehabilitation, into non-dialysis patients often falls behind that of HD (Heiwe and Jacobson, 2014). With reductions in physical function, an independent measure of mortality (Roshanravan et al., 2013), evident in the early stages of CKD (Hiraki et al., 2013), it appears fundamental that interventions designed to improve functional status (e.g., exercise (Heiwe and Jacobson, 2014)) are initiated promptly in the disease process and before it can progress. Furthermore, with non-dialysis patients often experiencing differing functional capacities to other disease populations (e.g., COPD), and indeed other CKD groups (e.g., those on HD (Hiraki et al., 2013; Segura-Ortí et al., 2017)), the identification of the reliability and validity of physical function tests specific to the non-dialysis population is vital in the correct interpretation of functional changes.
The aims of the current study were to: 1) determine the test-retest reliability and estimated MDC of the common physical function tests in patients with non-dialysis CKD; and 2) confirm the validity of the ISWT as a measure of exercise capacity ($VO_{2peak}$), and the STS as a measure of lower body strength.
METHODS

All assessments took place between December 2013 and June 2016 at Leicester General Hospital, Leicester, UK. Patients were recruited from nephrology outpatient clinics at the University Hospitals of Leicester NHS Trust. Patients gave written informed consent in accordance with the Declaration of Helsinki and local Research Ethics Committee approval was obtained. This analysis forms part of a larger body of work completed by our group (ISRCTN registration 36489137). To ensure accurate reporting of study measurement properties and analysis parameters, the ‘COnsensus-based Standards for the selection of health Measurement Instruments’ (COSMIN) checklist was adhered to (Mokkink et al., 2010).

Participants

Patients were recruited if they had: CKD stages 3b-5 (i.e. an estimated glomerular filtration rate (eGFR) of $\leq 44 \text{ ml/min}/1.73^2$) not requiring renal replacement therapy (e.g., HD); were aged $\geq 18$; no significant co-morbidity (e.g., unstable hypertension, potentially lethal arrhythmia, myocardial infarction within the previous six months) contraindicative to physical exercise; no significant physical impairment; and sufficient ability to give informed consent. Prior exercise and physical activity level was not a pre-requisite for inclusion.

The severity of comorbidity was recorded and scored according to Charlson Comorbidity Index (CCI). A higher CCI score indicated greater comorbidity, and CCI scores of 1–2 were classified as mild; scores of 3–4 as moderate; and scores $\geq 5$ as severe (Huang et al., 2014).

Physical function assessments
Assessments were performed at two time points (test 1 and 2) separated by a six week period in which patients were instructed to maintain their habitual lifestyle. The six week interval forms part of a control period in the main trial, thus we used pre- and post-data as an opportunity to assess the reliability of the tests employed. All researchers performing the tests followed strict operating procedures for each of the tests performed to reduce investigator bias. As different researchers were used, we do not present intra-rater reliability, but inter-rater reliability. Each test was performed in the same order (STS tests, ISWT, and e1RM). As CPET was used to screen patients for cardio-pulmonary contraindications, and due to the relative exhaustive aspect of the test, CPET was performed on a separate visit several days before. All tests, apart from CPET due to logistic and cost factors, had a familiarisation test several days prior to their first test to minimise learning effect (Holland et al., 2014). The familiarisation testing session involved an explanation of the test to the patient, short demonstration of the basic movement(s) by the researcher present, a practice repetition (if appropriate), and then the full test. On the day of testing, following a re-explanation of the test, patients performed each test once.

ISWT

The ISWT is maximally progressive test that involves walking at a pace externally dictated by an auditory tone (Holland et al., 2014; Singh et al., 1994). During the test, the patient walked a total of 10m back and forth and around two cones. The walking pace was increased by a rate of 0.17m/sec every minute for twelve stages until the patient could no longer keep up with the pace because of breathlessness, pain, or other symptoms. Only completed shuttles were counted. The outcome was distance walked (m).
The ISWT was preferred over other measures of exercise capacity, such as the 6-minute walk test (6MWT), as this test requires at least 30m of walking space (Holland et al., 2014); provisions not available at our research facility. However, along with reducing limitations associated with being self-paced, the ISWT may also provide superiority when prescribing exercise intensity as a % of peak performance (Holland et al., 2014), and thus may be more beneficial in a renal rehabilitation setting when tailoring individualised exercise interventions.

**STS tests**

The STS-5 and -60 tests were employed as measures of lower body strength and muscle endurance (Mong et al., 2010; Rikli and Jones, 2013; Segura-Ortí and Martínez-Olmos, 2011). The patient sat on a seat (43.2 cm from the ground). With their hands across their chest, patients were asked to: 1) perform five complete STS cycles as fast as possible (STS-5); and 2) perform as many complete STS cycles in 60 secs (STS-60). The STS-5 time was stopped when the patient was seated following their fifth repetition. If the patient was halfway through a stand when STS-60 time had expired, this was counted as one repetition (Rikli and Jones, 2013). The STS-5 test was preferred over other STS versions (e.g., STS-10) as its short duration reduces patient burden during testing sessions involving multiple other outcomes (Nilsagård, Andreasson, Carling, and Vesterlin, 2017).

**Lower limb strength**

The maximal strength (kg) of the quadriceps muscle was measured using a leg extension machine (TechnoGym, Italy). Performing a true 1RM test is associated with an increased injury risk and stress on the muscles and joints, particularly in untrained (Gail and Künzell, 2014) and clinical groups (Abdul-Hameed, Rangra, Shareef, and Hussain, 2012), therefore
we estimated 1RM (e1RM) for the leg extension exercise from a 5-rep maximum (5RM) (Brzycki, 1993; Dudgeon et al., 2010). During the test, weight was progressively increased by a minimum of 2.5 kg depending on participant feedback and ease of the previous 5 repetitions. To reduce cumulative fatigue and to ensure an accurate 5RM, patients were given a minimum 60 second rest in-between each attempt, although researcher discretion was also used. The 5RM was determined as the maximal weight the patient could lift five times with correct technique.

**CPET**

Peak exercise capacity (VO2peak) was assessed using CPET. Patients were asked to cycle for as long as possible at a revolutions per minute (RPM) ≥60. Following a 3 minute warm up, the resistance on the static ergometer (Lode Excalibur, Netherlands) increased from 30 Watts by 1 Watt every 3 secs in a ramp protocol. Throughout the test, an echocardiogram was performed and reviewed by an experienced exercise cardiac nurse or doctor. The test was stopped if: RPM <60; the patient reached volitional exhaustion; or at the discretion of the medical professional. Using online direct breath-by-breath measurement (Cortex Metalyzer, Cranlea, UK) of oxygen consumption (VO2), we calculated relative VO2peak (peak ml/kg/min).

**Statistical methods**

Test-retest relative reliability of data was assessed using the ICC (r). An ICC between .600-.749 is considered ‘fair’, ≥.750 ‘good’, whilst a value ≥.900 is considered ‘excellent’ for clinical measures (Cicchetti, 1994). Data is also represented graphically as Bland-Altman plots with mean bias and limits of agreement set at 95% confidence intervals (95CI) (Bland
and Altman, 1999). Here the difference of the two paired measurements is plotted against the mean of the two measurements.

The $SE_M$ and MDC were calculated as a measure of absolute reliability (Haley and Fragala-Pinkham, 2006; Stratford, 2004). The $SE_M$ was calculated as:

$$SE_M = SD \times \sqrt{1-r}.$$

This method of calculating $SE_M$ has been used previously (Chiu et al., 2016; Ries et al., 2009; Segura-Ortí and Martínez-Olmos, 2011). The MDC at an individual level ($MDC_{indv}$) was calculated at the 95CI. The equation used was:

$$MDC_{indv} = SE_M \times 1.96 \times \sqrt{2}.$$

The 1.96 represents the $z$-score at the 95CI. The ‘$SE_M \times 1.96$’ is multiplied by the $\sqrt{2}$ to account for errors associated with repeated measures (De Vet et al., 2001; Haley and Fragala-Pinkham, 2006; Segura-Ortí and Martínez-Olmos, 2011; Stratford, 2004). The MDC at group level (Busija et al., 2008; De Vet et al., 2001) ($MDC_{group}$) was calculated as:

$$MDC_{group} = MDC_{indv} / \sqrt{n}.$$

Construct validity between the ISWT and CPET derived VO$_{2peak}$, and between the STS-5 and e1RM was assessed using simple linear regression and data are represented as scatterplot graphs with a trend line showing $r$ and 95CI interval bands.
A minimum sample size of 39 patients was needed to estimate an ICC \( r \) of .600 (the minimal acceptable ICC in clinical investigations (Shoukri, Asyali, and Donner, 2004)) with a \( \beta \) of 0.80 at a significance level of \( P = 0.050 \) (Walter, Eliasziw, and Donner, 1998). Data was assessed using SPSS v24. Data are reported as mean (SD), if normally distributed, or as median (interquartile range, IQR). Distribution was assessed using the Kolmogorov-Smirnov test. Paired comparisons were tested using paired \( t \)-tests, or non-parametric Wilcoxon signed rank test as appropriate. Difference (with 95CI) is reported as mean (if data were normally distributed), or median (if data were non-normally distributed). Data for some individuals were not collected for different assessments due to missed measures or an inability to complete (see footnote of Table 2); consequently, with no comparable data, this patient was excluded (listwise approach) from analysis in that test.
RESULTS

Forty-one patients (23 females, 56%) with non-dialysis CKD were recruited. Mean patient age was 62 (SD: 11) years old with a body mass index of 30.1 (SD: 5.7). The majority of patients were Caucasian (66%). The mean eGFR was 25 (SD: 8) ml/min/1.73². Over half of patients (63%) had previously diagnosed hypertension and 22% of patients had diabetes mellitus type II. The mean CCI score was 2.5 (SD: 0.7), with the majority of patients (59%) classified as having a mild CCI score. Full patient clinical and demographic characteristics are shown in Table 1. Apart from anticipated exercise-induced fatigue during the tests, no adverse events of complaints were recorded.

Test-retest relative and absolute reliability

There was minimal difference between the two tests for the ISWT, STS-60, e1RM, and VO₂peak by CPET with ICC $r$ values of .973, .927, .927, (all rated as ‘excellent’) and .866 (rated ‘good’), respectively. Conversely, the ICC for the STS-5 displayed only ‘fair’ agreement ($r = .676$). Data analysis revealed no statistically significant differences between the two tests for each measure of function (Table 2). Figure 1A–E shows Bland-Altman plots with bias and limits of agreement (at 95CI) for each of these tests. Absolute reliability data (both SEM and MDC) are shown in Table 3.

The proportion of patients whom performed best in test 1 was as follows: ISWT (n = 16, 39%), STS-5 (n = 18, 44%), STS-60 (n = 14, 35%), e1RM (n = 9, 23%), and VO₂peak (n = 20, 54%). No change between test 1 and 2 was observed in n = 3 (7%) for the ISWT; n = 2 (5%) for the STS-60; n = 9 (23%) for the e1RM; and n = 1 (3%) for VO₂peak by CPET.
Validation of the ISWT and STS-5

We found a ‘strong’ correlation between the ISWT and VO$_{2\text{peak}}$ test with an $r$ of 0.73 at time point 1 ($P < 0.001$), and an $r$ of 0.74 at time point 2 ($P < 0.001$) (Figure 2A–B). The correlation between the STS-5 and the e1RM test was ‘poor’ at time point 1 ($r = 0.14$, $P = 0.371$), and (albeit better) ‘weak’ at time point 2 ($r = 0.47$, $P = 0.003$) (Figure 2C–D). When we compared the STS-5 with ISWT, we found ‘moderate’ correlation at time point 1 ($r = 0.55$, $P < 0.001$), and ‘strong’ correlation at time point 2 ($r = 0.74$, $P < 0.001$) (Figure 2E–F).
This study is the first investigation into the test-retest reliability and identification of the MDC for tests of physical function in a non-dialysis CKD population. Our results have large clinical relevance in CKD research and rehabilitation, and demonstrate that the ISWT, STS-60, e1RM, and VO₂peak via CPET have ‘good’ to ‘excellent’ reliability in non-dialysis CKD. Conversely, the STS-5 test performed poorly with only ‘fair’ agreement. Using SEM data, we were able to calculate the MDC at an individual and group level. We also confirmed the ISWT as a valid measure of VO₂peak, but found the STS-5 only weakly associated with lower limb strength.

The ISWT in our trial showed excellent test-retest reliability with an ICC of .973. This corresponds well with previous estimates in cardiac rehabilitation patients (.990) (Hanson, Taylor, and McBurney, 2015), patients with peripheral vascular disease (.990) (Cunha-Filho et al., 2008), and non-cystic fibrosis (.950) (Lee et al., 2015). The ISWT is used extensively in research and rehabilitation as a measure of exercise capacity, and as stated previously, was chosen above the 6MWT, another common field test of exercise capacity, for several reasons including space considerations and the limitations associated with self-paced tests. Furthermore, given its progressive nature, the ISWT is superior when prescribing exercise intensity (e.g., % of maximum) (Holland et al. 2014), and thus potentially more useful in a renal rehabilitation setting. Studies validating the ISWT against CPET-measured VO₂peak have found good agreement ($r = 0.74$ to 0.88) regardless of exercise modality (i.e. treadmill versus cycle ergometer) (Arnardóttir et al., 2006; Green et al., 2001; Holland et al., 2014; MacSween et al., 2001; Moloney et al., 2003). Our $r$ values of 0.73 and 0.74 represent a
strong agreement, and confirm the ISWT as a valid and easy means to assess exercise
capacity in a CKD population.

The STS-60 had excellent reliability with an ICC of .927. Similarly, Segura-Ortí and
Martínez-Olmos (2011) found an ICC of .970 in HD patients; however, other clinical
research into the reliability of the STS-60 is limited. Conversely, the STS-5 displayed
relatively poor reliability with an ICC of .676 (‘fair’). Whilst this contrasts with previous
estimates in clinical populations (ICC values between .940 and .990), (Jones et al., 2013;
Mong et al., 2010; Thomas and Hageman, 2002), similar reliability (ICC of .640) was
reported by Ostchega et al (Ostchega et al., 2000) in older adults. Unfortunately, the authors
in this trial offer no explanation for their moderate reliability score. The lack of reliability in
our study may be due to the large within-patient variability seen in our sample. Indeed,
several patients took considerably longer on their first attempt than on their second: one
patient performed 25.8 seconds quicker in the STS-5 at visit 2. When we removed this
patient, the ICC for the STS-5 increased to .752 (.577 - .860), and although defined as now
‘good’, this still lacked the reliability observed relative to our other tests. As the aim of this
study was to assess natural variation, and with no clinical justification to remove this patient,
we retained this data point in the analysis

As recommended by Nilsagård et al. (2017), the STS-5 was preferred over the STS-10,
another commonly reported STS variant, to: 1) ease patient burden (i.e. STS-5 takes ~50%
less time); and 2) to isolate the assessment of strength. Given the requirement to complete
double the repetitions, the STS-10 has been described as a measure of muscular endurance
(Nilsagård et al. 2017), whereas the STS-5 is purported as a measure of lower limb strength.
Therefore, to avoid multiple assessments of muscular endurance we used the STS-5.
Nonetheless, we observed a weak relationship between the STS-5 and lower limb strength (by e1RM), suggesting that, in our patients, the STS-5 may not be a comprehensive proxy measure of strength. Whilst this supports previous work which found the STS-5 had only a modest relationships with lower limb strength in COPD (Jones et al., 2013; Roig et al., 2011), it conflicts with research that suggests in both elderly and stroke patients, the STS-5 is a good marker of lower body strength (Mong et al., 2010). Interestingly, although we were unable to demonstrate any strong relationship between STS-5 and lower limb muscle strength, we did observe a strong to moderate association between the STS-5 and the ISWT. Similar observations have been reported by Jones et al (2013) in COPD patients. In CKD clinical practice and research, other methods to measure lower limb strength should be employed over the STS-5.

Limited research exists on the reliability of the e1RM estimated by a 5RM test. Our ICC value of .927 compares well with previous estimates in healthy recreational athletes (.900) (Gail and Künzell, 2014), and in untrained diabetic patients (.990) (Abdul-Hameed et al., 2012). Although dynamometry is considered the ‘gold standard’ for the assessment of strength, e1RM testing is commonly applied for ease and simplicity, and represents a valid means to assess leg muscle strength (Verdijk, Van Loon, Meijer, and Savelberg, 2009). We experienced no serious difficulties performing a 5RM test, and therefore in gymnasium-based rehabilitative setting, where dynamometers are unavailable, a 5RM test may be a safe and reliable tool in strength measurement in CKD.

As expected, the ‘gold standard’ of assessing exercise capacity (VO_{2peak}) by CPET showed good (bordering on ‘excellent’) test-retest reliability. Our ICC of .866 compares well with
previous testing in clinical groups, including patients with multiple sclerosis (.933) (Heine et al., 2015) and coronary heart disease (.970) (Coeckelberghs et al., 2016).

The MDC is the smallest change that falls outside the expected range of error (Segura-Ortí and Martínez-Olmos, 2011), and is invaluable to know as any change from a clinical intervention exceeding the MDC can be considered a ‘true’ change (Haley and Fragala-Pinkham, 2006). We were able to calculate the MDC at both an individual and group level. An MDC at individual level is should be used to determine whether a single patient has made an improvement, not attributable to error, following an intervention (Haley and Fragala-Pinkham, 2006; Lee et al., 2013). This may be useful to physical therapists, or other allied professionals, whom work with, and interpret changes of, individual patients. Conversely, as the majority of clinical trials compare changes between groups of patients, the MDC$_{indv}$ should be adapted to quantify the MDC of an outcome measure in a group of patients (De Vet et al., 2001). Often the MDC$_{group}$ is smaller than the corresponding MDC$_{indv}$ implying superiority at detecting change at group level (De Vet et al., 2001).

At an individual level, the MDC for the ISWT was 20m. The small MDC$_{group}$ (3m) found suggests that, at group level, small changes in ISWT scores can be reliably identified. To our knowledge, this is the first estimate of the MDC for this test in any clinical population.

For the STS-60, we found an MDC$_{indv}$ of 4 reps. Similar findings have been reported in HD patients (4 reps) (Segura-Ortí and Martínez-Olmos, 2011). This is perhaps due to the similar mean number of repetitions (26) between their HD patients and our non-dialysis CKD sample. Whilst such comparable functional capacity between HD and non-dialysis CKD patients may be unexpected, the majority of patients in the Segura-Ortí and Martínez-Olmos
trial were almost exclusively male (82%). Male CKD patients (including those on HD) display superior functional capacity than females (Hiraki et al., 2013); as our sample was majority female (56%), this may explain the comparable results observed. At a group level, the high reliability and low measurement error of the STS-60 means the MDC\textsubscript{group} is just 1 rep.

Whilst an MDC of 3.1 secs has been reported for the STS-5 in patients undergoing cardiac rehabilitation (Puthoff and Saskowski, 2013) and 5.0 secs in those with stroke (Pardo et al., 2013), our data revealed an MDC\textsubscript{indv} of 7.5 secs and an MDC\textsubscript{group} of 1.2 secs. Our larger MDC values could be due to the poor reliability for this test as discussed previously.

Whilst no MDC exists for 1RM testing, using dynamometry, in patients with knee osteoarthritis, the MDC using peak isokinetic knee extension torque was estimated at 3.5 kg (Kean et al., 2010). This value is smaller than our estimated MDC\textsubscript{indv} of 6.4 kg. A possible explanation for the larger value in our trial may be due to equipment limitations. The machine used only increased in increments of 2.5 kg and above, thus limiting the ability to increase weight in smaller increments. Whilst the e1RM has been shown to represent a valid measure of strength assessed by dynamometer (Verdijk et al., 2009), further research is needed to assess the validity in a CKD group.

The MDC\textsubscript{indv} for VO\textsubscript{2peak} was 2.8 ml/kg/min, and 0.5 ml/kg/min at group level. Brehm et al. (Brehm, Balemans, Becher, and Dallmeijer, 2014) previously estimated an MDC\textsubscript{indv} of 5.7 ml/kg/min in patients with cerebral palsy, whilst in patients with multiple sclerosis, Heine et al. (Heine et al., 2015) calculated the MDC\textsubscript{indv} at 4.6 ml/kg/min and the MDC\textsubscript{group} at 0.8 ml/kg/min. Like the ISWT, the small MDC\textsubscript{group} identified suggests that small changes in
VO₂peak can be identified. This is perhaps unsurprising given the highly technical analysis of CPET.

We have confirmed the ISWT as a valid assessment of exercise capacity, although disappointingly the STS-5 was not associated with lower limb strength. The low $r$ at time point 1 between the STS-5 and e1RM may be attributed to the outlier situated to the far right of the Bland-Altman plot. However, as stated above, with no clinical justification to remove this patient, we retained this data point in the analysis. Nonetheless, even at time point 2, the $r$ value (0.47) was still poor. The STS test as a proxy measure of lower limb strength (Mong et al., 2010; Rikli and Jones, 2013) is increasingly being questioned by researchers (McCarthy et al., 2004), and study has shown that only a moderate proportion of STS performance is attributable to strength, with factors such as balance and sensorimotor ability also contributing (McCarthy et al., 2004). Indeed, our data, and that of others (Jones et al., 2013), suggests that the STS-5 may actually be a better measure of exercise capacity than strength. As such, additional physiological factors may be contributing, and further research is required into the identifying these elements and assessing the validity of the STS-5.

Study limitations

Our patient population was an opportunistic sample taken from a larger body of work completed by our group. Nonetheless, the 41 patients included in the current analysis was above the minimum sample size of 39 patients required to estimate an ICC $r$ of .600 (the minimal acceptable ICC in clinical investigations). Consequently, our analysis was sufficiently powered to detect any differences between time points.
Another limitation in our trial was the six week test-retest interval. However, as our data is derived from the control period the larger trial, this period was unavoidable. Nevertheless, our trial was suitably powered and our data shows even over a six week period, the reliability of our function tests was good. Furthermore, previous test-retest estimates of the reliability of objective function such as the STS (Schaubert and Bohannon, 2005a; Schaubert and Bohannon, 2005b) and strength (Schaubert and Bohannon, 2005b) over a similar six week interval have also established good agreement.

In order to ensure accurate baseline data and to reduce learning effects (Holland et al., 2014; Johnson-Warrington, Sewell, Morgan, and Singh, 2015), patients completing the ISWT at visit 1 did so after a familiarisation test several days before. However, due to time constraints at the second visit, it was feasible that the ISWT, and other outcomes, were repeated only once. Whilst we observed good reliability between the two visits for the ISWT without re-familiarisation at visit 2, it is unclear from our data if lack of re-familiarisation has effected outcome measure scores at the second visit and whether one test is sufficient if the test has previously been conducted is still unclear (Holland et al., 2014). In intervention trials, repeating, or re-familiarising patients with physical function tests may prevent underestimation of the impact of an intervention (Holland et al., 2014). Interestingly however, due to impracticality and costing, patients did not undergo familiarisation CPET at baseline; however, our results show even without this familiarisation test, reliability for CPET is high, perhaps due to the technical accuracy of the method, and a familiarisation is perhaps unnecessary for this test.

In order to further improve the methodology of our study, further research should investigate the identification of the MDC in earlier stages of CKD (1 and 2), as well as potentially
identifying differences in gender or ethnicity. In our sample, a large proportion (46%) of our patients were obese (defined by BMI), with 34% ‘overweight’. Whilst more current obesity prevalence statistics are needed to fully determine the demographic of our sample, the external validity of the results may be reduced. Dynamometry, over 1-5RM testing, should be used to accurately measure strength; as such, associations with the STS-5 should be reassessed. Additional physiological processes that contribute to function, such as balance, should also be considered in future trials assessing performance. The minimally clinical important difference, which is the minimal meaningful important difference from a patient perspective, should also be investigated.
CONCLUSIONS

In conclusion, the ISWT, STS-60, e1RM, and CPET are reliable and valid tests of function in non-dialysis CKD patients. The MDC’s we have calculated in this trial are imperative in the interpretation of changes as a result of intervention trials or rehabilitation programmes. In order for a ‘true’ change to occur, i.e. that above individual performance variability or inherent measurement error, the value should exceed the MDC’s presented here. The MDC at an individual level is useful in evaluating changes in a single patient, whilst the MDC at group level should be employed in research comparing changes between groups. The MDC should be used to help clinicians, researchers, and rehabilitation professionals interpret reliable changes in their patients and clinical trials.

Acknowledgments

We are grateful to the Stoneygate Trust for part-funding of this work. The research was supported by the National Institute for Health Research (NIHR) Leicester Biomedical Research Centre (BRC). The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research Leicester BRC or the Department of Health. We would like to thank Ms. Amy Clarke, Mr. Darren Churchward, Mr. Patrick Highton, and Ms. Charlotte Grantham, researchers at the Leicester Kidney Exercise Team, for collection of some assessment outcome data.
REFERENCES


Hanson LC, Taylor NF, McBurney H 2015 The 10m incremental shuttle walk test is a highly reliable field exercise test for patients referred to cardiac rehabilitation: a retest reliability study. Physiotherapy. DOI: 10.1016/j.physio.2015.08.004

multiple sclerosis and low to mild disability. PloS One 10: e0122260. DOI: 10.1371/journal.pone.0122260


McIntyre CW, Selby NM, Sigrist M, Pearce LE, Mercer TH, Naish PF 2006 Patients receiving maintenance dialysis have more severe functionally significant skeletal muscle wasting than patients with dialysis-independent chronic kidney disease. Nephrology Dialysis Transplantation 21: 2210-6. DOI: 10.1093/ndt/gfl064


Table 1. Patient clinical and demographic characteristics

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>n = 41</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years) (SD) [min-max]</td>
<td>62 (11) [27-80]</td>
</tr>
<tr>
<td>Females, n (%)</td>
<td>23 (56%)</td>
</tr>
<tr>
<td>BMI (kg/m$^2$) (SD) [min-max]</td>
<td>30.1 (5.7) [16.4-41.5]</td>
</tr>
<tr>
<td>eGFR (ml/min/1.73$^2$) (SD) [min-max]</td>
<td>25 (8) [9-41]</td>
</tr>
<tr>
<td>Haemoglobin (g/l) (SD) [min-max]</td>
<td>118 (14) [90-161]</td>
</tr>
<tr>
<td>Albumin (mg/l) (SD) [min-max]</td>
<td>41 (2.7) [34-47]</td>
</tr>
</tbody>
</table>

**Ethnicity**

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caucasian</td>
<td>27 (66%)</td>
</tr>
<tr>
<td>South Asian</td>
<td>6 (15%)</td>
</tr>
<tr>
<td>Indian</td>
<td>4 (10%)</td>
</tr>
<tr>
<td>Black Caribbean</td>
<td>2 (5%)</td>
</tr>
<tr>
<td>Other</td>
<td>2 (5%)</td>
</tr>
</tbody>
</table>

**Cause of disease**

<table>
<thead>
<tr>
<th>Cause of disease</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetic nephropathy type II</td>
<td>5 (12%)</td>
</tr>
<tr>
<td>Interstitial nephritis</td>
<td>5 (12%)</td>
</tr>
<tr>
<td>IgA nephropathy</td>
<td>3 (7%)</td>
</tr>
<tr>
<td>Polycystic kidney disease</td>
<td>3 (7%)</td>
</tr>
<tr>
<td>Other</td>
<td>5 (12%)</td>
</tr>
<tr>
<td>Unknown / aetiology uncertain</td>
<td>20 (49%)</td>
</tr>
</tbody>
</table>
## Comorbidities

<table>
<thead>
<tr>
<th>Condition</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetes mellitus type II</td>
<td>9 (22%)</td>
</tr>
<tr>
<td>Hypertension</td>
<td>26 (63%)</td>
</tr>
<tr>
<td>CCI score (SD) [min-max]</td>
<td>2.5 (0.7) [2-4]</td>
</tr>
<tr>
<td>- Mild CCI score (1-2)</td>
<td>24 (59%)</td>
</tr>
<tr>
<td>- Moderate CCI score (3-4)</td>
<td>17 (41%)</td>
</tr>
<tr>
<td>- Severe CCI score (≥ 5)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Median n of morbidities [min-max]</td>
<td>3 [0-6]</td>
</tr>
</tbody>
</table>

Unless stated, data presented as mean and SD: standard deviation; BMI: body mass index; CCI: Charlson Comorbidity Index; eGFR: estimated glomerular filtration rate
Table 2. Relative reliability results of physical function tests in patients with CKD

<table>
<thead>
<tr>
<th>Test</th>
<th>n</th>
<th>Test 1 score</th>
<th>Test 2 score</th>
<th>Difference (95CI)</th>
<th>ICC (r) †</th>
<th>95CI for ICC (r)</th>
<th>P †</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISWT (m)</td>
<td>41 ‡</td>
<td>400 (SD: 186)</td>
<td>409 (SD: 185)</td>
<td>9 (-4 – 23)</td>
<td>.973</td>
<td>.950 -.986</td>
<td>0.164</td>
</tr>
<tr>
<td>STS-5 (secs)</td>
<td>41 §</td>
<td>11.0 (IQR: 8.8 – 13.4)</td>
<td>10.5 (IQR: 8.7 – 12.4)</td>
<td>-0.1 (-0.9 – 0.4)</td>
<td>.676</td>
<td>.468 -.813</td>
<td>0.248</td>
</tr>
<tr>
<td>STS-60 (repetitions)</td>
<td>40 ‡</td>
<td>26 (IQR: 20 – 29)</td>
<td>28 (IQR: 22 – 30)</td>
<td>2 (0 – 2)</td>
<td>.927</td>
<td>.866 -.961</td>
<td>0.093</td>
</tr>
<tr>
<td>e1RM (kg)</td>
<td>40 ¶</td>
<td>50.6 (IQR: 34.5 – 64.7)</td>
<td>51.8 (IQR: 36.6 – 64.7)</td>
<td>2.7 (0.0 – 4.3)</td>
<td>.927</td>
<td>.866 -.961</td>
<td>0.053</td>
</tr>
<tr>
<td>VO₂peak (ml/kg/min)</td>
<td>37 #</td>
<td>19.6 (SD: 5.4)</td>
<td>19.6 (SD: 5.7)</td>
<td>0.0 (-0.9 – 1.0)</td>
<td>.866</td>
<td>.755 -.929</td>
<td>0.936</td>
</tr>
</tbody>
</table>

95CI: 95% confidence interval; e1RM: estimated 1 rep maximum-leg extension strength; ICC: intraclass correlation coefficient; ISWT: incremental shuttle walk test; IQR: interquartile range; SD: standard deviation; STS: sit-to-stand; † refers to test 1 versus test 2 analysis; ‡ 1 (2%) patient removed due to arthritic pain in hips and knees during test, 1 (2%) patient had incomplete data missing at time point 1 or 2; § 1 (2%) patient removed due to arthritic pain in hips and knees during test, 1 (2%) patient had incomplete data missing at time point 1 or 2; ¶ 3 (7%) patients removed due to arthritic pain in hips and knees during test, 1 (2%) patient had incomplete data missing at time point 1 or 2; † 3 (7%) patients had incomplete data missing at time point 1 or 2; # 6 (14%) patients had incomplete data missing at time point 1 or 2
Table 3. SEM and MDC at both individual and group levels at 95CI for physical function tests in patients with CKD

<table>
<thead>
<tr>
<th>Test</th>
<th>$SE_M$</th>
<th>$MDC_{indv}$</th>
<th>$MDC_{group}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISWT (m)</td>
<td>7.1</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>STS-5 (secs)</td>
<td>2.7</td>
<td>7.5</td>
<td>1.2</td>
</tr>
<tr>
<td>STS-60 (reps)</td>
<td>1.3</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>e1RM (kg)</td>
<td>2.3</td>
<td>6.4</td>
<td>1.0</td>
</tr>
<tr>
<td>$VO_2^{peak}$ (ml/kg/min)</td>
<td>1.0</td>
<td>2.8</td>
<td>0.5</td>
</tr>
</tbody>
</table>

e1RM: estimated 1 rep maximum-leg extension strength; ISWT: incremental shuttle walk test; $MDC_{indv}$: minimal detectable change at individual level; $MDC_{group}$: minimal detectable change at group level; SEM: standard error of measurement; STS: sit-to-stand.
Figure 1. Bland-Altman plots showing test-retest reliability between test 1 and test 2

A

Figure 1. Bland-Altman plots showing test-retest reliability between test 1 and test 2

B

C

D

E

e1RM = estimated 1 repetition maximum-leg extension strength; ISWT = incremental shuttle walk test; STS = sit-to-stand. Top and bottom dotted lines represent upper and lower 95% confidence limits of agreement, respectively. Dashed line represents mean bias.
Figure 2. Linear regression line showing the ISWT against VO$_{2peak}$, and the STS-5 against e1RM and ISWT over the two testing visits

e1RM = estimated 1 repetition maximum-leg extension strength; ISWT = incremental shuttle walk test; STS-5 = sit-to-stand 5 repetition test. 95% confidence interval bands are shown as dotted lines that embrace the regression line in solid black fill.