



# Criterion Validity of Screening Tools and Field-Based Tests for Health-Related Physical Fitness in Inflammatory Bowel Disease

Karlijn Demers<sup>1,2,3,4</sup> · Bart C. Bongers<sup>3,5</sup> · Sander M. J. van Kuijk<sup>6</sup> · Guy Plasqui<sup>5</sup> · Daisy M. A. E. Jonkers<sup>4</sup> · Marieke J. Pierik<sup>2,4</sup> · Laurents P. S. Stassen<sup>1,3</sup>

Received: 24 July 2024 / Accepted: 4 October 2024 / Published online: 19 October 2024  
© The Author(s) 2024

## Abstract

**Background** Monitoring health-related physical fitness (HRPF) may benefit proactive Inflammatory Bowel Disease (IBD) management. However, knowledge regarding HRPF in patients with IBD is limited and gold standard tests are impractical for widespread use, necessitating simpler methods.

**Aim** This study evaluated the criterion validity of screening tools and field-based tests compared to gold standard tests for HRPF in patients with IBD.

**Methods** Adult patients with IBD completed screening tools, field-based tests, and gold standard tests for HRPF. Criterion validity was examined through (intraclass) correlation coefficients and Bland–Altman plots. Predictive capacity of the screening tools was examined with receiver operating curve analysis.

**Results** Among 53 included patients, screening tools demonstrated poor-to-moderate validity compared to the cardiopulmonary exercise test (CPET) for cardiorespiratory fitness. Very strong correlations were found for four-site skinfold thickness and multi-frequency bioimpedance analysis (BIA) with deuterium oxide dilution for body fat percentage (ICC = 0.90, ICC = 0.93), and between the steep ramp test and CPET ( $r = 0.95$ ) for cardiorespiratory fitness. The steep ramp test also correlated strongly with isokinetic quadriceps ( $r > 0.75$ ) and hamstring ( $r > 0.74$ ) strength. Hand-held dynamometry and the sit-to-stand test showed strong correlations with hamstring strength ( $r > 0.80$ ,  $r > 0.76$ ). Negligible correlations were found for field-based tests compared to isokinetic quadriceps and hamstring endurance.

**Conclusions** Four-site skinfold thickness and BIA showed good agreement with the gold standard for body fat measurement. The steep ramp test demonstrated strong correlations with the gold standard tests for cardiorespiratory fitness and quadriceps and hamstring strength, while hand-held dynamometry and the sit-to-stand test showed strong correlations with hamstring strength.

**Keywords** Body composition · Cardiorespiratory fitness · Muscular strength · Muscular endurance · Validation · IBD

✉ Karlijn Demers  
karlijn.demers@mumc.nl; k.demers@maastrichtuniversity.nl

<sup>1</sup> Department of Surgery, Maastricht University Medical Center+, P. Debyelaan 25, 6229 HX Maastricht, The Netherlands

<sup>2</sup> Department of Gastroenterology-Hepatology, Maastricht University Medical Center +, P. Debyelaan 25, 6229 HX Maastricht, The Netherlands

<sup>3</sup> Department of Surgery, Institute of Nutrition and Translational Research in Metabolism (NUTRIM), Maastricht University, PO Box 616, 6200 MD Maastricht, The Netherlands

<sup>4</sup> Department of Gastroenterology-Hepatology, Institute of Nutrition and Translational Research in Metabolism (NUTRIM), Maastricht University, PO Box 616, 6200 MD Maastricht, The Netherlands

<sup>5</sup> Department of Nutrition and Movement Sciences, Institute of Nutrition and Translational Research in Metabolism (NUTRIM), Maastricht University, PO Box 616, 6200 MD Maastricht, The Netherlands

<sup>6</sup> Department of Clinical Epidemiology and Medical Technology Assessment (KEMTA), Maastricht University Medical Center +, PO Box 5800, 6202 AZ Maastricht, The Netherlands

## Introduction

Inflammatory bowel disease (IBD) is a chronic inflammatory condition that primarily affects the gastrointestinal tract but also has systemic and extra-intestinal implications, such as fatigue, joint pain, and reduced psychosocial functioning, influencing various aspects of overall health and well-being [1, 2]. Therefore, comprehensive disease control of IBD extends beyond achieving endoscopic remission and intestinal symptom relief, necessitating a holistic proactive approach [3]. Integrating the assessment of health-related physical fitness (HRPF) parameters within the monitoring framework of IBD may offer valuable insights into the patient's functional capacity and overall health status. Early detection of alterations in HRPF parameters could enable healthcare providers to intervene promptly, tailor interventions to individual needs, and promote enhanced patient engagement in managing their condition. Ultimately, this approach could potentially improve health outcomes and enhance well-being for patients with IBD. Yet, the existing evidence on the status and impact of HRPF components in patients with IBD remains notably limited [4].

HRPF is considered a multidimensional construct that includes body composition, cardiorespiratory fitness, muscular strength, muscular endurance, and flexibility [5]. In patients with IBD, factors such as a sedentary lifestyle, poor nutritional status, chronic inflammation, and medication usage may exert adverse effects on these HRPF components [6–8]. A critical need exists for a deeper understanding of the relevance of HRPF parameters in the context of IBD, establishing its correlation with patient- and disease-specific factors and examining its implications on clinical and patient-reported outcomes. This necessitates an accurate and structural assessment of HRPF parameters in clinical practice as well as in interventional studies to additionally be able to explore the influence of physical activity and physical exercise training on HRPF among patients with IBD.

Unfortunately, the most accurate assessment methods (*i.e.*, gold standard tests) for the components of HRPF are not always feasible to perform in clinical practice because the procedures are time-consuming and necessitate advanced equipment, trained personnel, and medical supervision. Hence, accurate but less demanding assessment methods are warranted to facilitate routine assessment of HRPF in clinical practice and deliver more practical endpoints for interventional studies. Therefore, the objective of this study was to evaluate the criterion validity of various screening tools and field-based tests for components of HRPF (*i.e.*, body composition, cardiorespiratory fitness, muscular strength, and muscular endurance) compared to gold standard tests in patients with IBD.

## Materials and Methods

### Study Design and Patient Population

We conducted a cross-sectional study in which consecutive patients with Crohn's disease (CD) or ulcerative colitis (UC) were recruited at the Gastroenterology outpatient clinic of the Maastricht University Medical Center + between August 2022 and October 2023. Eligible patients were those aged 18 years or above with an American Society of Anesthesiologists physical status  $\leq$  II, either in remission or exhibiting mild-to-moderate clinical disease activity. Clinical disease activity was assessed at time of inclusion using the Harvey Bradshaw Index (HBI) for CD and the Simple Clinical Colitis Activity Index (SCCAI) for UC, both of which are diagnostic questionnaires measuring symptom severity (e.g., well-being, pain, bowel frequency, bleeding), with higher scores indicating greater clinical disease severity [9, 10]. Remission was defined as a HBI  $<$  5 or a SCCAI  $<$  3, mild disease activity as a HBI 5–7 or a SCCAI 3–5, and moderate disease activity as a HBI 8–16 or a SCCAI 6–11. Exclusion criteria were severe clinical disease activity (HBI  $>$  16 or SCCAI  $>$  11), or injuries, severe (neuro)muscular, rheumatic, or orthopedic conditions potentially affecting study evaluations. Pregnant or lactating women and (competitive or elite) athletes were also excluded. The Physical Activity Readiness Questionnaire was administered to identify any potential health risks associated with exercise performance [11].

Participants completed several questionnaires and performed field-based tests as well as gold standard tests for the HRPF components (Fig. 1). Field-based tests and gold standard tests were performed on two separate days, allowing a maximum interval of 14 days in between. All tests were carried out by a single trained clinical researcher (KD). Selection of questionnaires and field-based tests was based on a comprehensive review of the literature and expert opinion. The flexibility component was not incorporated in the study design since defining a single universally accepted gold standard test is difficult due to its high joint-specific nature. Instead, field-based tests focusing on different body parts are frequently used to provide a comprehensive evaluation of flexibility [12, 13].

### Data Collection

Baseline characteristics, including sex, age at inclusion, Charlson Comorbidity Index, disease entity and duration, disease classification according to the Montreal classification, medication use, and previous intestinal resections, were collected from electronic patient records. In addition, fecal calprotectin values were collected if they were available



**Fig. 1** Schematic overview of the study design. Created in BioRender. Demers, K. (2024) BioRender.com/h33o532. *BMI* body mass index, *CPET* cardiopulmonary exercise test, *D<sub>2</sub>O* deuterium oxide,

*DASI* Duke activity status index, *M-DASI* modified Duke activity status index, *BIA* bioelectrical impedance analysis, *VSAQ* veterans-specific activity questionnaire. \*field-based test

within an 8-week period around inclusion, with levels below 250 ug/g indicating biochemical remission. Information on smoking status was acquired during the study visit.

**Gold Standard Tests**

The deuterium oxide (D<sub>2</sub>O) dilution method was used as gold standard test to assess body fat percentage and was performed according to the Maastricht protocol [14, 15]. This method relies on the principle of dilution, where deuterium, a stable isotope of hydrogen, is introduced into the body's water compartment. Deuterium evenly disperses in body water and allows for the estimation of total body water based on the degree of dilution observed in bodily fluids, such as urine. The participant collected a background urine sample prior to consuming approximately 75 ml of deuterium-enriched water in the evening, which resulted in an enrichment of 50–150 ppm. After an overnight equilibration, a urine sample of the second voiding was collected the next morning. Isotope enrichment in the urine samples was analyzed with isotope ratio mass spectrometry. A hydration fraction of fat-free mass of 0.73 was applied to determine the total amount of fat-free mass, fat mass, and subsequently, body fat percentage [16].

Cardiorespiratory fitness was assessed by measuring oxygen uptake at peak exercise (VO<sub>2peak</sub>) using a cardiopulmonary exercise test (CPET) on an electronically braked cycle ergometer [17]. Resting values were collected during

a 2-min seated rest period on the cycle ergometer, followed by a 3-min warm-up of unloaded cycling. Subsequently, the work rate was continuously increased in a ramp-like manner depending on the participant's estimated fitness level, aiming to achieve a maximal effort within eight to twelve minutes. Throughout the test, participants had to maintain a pedaling frequency of between 70 and 80 revolutions per minute. The test was terminated when the pedaling frequency dropped below 60 revolutions per minute, despite strong verbal encouragement, indicating the point of peak exercise. During the test, participants breathed through an oro-nasal face-mask (V2 series, Hans Rudolph Inc., Shawnee, KS, USA) connected to an ergospirometry system (Vyntus CPX, Vyaire Medical Inc., Mettawa, IL, USA) calibrated for respiratory gas analysis and volume measurements. Additionally, participants were fitted with a 12-lead electrocardiogram (Vyntus ECG, Vyaire Medical Inc., Mettawa, IL, USA) to measure heart rate and rhythm. Blood pressure was monitored every 2 min (Tango M2, SunTech Medical Inc., Morrisville, NC, US), and continuous monitoring of SpO<sub>2</sub> was conducted at the ear (8000Q2 Oximetry Sensor, NONIN Medical Inc., Plymouth, Minnesota, USA). Respiratory gas analysis values at peak exercise were calculated as the average value over the final 30 s before test termination. The VO<sub>2peak</sub> was considered (near-)maximal in case of a respiratory exchange ratio at peak exercise (RER<sub>peak</sub>) greater than 1.10 and/or a heart rate at peak exercise (HR<sub>peak</sub>) exceeding 85% of the predicted maximum heart rate (i.e., 208–0.7 × participant's

age). In addition, participants needed to show subjective signs of maximal effort, such as unsteady cycling, sweating, facial flushing, and clear unwillingness to continue despite encouragement. The level of perceived exertion before and directly after the CPET was assessed with the 6–20 Borg rating of perceived exertion scale [18].

Muscular strength and endurance of the quadriceps and hamstring muscles of the dominant leg were assessed using the Biodex System 4 Pro dynamometer (Biodex Medical System Inc, Shirley, NY, US). [19] Prior to the assessment, participants engaged in a 5-min submaximal cycling warm-up on a cycle ergometer. Subsequently, participants were positioned in the dynamometer chair using shoulder, leg, and abdominal straps to minimize compensatory movements. The dominant leg was fixated to the Biodex system just above the ankle joint. Initially, participants performed a submaximal isokinetic quadriceps followed by a hamstring contraction to acquaint them with the testing procedure. Subsequently, participants completed three series of maximal repetitions, including three repetitions at 60°/s, five at 90°/s, and 30 at 180°/s, with a 60-s rest period between each series. A single repetition was defined as a maximal isokinetic contraction of the quadriceps muscles, followed by a contraction of the hamstring muscles. Maximal isokinetic peak torque, expressed in Newton-meter (Nm) normalized for body mass, was determined as the average of the two highest peak torques. The evaluation of muscular endurance involved two fatigue index methods derived from the performance of 30 repetitions at 180°/s: the peak torque fatigue index and the work fatigue index [20, 21]. The peak torque fatigue index assesses the decline in peak torque output, providing insight into the sustainability of maximal strength within the muscles, while the work fatigue index measures the decline in overall energy output. Employing both peak torque and work fatigue indexes allowed for the examination of diverse facets of muscular endurance. The following formulas were used:

Peak torque (in Nm) fatigue index: =  $100 - (\text{sum of peak torque from last 5 repetitions} / \text{sum of peak torque from the highest consecutive 5 repetitions}) \times 100$ ;

Work (in Joule) fatigue index =  $100 - (\text{work performed during last 10 repetitions} / \text{work performed during first 10 repetitions}) \times 100$ .

### Screening Tools

Patients self-assessed their physical fitness on a numeric rating scale (NRS), which ranged from 1, indicating 'not fit', to 10, signifying 'very fit'. In addition, the Duke Activity Status Index (DASI), the Modified DASI (M-DASI), and the Veterans-Specific Activity Questionnaire (VSAQ) were used as screening tools to estimate cardiorespiratory fitness [22–24]. The DASI presents 12 items on daily activities to

predict  $\text{VO}_{2\text{peak}}$ . The M-DASI is a modified version of the DASI that was developed with four questions identified to have dominance in the prognostic significance of the DASI [23]. The total score of the M-DASI reflects the number of positive responses to each of the questions. The VSAQ estimates cardiorespiratory fitness in metabolic equivalents of tasks by measuring an individual's engagement in various physical activities based on barriers and limitations that they might experience. Subsequently, these metabolic equivalents of tasks were translated into estimated  $\text{VO}_{2\text{peak}}$ , utilizing the conversion ratio of one metabolic equivalents of task to 3.5 ml of oxygen per kilogram of body weight per minute (ml/kg/min) [25].

### Field-Based Tests

The field-based tests comprised body mass index (BMI), body fat percentage measured by using four-site skinfold thickness, and multi-frequency bioelectrical impedance analysis (BIA), waist circumference and mid-upper arm circumference (MUAC) measurements, handgrip strength and endurance, quadriceps and hamstring muscular strength measured with hand-held dynamometry, the 60-s sit-to-stand test, and the steep ramp test.

Body mass was assessed to the nearest 0.1 kg without shoes and in lightweight clothing, with emptied pockets, utilizing an electronic scale (Seca ghmb & co., Hamburg, Germany). Body height was measured to the closest centimeter using a stadiometer (Seca ghmb & co., Hamburg, Germany). BMI was calculated as body mass (in kg) divided by body height (in m) squared. Skinfold thickness was measured at four specific sites (biceps, triceps, subscapular, and suprailiac) using a Harpenden skinfold caliper (HSB-BI, Baty International Ltd, West Sussex, UK). Subsequently, the cumulative sum of these measurements was applied in the age-specific skinfold prediction equation by Durnin & Womersley to estimate body fat percentage [26]. Circumference measurements were performed using a flexible, non-stretchable plastic tape. All measurements were performed three times for each site, where after the average of the two closest measurements was calculated. Waist circumference was measured at the horizontal plane midway between the highest point of the iliac crest and the lowest rib at the end of a normal expiration. MUAC was measured at the posterior side of the non-dominant mid-upper arm. Evaluation of upper arm composition relied on standardized equations derived from MUAC and triceps skinfold (TS) measurements: [27]

$$\text{Total upper arm area (cm}^2\text{)} = \text{MUAC (cm)}^2 / (4 \times \pi);$$

$$\text{Arm muscle area (cm}^2\text{)} = (\text{MUAC (cm)} - (\text{TS (cm)} \times \pi))^2 / (4 \times \pi);$$

$$\text{Upper arm fat area (cm}^2\text{)} = \text{total upper arm area (cm}^2\text{)} - \text{arm muscle area (cm}^2\text{)};$$

Upper arm fat index = (upper arm fat area/total upper arm area)  $\times$  100.

Furthermore, body fat percentage was assessed through multi-frequency BIA (InBodyS10, InBody Co., Ltd., Seoul, Korea) based on the electrical conduction of alternating currents through the body, following the manufacturer's guidelines. Participants fasted for a minimum of four hours before the analysis and were questioned about alcohol and excessive caffeine intake within the past 24 h, recent physical activity within the last hour, and menstrual status (if applicable) prior to the measurement.

The steep ramp test was conducted using an electronically braked cycle ergometer (Lode Corival Rehab, Lode B.V., Groningen, the Netherlands). The test started with a 3-min warm-up phase of unloaded cycling. Following this, the work rate was incrementally increased by 25 W every 10 s in a ramp-like manner. Participants were instructed to keep cycling until exhaustion, with a pedaling frequency of 70 to 80 rpm. The main outcome of the test was the work rate attained at peak exercise ( $WR_{\text{peak}}$ ) normalized for body mass, marking the moment when the pedaling frequency dropped below 60 rpm, despite strong verbal encouragement. Throughout the test, continuous measurements were taken for heart rate (Polar H9, Polar Electro Oy, Kempele, Finland) and peripheral oxygen saturation ( $SpO_2$ ) (NONIN Onyx Vantage 9590, NONIN Medical Inc., Plymouth, Minnesota, USA). Perceived exertion prior to and directly after the steep ramp test was measured using the 6–20 Borg scale for rating of perceived exertion. [18].

Handgrip strength of the dominant hand was measured using the electronic JAMAR dynamometer (5030 J1, Sammons Preston Rolyan, Bollingbrook, IL, US). Participants were seated with their shoulders at 0-degree abduction and neutral rotation, elbows flexed to 90 degrees, and their forearm and wrist in a neutral position. To measure handgrip strength, a maximal isometric contraction was performed three times, with the best result counting, expressed in kg normalized for body mass. Subsequently, participants were instructed to perform a maximal isometric handgrip contraction and maintain this contraction for as long as possible to measure handgrip endurance [28]. The test was terminated if the strength declined to less than 50% of their individual handgrip strength. The total time was recorded, rounded to the nearest 0.01 s. The onset handgrip strength of the endurance test was allowed to deviate by a maximum of 20% from handgrip strength.

Muscular strength of the quadriceps and hamstring muscles of the dominant leg was manually assessed using a hand-held dynamometer (MicroFET2, Hoggan Scientific LLC, Salt Lake City, Utah, US). Participants were seated on a bench, with their legs dangling over the end, with knees and hips bend to 90 degrees. The hand-held dynamometer was held in the palm of the hand of the clinical researcher

and placed on the participants' lower limb, just above the ankle joint. To ensure a stable position, the clinical researcher was positioned with one knee on the ground. Participants performed three 5-s maximal isometric contractions for the quadriceps muscles as well as for the hamstring muscles. Maximal isometric torque was defined as the average of the highest two out of three peak torques and expressed in Newton normalized for body mass.

The 60-s sit-to-stand test was conducted using a chair of standard height (46 cm) and without armrests. Participants were instructed to repeatedly stand up and sit down as fast as possible within one minute while keeping their arms folded across their chest. The total count of repetitions performed within one minute was recorded.

### Statistical Analyses

Statistical analyses were performed using IBM SPSS Statistics (IBM Corp., Armonk, NY, USA). Patient and clinical characteristics were presented using descriptive statistics. Continuous variables were presented as mean (standard deviation (SD)) or median (1st quartile (Q1), 3rd quartile (Q3)), according to normality and the presence of outliers. Criterion validity of the questionnaires and field-based tests was evaluated by Pearson's correlation coefficients or Spearman's rank correlation, according to normality and the presence of outliers. The sample size was determined according to the rule-of-thumb recommendation for validation studies, which suggests including a minimum of 50 participants. [29] Correlation coefficients ranging from 0 to 0.30 were considered negligible, from 0.30 to 0.50 as weak, from 0.50 to 0.70 as moderate, from 0.70 to 0.90 as strong, and from 0.90 to 1.00 as very strong [30]. The  $R^2$  was calculated to evaluate the variance in the gold standard test that could be explained by the field-based test. Additionally, when two tests quantified a construct using the same unit, agreement was quantified with a two-way mixed Intraclass Correlation Coefficient (ICC) and visually inspected using a Bland–Altman plot [31]. The predictive accuracy of the screening tools for cardiorespiratory fitness was examined based on the tool's ability to correctly classify individuals as having a  $VO_{2\text{peak}}$  below 80% of the predicted value based on age- and sex-specific reference values from the Lowlands Fitness Registry [32]. A Receiver Operating Characteristic (ROC) curve analysis was performed for each screening tool, and the Area Under the Curve (AUC) of the ROC was calculated with corresponding 95% confidence intervals (CI). To investigate the potential synergistic effects of combining multiple screening tools or field-based tests in enhancing the predictive accuracy for the gold standard measurements, a post hoc analysis was conducted using stepwise multiple linear regression analysis. At each step, the screening tool or field-based test with the highest  $p$ -value was removed

from the model. A  $p$ -value  $< 0.05$  was considered statistically significant.

## Ethical Statement

This study was conducted in compliance with the Declaration of Helsinki and was approved by the Medical Ethical Committee of the Maastricht University Medical Center + (registration no. 22–012). All patients gave written informed consent.

## Results

### Study Population

In total, 53 patients were included in this study, whose patient and clinical characteristics are presented in Table 1. Of these, 29 (54.7%) were diagnosed with CD and 24 (45.3%) with UC. The mean age of the study population was 43.8 years and 45.3% were female. The majority of patients were in remission, with 75.5% in clinical remission and 85.4% in biochemical remission. Biologic agents were the predominant form of treatment, used by 58.5% of the patients. A prior intestinal resection was reported in 41.4% of the patients with CD, while none of the patients with UC had a prior resection.

### Outcomes of Screening Tools and Field-Based Tests

Table 2 presents the results of the screening tools, field-based tests, and gold standard tests for the assessment of HRPF. The screening tool self-perceived fitness was missing in 3 patients. In addition, one assessment of body fat percentage with D<sub>2</sub>O was excluded due to the patient's failure to comply with the urine collection protocols, and the VO<sub>2peak</sub> data from another patient was excluded due to the inability to achieve a maximal effort at the CPET. Additional outcomes of the steep ramp test and the CPET can be retrieved from Supplemental Table 1.

### Validity and Predictive Capacity of the Screening Tools for Cardiorespiratory Fitness

The ICC for the DASI-estimated VO<sub>2peak</sub> and the VSAQ-estimated VO<sub>2peak</sub> with the CPET VO<sub>2peak</sub> were low (ICC 0.27 (95% CI 0.00–0.51) and ICC 0.35 (95% CI – 0.08–0.65), respectively), indicating a poor agreement between these questionnaires and CPET measures of cardiorespiratory fitness (Table 3). Spearman's correlation coefficients showed a weak positive association for self-perceived physical fitness ( $r=0.32$ ) and moderate positive associations for the DASI ( $r=0.51$ ), the M-DASI ( $r=0.52$ ), and the VSAQ ( $r=0.62$ )

with CPET. Scatterplots showing the relationship between the screening tools and CPET VO<sub>2peak</sub> can be found in Supplemental Fig. 1. The VSAQ demonstrated the highest predictive accuracy for identifying individuals with a VO<sub>2peak</sub> less than 80% of the predicted value with an AUC of 0.727 (Fig. 2) compared to the other screening tools.

Bland–Altman plots were created to visualize the agreement between the estimated VO<sub>2peak</sub> from the DASI and VSAQ and CPET-VO<sub>2peak</sub> (Fig. 3). For the DASI, a mean (SD) bias of – 0.62 (8.65) ml/kg/min was observed with signs of systematic bias compared to CPET. The DASI tended to overestimate at lower VO<sub>2peak</sub> values and underestimate at higher VO<sub>2peak</sub> values. The 95% limits of agreement ranged widely from – 17.58 to 16.34 ml/kg/min, indicating substantial variability. For the VSAQ, a mean (SD) bias of 11.29 (9.56) ml/kg/min was found, indicating a general tendency for the VSAQ to overestimate VO<sub>2peak</sub> values without signs of systematic bias. The 95% limits of agreement also varied substantially, ranging from – 7.45 to 30.02 ml/kg/min.

### Validity of the Field-Based Tests

Criterion validity of the field-based tests for the components of HRPF are displayed in Table 4, while scatterplots visualizing these relationships can be retrieved from Supplemental Figs. 1–6. With regard to body composition, both BIA and four-site skinfold thickness showed good agreement in the measurements of body fat percentage in relation to D<sub>2</sub>O dilution (BIA: ICC 0.93, 95% CI 0.82–0.97; four-site skinfold thickness: ICC 0.90, 95% CI 0.84–0.94), with a very strong linear relationship and high degree of explained variance (BIA:  $r=0.95$ ,  $R^2=0.90$ ; four-site skinfold thickness:  $r=0.90$ ,  $R^2=0.82$ ). Bland–Altman plots demonstrated a minor negative mean bias using both four-site skinfold thickness and BIA (– 0.61% and – 1.80%, respectively) as compared to D<sub>2</sub>O dilution, without signs of systematic error (Fig. 4). The limits of agreement for four-site skinfold thickness ranged from – 8.31% to + 7.08%, whereas for BIA they were narrower, ranging from – 7.54% to + 3.93%. Strong linear correlations were also observed for upper arm fat index ( $r=0.80$ ,  $R^2=0.63$ ) and upper arm fat area ( $r=0.76$ ,  $R^2=0.58$ ). BMI showed a moderate correlation ( $r=0.60$ ,  $R^2=0.36$ ), while weak correlations were observed for waist circumference ( $r=0.47$ ,  $R^2=0.23$ ) and MUAC ( $r=0.42$ ,  $R^2=0.18$ ).

For cardiorespiratory fitness, a very strong linear correlation was found between the WR<sub>peak</sub> attained at the steep ramp test and VO<sub>2peak</sub> achieved during CPET ( $r=0.95$ ,  $R^2=0.89$ ). Additionally, a moderate correlation was observed for the 60-s sit-to-stand test ( $r=0.69$ ,  $R^2=0.48$ ).

Regarding quadriceps muscular strength, strong linear correlations were identified between the WR<sub>peak</sub> achieved

**Table 1** Patient and clinical characteristics of included patients

Patient and clinical characteristics	Total (n=53)	CD (n=29)	UC (n=24)
Age at inclusion, <i>mean (SD)</i>	43.8 (14.9)	42.2 (13.6)	45.7 (16.3)
Sex, female, <i>n (%)</i>	24 (45.3)	14 (48.3)	10 (41.7)
Charlson Comorbidity Index, <i>n (%)</i>			
0	31 (58.5)	16 (55.2)	15 (62.5)
1–2	19 (35.8)	13 (44.8)	6 (25.0)
> 2	3 (5.7)	0 (0)	3 (12.5)
Disease duration (years), <i>median (Q1, Q3)</i>	11.3 (7.0, 22.7)	11.3 (7.2, 25.4)	11.1 (5.9, 17.9)
Montreal age at diagnosis, <i>n (%)</i>			
A1: ≤ 16 years	7 (13.2)	5 (17.2)	2 (8.3)
A2: 17–40 years	36 (67.9)	19 (65.5)	17 (70.8)
A3: > 40 years	10 (18.9)	5 (17.2)	5 (20.8)
Montreal disease location [CD], <i>n (%)</i>			
L1: ileal		10 (34.5)	
L2: colonic		4 (13.8)	
L3: ileocolonic		15 (51.7)	
Perianal disease		7 (13.2)	
Upper gastrointestinal disease		3 (10.3)	
Montreal disease behavior [CD], <i>n (%)</i>			
B1: non-stricturing, non-penetrating		14 (48.3)	
B2: stricturing		6 (20.7)	
B3: penetrating		9 (31.0)	
Montreal disease extension [UC], <i>n (%)</i>			
E1: proctitis			1 (4.2)
E2: left-sided colitis			9 (37.5)
E3: pancolitis			14 (58.3)
HBI score [CD], <i>median (Q1, Q3)</i>		3 (2, 5)	
SCCAI score [UC], <i>median (Q1, Q3)</i>			1 (0, 1)
Clinical disease activity <sup>a</sup> , <i>n (%)</i>			
Remission	40 (75.5)	20 (69.0)	20 (83.3)
Mild disease activity	12 (22.6)	8 (27.6)	4 (16.7)
Moderate disease activity	1 (1.9)	1 (3.4)	0 (0)
Fecal calprotectin (µg/g), <i>median (Q1, Q3)</i>	38.0 (16.0, 158.3)	68.5 (26.5, 222.5)	22.5 (14.9, 46.8)
Biochemical disease activity <sup>b</sup> , <i>n (%)</i>			
Remission	41 (85.4)	21 (80.8)	20 (90.9)
Active disease	7 (14.6)	5 (19.2)	2 (9.1)
Current smoking status, <i>n (%)</i>			
Smokes daily	4 (7.5)	3 (10.3)	1 (4.2)
Smokes occasionally	3 (5.7)	2 (6.9)	1 (4.2)
Ex-smoker	21 (39.6)	10 (34.5)	11 (45.8)
Never smoked	25 (47.2)	14 (48.3)	11 (45.8)
Current IBD medication, <i>n (%)</i>			
None	10 (18.9)	7 (24.1)	3 (12.5)
Mesalazine only	5 (9.4)	2 (6.9)	3 (12.5)
(Topical) corticosteroids	2 (3.8)	1 (3.4)	1 (4.2)
Immunomodulators	5 (9.4)	2 (6.9)	3 (12.5)
Biologic agents	31 (58.5)	17 (58.6)	14 (58.3)
Prior intestinal resection, <i>n (%)</i>	12 (22.6)	12 (41.4)	0 (0)

CD Crohn's disease, HBI Harvey Bradshaw Index, IBD inflammatory bowel disease, Q1 1st quartile, Q3 3rd quartile, SCCAI simple clinical colitis activity index, SD standard deviation, UC ulcerative colitis

<sup>a</sup>Clinical disease activity according to the HBI for CD and the SCCAI for UC: remission was defined as HBI < 5 or SCCAI < 3, mild disease activity as HBI 5–7 or SCCAI 3–5, and moderate disease activity as HBI 8–16 or SCCAI 6–11;

<sup>b</sup>Available in n = 48; remission was defined as fecal calprotectin < 250 µg/g;

**Table 2** Main outcomes of gold standard tests, screening tools, and field-based tests for components of health-related physical fitness

	Total (n = 53)	CD (n = 29)	UC (n = 24)
<b>Gold standard tests</b>			
D <sub>2</sub> O dilution-derived body fat (%) <sup>a</sup> , mean (SD)	30.6 (9.1)	31.7 (10.0)	29.2 (7.9)
CPET VO <sub>2peak</sub> (ml/kg/min) <sup>b</sup> , mean (SD)	33.1 (9.7)	31.0 (9.6)	35.8 (9.3)
CPET VO <sub>2peak</sub> percentage of predicted (%) <sup>b</sup> , mean (SD)	84.6 (17.8)	79.3 (16.9)	91.4 (16.8)
Quadriceps peak torque, 60°/s (Nm/kg), mean (SD)	1.8 (0.6)	1.8 (0.6)	1.8 (0.6)
Quadriceps peak torque, 90°/s (Nm/kg), mean (SD)	1.6 (0.5)	1.7 (0.5)	1.6 (0.5)
Quadriceps peak torque, 180°/s (Nm/kg), mean (SD)	1.2 (0.4)	1.2 (0.4)	1.2 (0.3)
Hamstring peak torque, 60°/s (Nm/kg), mean (SD)	1.0 (0.3)	1.1 (0.3)	1.0 (0.3)
Hamstring peak torque, 90°/s (Nm/kg), mean (SD)	1.0 (0.3)	1.0 (0.3)	0.9 (0.3)
Hamstring peak torque, 180°/s (Nm/kg), mean (SD)	0.8 (0.2)	0.8 (0.2)	0.8 (0.2)
Quadriceps peak torque fatigue index, median (Q1, Q3)	33.5 (28.8, 40.8)	35.9 (29.9, 45.6)	30.4 (26.2, 33.7)
Hamstring peak torque fatigue index, median (Q1, Q3)	25.6 (18.0, 30.5)	27.0 (22.4, 33.4)	21.7 (17.8, 29.2)
Quadriceps work fatigue index, median (Q1, Q3)	40.1 (30.8, 46.7)	42.4 (35.3, 50.0)	33.4 (29.2, 41.9)
Hamstring work fatigue index, median (Q1, Q3)	33.8 (25.7, 39.9)	37.4 (30.1, 43.4)	30.6 (22.6, 35.9)
<b>Screening tools</b>			
Self-perceived fitness (NRS) <sup>b</sup> , median (Q1, Q3)	7.0 (6.0, 7.0)	6.5 (6.0, 7.0)	7 (6.25, 7.0)
DASI-estimated VO <sub>2peak</sub> (ml/kg/min), median (Q1, Q3)	34.6 (30.3, 34.6)	34.6 (29.1, 34.6)	34.6 (31.4, 34.6)
M-DASI (points), median (Q1, Q3)	4 (3, 4)	4 (3, 4)	4 (3, 4)
VSAQ-estimated VO <sub>2peak</sub> (ml/kg/min), median (Q1, Q3)	45.0 (36.3, 54.7)	39.6 (35.9, 52.0)	49.0 (36.9, 56.2)
<b>Field-based tests</b>			
BMI (kg/m <sup>2</sup> ), mean (SD)	26.3 (4.4)	27.4 (5.1)	25.1 (2.9)
Waist circumference (cm), mean (SD)	91.3 (12.4)	93.4 (14.0)	88.6 (9.7)
Upper arm anthropometry			
MUAC (cm), mean (SD)	30.6 (3.4)	31.3 (3.7)	29.7 (2.9)
Total upper arm area (cm <sup>2</sup> ), mean (SD)	75.2 (17.0)	78.9 (18.8)	70.7 (13.6)
Arm muscle area (cm <sup>2</sup> ), mean (SD)	53.0 (12.4)	54.7 (11.4)	50.9 (13.4)
Upper arm fat area (cm <sup>2</sup> ), mean (SD)	22.2 (12.4)	18.6 (14.7)	19.8 (8.70)
Upper arm fat index, mean (SD)	28.7 (11.6)	29.2 (12.2)	23.5 (11.0)
Four-site skinfold thickness-derived body fat (%), mean (SD)	29.9 (8.7)	30.7 (9.5)	28.8 (7.7)
BIA-derived body fat (%), mean (SD)	28.6 (9.3)	30.1 (10.5)	26.8 (7.4)
Handgrip strength (kg/kg), mean (SD)	0.5 (0.1)	0.5 (0.1)	0.5 (0.1)
Handgrip endurance (s), mean (SD)	27.7 (15.3)	27.0 (11.9)	28.6 (18.9)
Hand-held dynamometry quadriceps peak torque (N/kg), mean (SD)	4.6 (1.0)	4.5 (0.9)	4.7 (1.1)
Hand-held dynamometry hamstring peak torque (N/kg), mean (SD)	2.9 (0.7)	2.9 (0.7)	2.9 (0.7)
60-s sit-to-stand test (repetitions), mean (SD)	42.9 (12.4)	41.2 (12.8)	46 (10.9)
Steep ramp test WR <sub>peak</sub> (W/kg), mean (SD)	4.3 (1.0)	4.2 (1.1)	4.5 (1.0)

BMI body mass index, CD Crohn’s disease, CPET cardiopulmonary exercise test, D<sub>2</sub>O deuterium oxide, DASI Duke activity status index, IBD inflammatory bowel disease, M-DASI modified Duke activity status index, METs metabolic equivalent of tasks, BIA bioelectrical impedance analysis, MUAC mid-upper arm circumference, Q1 1st quartile, Q3 3rd quartile, SD standard deviation, UC ulcerative colitis, VO<sub>2peak</sub> oxygen uptake at peak exercise, VSAQ veterans-specific activity questionnaire, WR<sub>peak</sub> work rate at peak exercise

<sup>a</sup>Available in n = 52

<sup>b</sup>Available in n = 50

during the steep ramp test and quadriceps isokinetic peak torque at 60°/s, 90°/s, and 180°/s, with *r* ranging from 0.75 to 0.81 and *R*<sup>2</sup> ranging from 0.56 to 0.65. In addition, moderate correlations were found for the 60-s sit-to-stand test (*r* = 0.66–0.70, *R*<sup>2</sup> = 0.43–0.49), handgrip strength (*r* = 0.59–0.68, *R*<sup>2</sup> = 0.35–0.46), and hand-held dynamometry quadriceps peak torque (*r* = 0.58–0.61, *R*<sup>2</sup> = 0.34–0.38).

For hamstring muscular strength, strong correlations were found between hamstring peak torque measured with hand-held dynamometry and hamstring isokinetic peak torque at 60°/s, 90°/s, and 180°/s (*r* = 0.80–0.81, *R*<sup>2</sup> = 0.65–0.66). Strong correlations were also found for the 60-s sit-to-stand test (*r* = 0.76–0.77, *R*<sup>2</sup> = 0.57–0.60) and the steep ramp test (*r* = 0.74–0.79, *R*<sup>2</sup> = 0.54–0.62), while moderate



**Table 3** Intraclass correlation coefficients and Spearman's correlation coefficients between cardiorespiratory fitness estimated from the screening tools (estimated  $VO_{2peak}$ ) and cardiorespiratory fitness assessed at the CPET ( $VO_{2peak}$ )

Screening tool	ICC (95% CI)	Spearman's $\rho$
Self-perceived physical fitness	–	0.32
DASI	0.27 (0.00–0.51)	0.51
M-DASI	–	0.52
VSAQ	0.35 (–0.08–0.65)	0.62

CI confidence interval, CPET cardiopulmonary exercise test, DASI Duke activity status index, ICC intraclass correlation coefficient, M-DASI modified Duke activity status index, UC ulcerative colitis,  $VO_{2peak}$  oxygen uptake at peak exercise, VSAQ veterans-specific activity questionnaire

correlations were found for handgrip strength ( $r=0.53$ – $0.58$ ,  $R^2=0.27$ – $0.34$ ). Negligible correlations, with  $r$  ranging from  $-0.07$  to  $0.18$ , were identified for handgrip endurance, the 60-s sit-to-stand test, and the steep ramp test in relation to both indices of quadriceps and hamstring muscular endurance.

### Post Hoc Analysis of Combined Screening Tools and Field-Based Tests

Considering the very strong correlations from single field-based tests for body composition and cardiorespiratory fitness, this analysis focused on examining the predictive value of combining screening tools for cardiorespiratory fitness and field-based tests for muscular strength and endurance. Combining various screening tools to predict cardiorespiratory fitness resulted in a modest enhancement of the explained variance ( $R^2$ ), which rose from 31.6% with the use

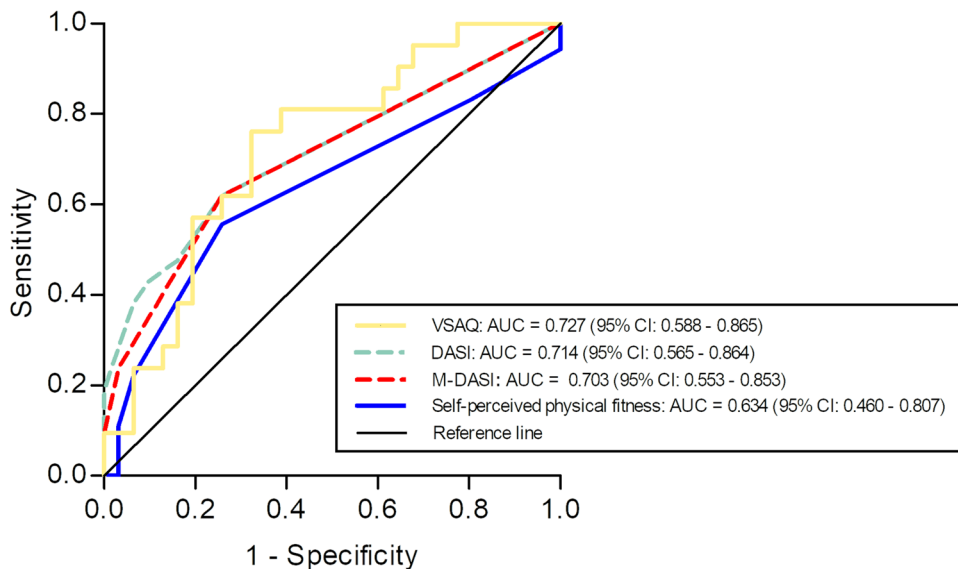
of the VSAQ alone to 41.1% when multiple screening tools were used (Supplemental Table 2). For quadriceps muscular strength, the combination of tests demonstrated limited additional value with a marginal increase of less than 5% in the explained variance when combining multiple tests compared to the singular use of the steep ramp test (Supplemental Table 3). For hamstring muscular strength, the explained variance increased from 64.7 to 66.2% when solely utilizing hand-held dynamometry to 71.1–73.0% when incorporating both hand-held dynamometry and the sit-to-stand test, further rising to 71.6–75.3% with the inclusion of all four tests (Supplemental Table 4). The predictive capacity of the field-based tests remained extremely low for indices of muscular endurance (Supplemental Table 5–8).

### Discussion

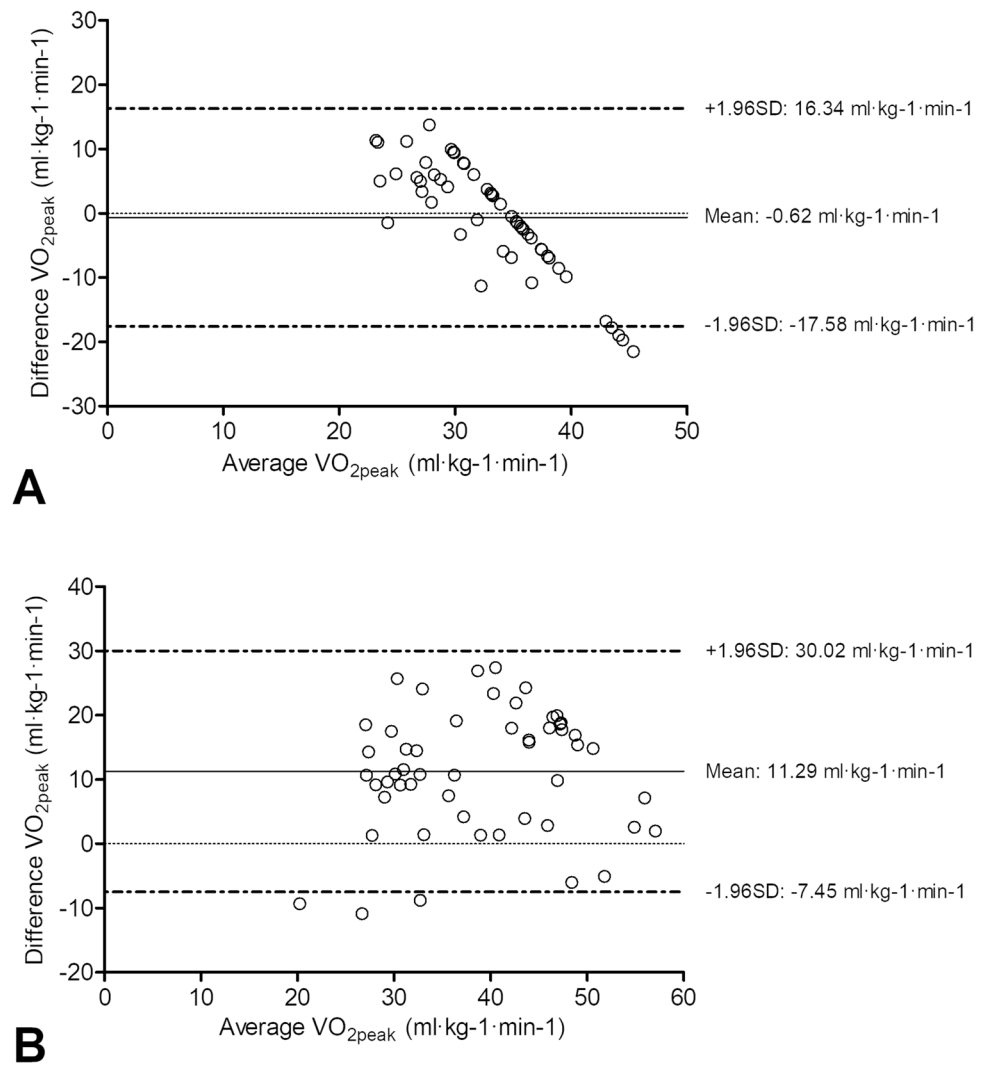
The purpose of this study was to evaluate the criterion validity of various field-based tests for assessing the components of HRF in patients with IBD. Substantial agreement or correlations between several field-based tests and established gold standard tests were found for body composition, cardiorespiratory fitness, and muscular strength, but not for muscular endurance.

For body composition, both BIA and four-site skinfold thickness exhibited good agreement and very strong correlations with the  $D_2O$  dilution method in determining body fat percentage. In the assessment of cardiorespiratory fitness,  $WR_{peak}$  attained at the steep ramp test was very strongly correlated with the gold standard CPET  $VO_{2peak}$ . Moreover, the steep ramp test also exhibited strong correlations in evaluating quadriceps and hamstring muscular strength, as compared to the gold standard of isokinetic peak torque

**Fig. 2** ROC curve analysis showing the predictive accuracy of the of VSAQ (yellow line), DASI (green dotted line), M-DASI (red dotted line), and self-perceived physical fitness (blue line) in detecting a CPET  $VO_{2peak} < 80\%$  of predicted<sup>a</sup>. AUC area under the curve, CI confidence interval, DASI Duke activity status index, M-DASI modified Duke activity status index, ROC receiver operating characteristic, VSAQ veterans-specific activity questionnaire, <sup>a</sup> $VO_{2peak}$  reference values from the Lowlands Fitness Registry



**Fig. 3** Bland–Altman plots showing the agreement between DASI-estimated  $VO_{2peak}$  versus CPET  $VO_{2peak}$  (graph **A**) and VSAQ-estimated  $VO_{2peak}$  versus CPET  $VO_{2peak}$  (graph **B**). The solid line indicates the mean bias, and the dotted lines indicate 95% limits of agreement. *DASI* Duke activity status index,  $VO_{2peak}$  oxygen uptake at peak exercise, *VSAQ* veteran specific activity questionnaire, *SD* standard deviation



dynamometry, highlighting its effectiveness in simultaneous evaluation of both cardiorespiratory fitness and muscular strength parameters. For quadriceps muscular strength, combining tests yielded minimal added value, with less than a 5% increase in explained variance compared to the steep ramp test alone, making the additional effort disproportionate to the limited improvement in predictive accuracy. Additionally, strong correlations with hamstring muscular strength were also observed for hand-held dynamometry and the 60-s sit-to-stand test, with further improvements in explained variance when tests were combined. However, this study identified a noteworthy limitation in the efficacy of both individual and combined field-based tests for assessing muscular endurance, as shown by the negligible correlations that were found with gold standard measures. Furthermore, poor-to-moderate validity was found for the individual patient-reported screening tools (self-perceived physical fitness, DASI, M-DASI, and VSAQ) as well as for their combined use in assessing cardiorespiratory fitness (CPET

$VO_{2peak}$ ). This underscores a limitation in the validity of these self-reported measures and highlights the importance of utilizing objective and precise testing methods to accurately identify areas of HRPF that may need improvement or closer monitoring.

For evaluating body fat percentage, both BIA and four-site skinfold thickness demonstrated good agreement with the  $D_2O$  dilution method, aligning with existing research, generally considering these methods as valid alternatives to assess body composition [33]. Nonetheless, prior research showed the overall validity and reliability of these methods is influenced by a variety of factors such as individual differences (e.g., hydration status, fat distribution), technical aspects (e.g., operator skills and technique, protocol variability), and environmental conditions (e.g., room temperature) [33, 34]. In our study, although the mean biases were minor ( $-0.61\%$  for four-site skinfold thickness and  $-1.80\%$  for BIA), the 95% limits of agreement were notably broad ( $-8.31\%$  to  $7.08\%$  for four-site

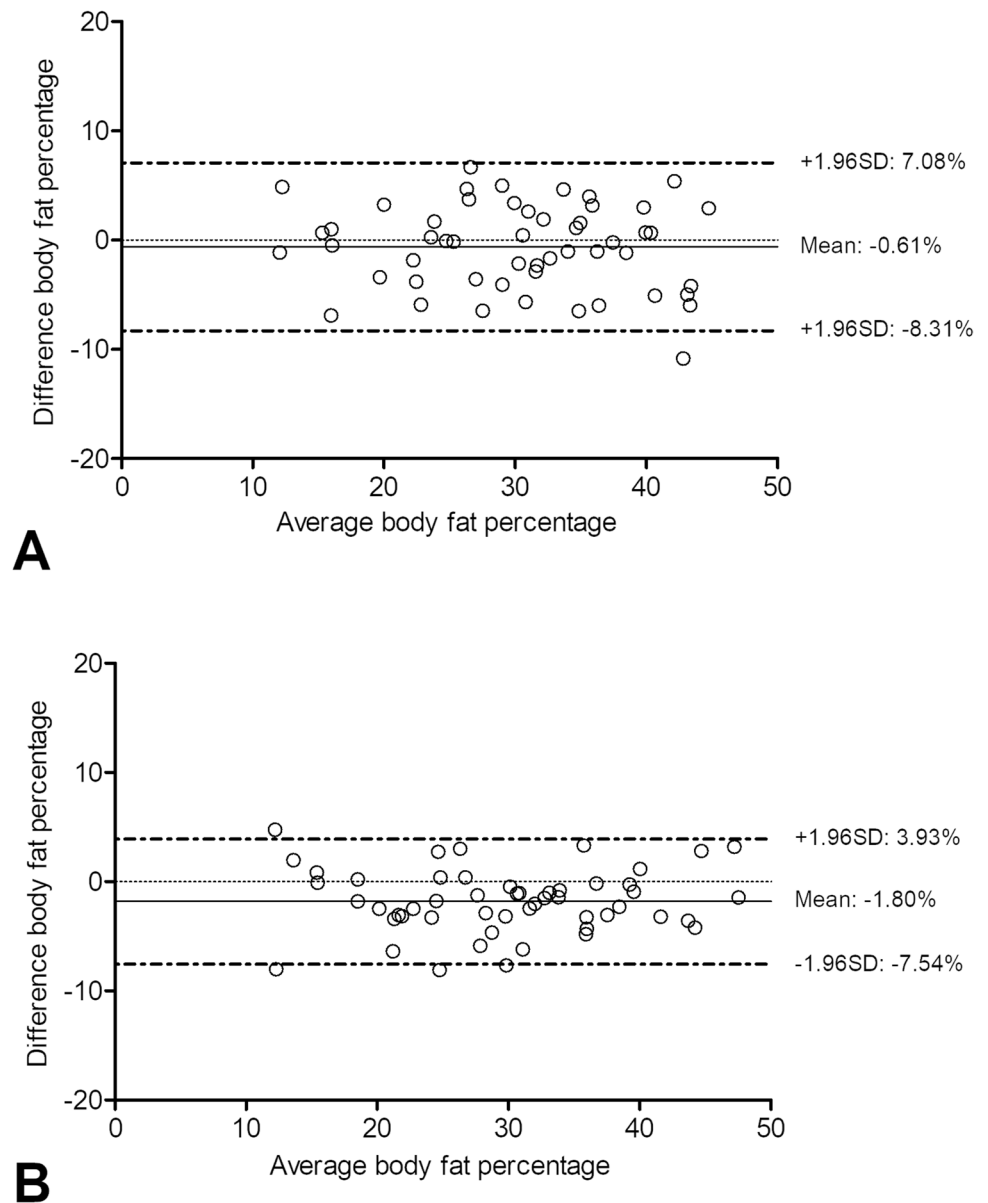
**Table 4** Criterion validity parameters (intraclass correlation coefficients, correlation coefficients, and coefficients of determination) of field-based tests compared to gold standard tests for body composition, cardiorespiratory fitness, muscular strength, and muscular endurance

Field-based tests	Gold standard tests														
	Body composition		Cardiorespiratory fitness		Muscular strength		Muscular endurance		Muscular endurance						
	D <sub>2</sub> O dilution	CPET	Isokinetic quadriceps peak torque (60°/s, 90°/s, 180°/s)	Isokinetic hamstring peak torque (60°/s, 90°/s, 180°/s)	Quadriceps peak torque fatigue index	Hamstring peak torque fatigue index	Quadriceps work fatigue index	Hamstring work fatigue index							
ICC (95% CI)	Pearson's <i>r</i>	R <sup>2</sup>	Pearson's <i>r</i>	R <sup>2</sup>	Pearson's <i>r</i>	R <sup>2</sup>	Pearson's <i>r</i>	R <sup>2</sup>	Pearson's <i>r</i>	Spearman's <i>r</i>	Pearson's <i>r</i>	R <sup>2</sup>	Pearson's <i>r</i>	Spearman's <i>r</i>	
BMI		0.60	0.36												
Waist circumference		0.47	0.23												
MUAC		0.42	0.18												
Upper arm fat area		<b>0.76</b>	0.58												
Upper arm fat index		<b>0.80</b>	0.63												
Four-site skinfold thickness		<b>0.90</b>	0.82												
BIA		<b>0.93 (0.82–0.97)</b>	0.90												
Handgrip strength															
Handgrip endurance															
Hand-held dynamometry quadriceps peak torque															
Hand-held dynamometry hamstring peak torque															
60-s sit-to-stand test															
Sleep ramp test															

Strong intraclass correlation coefficients and correlation coefficients (> 0.70) are highlighted in bold

BMI body mass index, CPET cardiopulmonary exercise test, D<sub>2</sub>O deuterium oxide, BIA bioelectrical impedance analysis, MUAC mid-upper arm circumference, VO<sub>2peak</sub> oxygen uptake at peak exercise

**Fig. 4** Bland–Altman plots showing agreement between body fat percentage obtained from four-site skinfold thickness versus D<sub>2</sub>O dilution (graph **A**) and body fat percentage obtained from BIA versus D<sub>2</sub>O dilution (graph **B**). The solid line indicates the mean bias, and the dotted lines indicate 95% limits of agreement. D<sub>2</sub>O deuterium oxide, BIA bioelectrical impedance analysis, SD standard deviation



skinfold thickness and  $-7.54\%$  to  $3.93\%$  for BIA). These wide limits of agreement indicate considerable individual variability, suggesting that such methods should be interpreted with caution when used for precise individual assessments, despite their general effectiveness for population-level assessment. Furthermore, strong linear correlations were also observed in our study for upper arm fat area and upper arm fat index. However, its validity and reliability in other populations remains relatively unknown in the current literature. While BMI and circumference measurements are commonly used due to their simplicity and cost-effectiveness, both this study and prior research showed their limitations as indicators of body composition due their inability to distinguish between different types of body tissues [35].

The very strong correlation found between steep ramp test performance ( $WR_{peak}$ ) and  $VO_{2peak}$  achieved during CPET in this study also align with previous study findings, underscoring the high validity of the steep ramp test as a proxy for measuring cardiorespiratory fitness across varied demographic groups and clinical conditions [36]. Our findings demonstrating a moderate correlation between the 60-s sit-to-stand test and CPET  $VO_{2peak}$  are also consistent with observations made in other patient populations, in which various adaptations of the 60-s sit-to-stand test were adopted, including the 30-s and the 5-repetition variant [37–39]. The steep ramp test’s higher correlation with CPET  $VO_{2peak}$  in comparison with the sit-to-stand test can be attributed to its greater similarity with CPET in exercise modality and higher cardiovascular and muscular demand.

The high muscular demand of the steep ramp test is further underscored by its strong correlation with quadriceps and hamstring peak torque in our study, although the validity of the steep ramp test for assessing muscular strength has not yet been investigated in other populations. As the steep ramp test is designed to rapidly increase the intensity of exercise, it challenges both the cardiovascular system and the muscle groups involved in cycling, with the quadriceps being crucial for the downstroke, and the hamstring for the upstroke phase of the pedaling action. This highlights the potential of the steep ramp test as a multifaceted tool for assessing both cardiorespiratory fitness and muscular strength.

In our study, the 60-s sit-to-stand test showed a moderate correlation for quadriceps peak torque, but strong correlation with hamstring peak torque. Although studies assessing the validity of the 60-s sit-to-stand test in other populations are limited, comparable findings were seen for various adaptations of the 60-s sit-to-stand test for quadriceps muscular strength in patients with chronic obstructive pulmonary disease and in healthy individuals [39, 40]. Stronger correlations between the 60-s sit-to-stand test and hamstring strength as compared to quadriceps strength in our study could be attributed to the greater involvement and functional demand placed on the hamstrings during the test, including their involvement in hip extension, eccentric control during descent, and stabilization of the pelvis and lower back. In our study, a moderate correlation was observed between handgrip strength and isokinetic peak torque of both the quadriceps and hamstring muscles. Previous research has demonstrated varying correlations between handgrip strength and lower limb strength, with correlations ranging from weak to strong across different populations [41–44]. This variability highlights the challenges in interpreting the relationship between handgrip strength and the strength of lower body muscles, as it can be affected by, for instance, characteristics of the population studied and the methodologies used for testing. Additionally, comparing these two muscle groups is complex due to their physiological differences. Specifically, the muscles involved in handgrip are smaller and designed for fine motor skills, whereas lower body muscles, such as the quadriceps and hamstring, are larger and built for power and stability in weight-bearing and physical activities. In examining the criterion validity of hand-held dynamometry in assessing muscular strength, we noted a moderate correlation with quadriceps strength and a strong correlation with hamstring strength. This suggests a decrease in measurement validity as muscle strength increases, a finding that aligns with prior research. A systematic review by Stark et al. showed that the majority of the 19 included studies showed moderate-to-good reliability of hand-held dynamometry when compared to isokinetic dynamometry, however the lack of homogeneity in testing procedures (*e.g.* patient positioning, force application, tester

technique) made comparisons challenging [45]. In addition, weaker muscle groups (such as those involved in upper body strength or the plantar and dorsiflexion of the ankle) tend to yield higher correlation and reliability compared to stronger muscle groups, such as the knee extensors [46, 47]. Hereby, the strength of the testing practitioner plays a significant role in influencing the validity of hand-held dynamometry outcomes and should be considered when applying this test, particularly in the context of stronger muscle groups.

We found negligible correlations, with Spearman's  $\rho$  ranging from  $-0.07$  to  $0.18$ , between the handgrip endurance, the 60-s sit-to-stand test, and the steep ramp test and indices of quadriceps and hamstring muscular endurance, indicating that these field-based tests may not serve as reliable indicators for assessing endurance of these muscles. This raises questions regarding the gold standard test protocol for measuring muscular endurance. Although various testing protocols have been investigated in the literature, there seems to be a lack of clear consensus on the most effective method [20, 21]. Muscular endurance, often referred to as the concept of muscle fatigue, is thought to vary significantly depending on the specific activity or task being performed, with different activities stressing the muscles in unique ways, leading to various forms of fatigue [48]. Moreover, the perceived sense of effort and actual performance, a psychophysiological phenomenon, can influence motor performance, highlighting the role of cognitive and psychological factors in muscle fatigue. There is a need for more research into the accurate assessment of muscular endurance in patients with IBD and its relevance in relation with clinical disease outcomes and subjective well-being, taking the potential specific physical challenges faced by patients with IBD into account.

The use of patient-reported measures to stratify patients at risk of low cardiorespiratory fitness levels appears promising, as it allows for the prioritization of patients who may benefit most from further diagnostic evaluation or targeted interventions. However, we found suboptimal validity of such patient-reported measures, including self-perceived physical fitness, the (M-)DASI, and the VSAQ, which aligns with previous research. Our study revealed a weak correlation ( $r=0.32$ ) between self-perceived fitness, rated by the patient on an NRS from 1 to 10, and CPET-derived  $VO_{2peak}$ . This is in line with results from a Danish population-based cohort study, which found correlation coefficients of 0.40 for women and 0.37 for men. [49] Although the Danish study utilized a slightly different method by asking patients to rate their fitness in comparison to others of the same age and sex, based on a set of five predefined categories, these consistent findings might indicate that patients' perception of fitness does not align well with objectively measured fitness. With regard to the DASI and VSAQ, moderate validity was observed compared to CPET  $VO_{2peak}$  in the validation

phase for both questionnaires [22, 24]. Moreover, wide 95% limits of agreement and substantial bias were found for both questionnaires in more recent studies [50, 51]. Furthermore, the screening tools showed suboptimal ability to detect a  $VO_{2peak} < 80\%$  of predicted based on age- and sex-specific reference values, with AUC values ranging from 0.634 for self-perceived physical fitness to 0.727 for the VSAQ. This indicates that these patient-reported measures have a limited ability to accurately distinguish between adequate and inadequate levels of cardiorespiratory fitness compared to the objective CPET testing. A notable limitation of the VSAQ is its reliance on activities such as basketball and cross-country skiing, which are less commonly performed in the Netherlands. To address this, the FitMáx©-questionnaire was recently developed as a self-reported questionnaire to estimate  $VO_{2peak}$  in the Dutch population, based on more locally relevant activities like walking, stair climbing, and cycling [52]. The FitMáx©-questionnaire showed better agreement and correlation with CPET- $VO_{2peak}$  in a heterogeneous patient population and in a cohort of cancer survivors [52, 53]. This suggests that it could be a more effective tool for estimating  $VO_{2peak}$ , however, its validity within the IBD population warrants further investigation. For current application, we recommend using such screening tools with caution in patients with IBD and not as direct replacements for comprehensive objective assessments in clinical or research settings.

Patients with IBD often experience impaired subjective disease control and debilitating symptoms such as fatigue or persistent abdominal pain, even in the absence of mucosal inflammation [54]. Achieving the long-term treatment objectives recommended by the STRIDE-II consortium, which aim for the absence of disability and the restoration of quality of life, necessitates a holistic approach that extends beyond medical or surgical treatment alone [3]. Incorporating assessment of contributing factors into routine care, such as physical fitness, is warranted to identify patients who may benefit from targeted interventions, which in turn may contribute to better overall patient and disease outcomes. This proactive approach aligns with the P4 (*Predictive, Preventive, Personalized, and Participatory*) health concept, reflecting healthcare approach that is proactive, personalized, and patient-centered [55]. Yet, existing evidence on the status HRPF components in patients with IBD is limited and a recent scoping review revealed a lack of use of validated assessment methods in the available interventional studies [4]. Therefore, findings of the current study hold significant clinical relevance, especially in situations where the utilization of advanced gold standard tests for assessing HRPF components is not feasible. The strong correlations observed between field-based tests and gold standard measures for various HRPF components suggest that these simpler, and less resource-intensive tests may serve as valid alternatives

for the assessment of HRPF. However, other psychometric properties, such as reliability (*i.e.*, the consistency of test results across repeated assessments) and responsiveness (*i.e.*, the ability of the test to detect changes over time), were not evaluated in this study and should be explored in future studies prior to their implementation in clinical and research settings for individuals with IBD. Furthermore, future research is warranted to determine the clinical utility and interpretability of these field-based assessments, which could ultimately enhance the evidence base regarding the relevance of HRPF parameters within the context of IBD. This entails establishing the relationship with patient- and disease-specific factors, but also examining its implications on clinical and patient-reported outcomes. Besides, this assessment should extend to interventional studies that delve into the influence of physical activity and physical exercise training on these HRPF components in patients with IBD.

A limitation of this study is that our population primarily consisted of patients in remission or with mild disease activity, limiting the generalizability of our findings to the broader IBD population, especially those with moderate to severe disease. Therefore, caution is advised when applying these results to the wider IBD population, highlighting the need for further validation across varying levels of disease severity. Next, all measurements were conducted by a single researcher. While this ensured consistency in testing procedures, the agreement and correlations observed in our study may differ when these measurements are applied by multiple practitioners with varying levels of expertise and technique, as inter-rater variability could potentially impact the outcomes of these tests. Additionally, agreement between tests was assessed only for a subset of screening tools and field-based tests that measured constructs using the same units, whereas for other tests, correlation coefficients were reported, providing no information on systematic bias between the tests. Furthermore, the selection of field-based tests in this study was based on a comprehensive review of the literature, expert opinions, and expected practical applicability. This implies that there may be other potential field-based tests and methods for assessing HRPF components in patients with IBD that were not considered in this study. Furthermore, this study focused only on evaluating the criterion validity of the screening tools and field-based tests and did not investigate other psychometric properties.

## Conclusion

This study evaluated the criterion validity of various screening tools and field-based tests for assessing components of HRPF compared to gold standard tests in patients with IBD. Four-site skinfold thickness and BIA showed good agreement with the gold standard for body fat measurement. The

steep ramp test demonstrated strong correlations with the gold standard tests for assessing cardiorespiratory fitness and muscular strength, while hand-held dynamometry and the sit-to-stand test showed strong correlations with hamstring strength. However, future research is warranted to explore additional psychometric properties, such as reliability, responsiveness, and their clinical interpretation to better understand the usefulness of these tests in clinical and research settings. This may pave the way for integrating HRRPF assessments within the monitoring framework of IBD, enhancing the understanding of the role of HRRPF in IBD management, and ultimately leading to the development of more holistic and proactive management strategies.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s10620-024-08682-5>.

**Acknowledgments** This work was supported by the Dutch Gastroenterology and Hepatology foundation (Maag, Lever, Darmstichting).

**Author contributions** Karlijn Demers (KD), Bart C. Bongers (BB), Sander M.J. van Kuijk (SK), Guy Plasqui (GP), Daisy M.A.E. Jonkers (DJ), Marieke J. Pierik (MP), and Laurents P.S. Stassen (LS) conceived this study. KD performed the measurements under the supervision of BB. GP assisted and supervised the deuterium oxide dilution body composition analyses. KD analyzed the data and wrote the manuscript. SK assisted in the statistical analyses. BB, SK, GP, DJ, MP, and LS contributed to interpreting the results and critically revised the manuscript for important intellectual content. All authors revised the manuscript and approved the final version

**Data availability** The datasets generated and analyzed during the present study are available from the corresponding author upon reasonable request.

## Declarations

**Conflict of interest** Daisy M.A.E. Jonkers reports grant from the public-private partnership grants of Dutch Top Institute of Food and Nutrition (TIFN), Top Knowledge Institute (TKI) Agri&Food and Health Holland, by the Carbokinetics program as part of the NWO-CCC Partnership Program, by Organic A2BV/Mothersfinest BV and, EU/FP7 SysmedIBD/305564, BIOM/305479 and Character/305676, H2020 DISCOVERIE/848228, all outside the submitted work. Marieke J. Pierik reports grants and non-financial support from Falk Pharma, grants from European commission, grants from ZONMW (Dutch national research fund), grants and non-financial support from Takeda, grants and non-financial support from Johnson and Johnson, grants and non-financial support from Abbvie, non-financial support from Ferring, non-financial support from Immunodiagnostics, non-financial support from MSD, all outside the submitted work. Laurents P.S. Stassen serves on the advisory board of Diagnostic Green GmbH and has served as a speaker and received research support from Takeda, outside the submitted work. For the remaining authors none were declared.

**Ethical approval** This study was conducted in compliance with the Declaration of Helsinki and was approved by the Medical Ethical Committee of the Maastricht University Medical Center + (registration no. 22–012).

**Consent to participate** Informed consent was obtained from all individual participants included in the study.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License, which permits any non-commercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc/4.0/>.

## References

- Kappelman MD, Long MD, Martin C et al. Evaluation of the Patient-Reported Outcomes Measurement Information System in a Large Cohort of Patients with Inflammatory Bowel Diseases. *Clin Gastroenterol and Hepatol*. 2014;12:1315–1323.e2.
- Burisch J, Jess T, Martinato M, Lakatos PL, ECCO-EpiCom. The burden of inflammatory bowel disease in Europe. *J Crohns Colitis*. 2013;7:322–337.
- Turner D, Ricciuto A, Lewis A et al. STRIDE-II: An Update on the Selecting Therapeutic Targets in Inflammatory Bowel Disease (STRIDE) Initiative of the International Organization for the Study of IBD (IOIBD): Determining Therapeutic Goals for Treat-to-Target strategies in IBD. *Gastroenterology*. 2021;160:1570–1583.
- Demers K, Bak MTJ, Bongers BC et al. Scoping review on health-related physical fitness in patients with inflammatory bowel disease: Assessment, interventions, and future directions. *World J Gastroenterol*. 2023;29:5406–5427.
- Caspersen CJ, Powell KE, Christenson GM. Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research. *Public Health Rep*. 1985;100:126–131.
- Schakman O, Gilson H, Thissen JP. Mechanisms of glucocorticoid-induced myopathy. *J Endocrinol*. 2008;197:1–10.
- Vogelaar L, van den Berg-Emons R, Bussmann H, Rozenberg R, Timman R, van der Woude CJ. Physical fitness and physical activity in fatigued and non-fatigued inflammatory bowel disease patients. *Scand J Gastroenterol*. 2015;50:1357–1367.
- Dhaliwal A, Quinlan JL, Overthrow K et al. Sarcopenia in Inflammatory Bowel Disease: A Narrative Overview. *Nutrients*. 2021;13:656.
- Harvey RF, Bradshaw JM. A simple index of crohn's-disease activity. *Lancet* 1980;1:514.
- Walmsley RS, Ayres RC, Pounder RE, Allan RN. A simple clinical colitis activity index. *Gut* 1998;43:29–32.
- Thomas S, Reading J, Shephard RJ. Revision of the physical activity readiness questionnaire (par-q). *Can J Sport Sci* 1992;17:338–345.
- Corbin CB, Noble L. Flexibility. *JOPERD*. 1980;51:23–60.
- Swain DP, Brawner CA, and American College of Sports Medicine. ACSM's resource manual for guidelines for exercise testing and prescription: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2012.
- Westerterp KR, Wouters L, van Marken Lichtenbelt WD. The Maastricht protocol for the measurement of body composition and energy expenditure with labeled water. *Obes Res*. 1995;3:49–57.
- Wagner DR, Heyward VH. Techniques of body composition assessment: A review of laboratory and field methods. *Res Q Exerc Sport* 1999;70:135–149.

16. Siri WE. Body composition from fluid spaces and density: analysis of methods 1961. *Nutrition* 1993;9:480–492.
17. Ross R, Blair SN, Arena R et al. Importance of assessing cardiorespiratory fitness in clinical practice: A case for fitness as a clinical vital sign: A scientific statement from the American Heart Association. *Circulation* 2016;134:e653–e699.
18. Borg GA. Psychophysical bases of perceived exertion. *Med Sci Sports Exerc* 1982;14:377–381.
19. Drouin JM, Valovich-mcLeod TC, Shultz SJ, Gansneder BM, Perrin DH. Reliability and validity of the biodex system 3 pro isokinetic dynamometer velocity, torque and position measurements. *Eur J Appl Physiol* 2004;91:22–29.
20. Pincivero DM, Lephart SM, Karunakara RA. Reliability and precision of isokinetic strength and muscular endurance for the quadriceps and hamstrings. *Int J Sports Med*. 1997;18:113–117.
21. Pincivero DM, Gandaio CM, Ito Y. Gender-specific knee extensor torque, flexor torque, and muscle fatigue responses during maximal effort contractions. *Eur J Appl Physiol*. 2003;89:134–141.
22. Hlatky MA, Boineau RE, Higginbotham MB et al. A brief self-administered questionnaire to determine functional capacity (the Duke Activity Status Index). *Am J Cardiol*. 1989;64:651–654.
23. Riedel B, Li MH, Lee CHA et al. A simplified (modified) Duke Activity Status Index (M-DASI) to characterise functional capacity: a secondary analysis of the Measurement of Exercise Tolerance before Surgery (METS) study. *Br J Anaesth*. 2021;126:181–190.
24. Myers J, Bader D, Madhavan R, Froelicher V. Validation of a specific activity questionnaire to estimate exercise tolerance in patients referred for exercise testing. *American Heart Journal*. 2001;142:1041–1046.
25. Liguori, Gary, and American College of Sports Medicine. ACSM's guidelines for exercise testing and prescription. Lippincott Williams & Wilkins; 2020.
26. Durnin JVGA, Womersley J. Body fat assessed from total body density and its estimation from skinfold thickness: measurements on 481 men and women aged from 16 to 72 Years. *Br J Nutr*. 1974;32:77–97.
27. Frisancho AR. *Anthropometric standards for the assessment of growth and nutritional status*. Ann Arbor: University of Michigan Press; 1990.
28. Bautmans I, Mets T. A fatigue resistance test for elderly persons based on grip strength: reliability and comparison with healthy young subjects. *Aging Clin Exp Res*. 2005;17:217–222.
29. de Vet HCW, Terwee CB, Mokkink LB, Knol DL. *Measurement in medicine: A practical guide*. Cambridge: Cambridge University Press; 2011.
30. Mukaka MM. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Med J* 2012;24:69–71.
31. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1:307–310.
32. van der Steeg GE, Takken T. Reference values for maximum oxygen uptake relative to body mass in Dutch/Flemish subjects aged 6–65 years: the LowLands Fitness Registry. *Eur J Appl Physiol*. 2021;121:1189–1196.
33. Marin-Jimenez N, Cruz-Leon C, Sanchez-Oliva D et al. Criterion-Related Validity of Field-Based Methods and Equations for Body Composition Estimation in Adults: A Systematic Review. *Curr Obes Rep*. 2022;11:336–349.
34. Bioelectrical impedance analysis in body composition measurement. National Institutes of Health Technology Assessment Conference Statement. *Am J Clin Nutr*. 1996;64:524S–532S.
35. Sommer I, Teufer B, Szelag M et al. The performance of anthropometric tools to determine obesity: a systematic review and meta-analysis. *Scientific Reports*. 2020;10:12699.
36. Trul-Kreuz IA, Akkerman M, Kieboom EAM, Nieuwenhuis MK, Houdijk H, Bongers BC. Clinimetric properties of the steep ramp test to assess cardiorespiratory fitness, its underlying physiological responses, and its current applications: a scoping review. *Arch Phys Med Rehabil*. 2024. <https://doi.org/10.1016/j.apmr.2024.02.717>.
37. Fuentes-Abolafio IJ, Escriche-Escuder A, Bernal-López MR et al. Estimation of Functional Aerobic Capacity Using the Sit-to-Stand Test in Older Adults with Heart Failure with Preserved Ejection Fraction. *J Clin Med*. 2022. <https://doi.org/10.3390/jcm11102692>.
38. Díaz-Balboa E, González-Salvado V, Rodríguez-Romero B et al. Thirty-second sit-to-stand test as an alternative for estimating peak oxygen uptake and 6-min walking distance in women with breast cancer: a cross-sectional study. *Support Care Cancer*. 2022;30:8251–8260.
39. Bohannon RW, Crouch R. 1-Minute Sit-to-Stand Test: Systematic Review of procedures, performance, and clinimetric properties. *J Cardiopulm Rehabil Prev*. 2019;39:2–8.
40. Bohannon RW, Bubela DJ, Magasi SR, Wang YC, Gershon RC. Sit-to-stand test: Performance and determinants across the age-span. *Isokinet Exerc Sci*. 2010;18:235–240.
41. Felicio DC, Pereira DS, Assumpção AM et al. Poor correlation between handgrip strength and isokinetic performance of knee flexor and extensor muscles in community-dwelling elderly women. *Geriatr Gerontol Int*. 2014;14:185–189.
42. Ostolin TLVDP, Gonze BdB, de Oliveira Vieira W et al. Association between the handgrip strength and the isokinetic muscle function of the elbow and the knee in asymptomatic adults. *SAGE Open Medicine*. 2021;9:2050312121993294.
43. Yeung SSY, Reijnierse EM, Trappenburg MC et al. Handgrip Strength Cannot Be Assumed a Proxy for Overall Muscle Strength. *J Am Med Dir Assoc*. 2018;19:703–709.
44. Bohannon RW, Magasi SR, Bubela DJ, Wang YC, Gershon RC. Grip and Knee extension muscle strength reflect a common construct among adults. *Muscle Nerve*. 2012;46:555–558.
45. Stark T, Walker B, Phillips JK, Fejer R, Beck R. Hand-held Dynamometry Correlation With the Gold Standard Isokinetic Dynamometry: A Systematic Review. *PM R*. 2011;3:472–479.
46. Douma RKW, Soer R, Krijnen WP, Reneman M, van der Schans CP. Reference values for isometric muscle force among workers for the Netherlands: a comparison of reference values. *BMC Sports Sci Med Rehabil*. 2014;6:10.
47. Lu YM, Lin JH, Hsiao SF, Liu MF, Chen SM, Lue YJ. The relative and absolute reliability of leg muscle strength testing by a hand-held dynamometer. *J Strength Cond Res*. 2011;25:1065–1071.
48. Enoka RM, Stuart DG. Neurobiology of muscle fatigue. *J Appl Physiol*. 1992;72:1631–1648.
49. Lerche L, Olsen A, Petersen KEN et al. Validity of physical activity and cardiorespiratory fitness in the Danish cohort “Diet, Cancer and Health-Next Generations.” *Scand J Med Sci Sports*. 2017;27:1864–1872.
50. Li MHG, Bolshinsky V, Ismail H, Ho K-M, Heriot A, Riedel B. Comparison of Duke Activity Status Index with cardiopulmonary exercise testing in cancer patients. *J Anesth*. 2018;32:576–584.
51. da Silva SC, Monteiro WD, Cunha FA, Myers J, Farinatti PTV. Determination of Best Criteria to Determine Final and Initial Speeds within Ramp Exercise Testing Protocols. *Pulm Med*. 2012. <https://doi.org/10.1155/2012/542402>.
52. Meijer R, van Hooff M, Papen-Botterhuis NE et al. Estimating VO<sub>2</sub>(peak) in 18–90 Year-Old Adults: Development and Validation of the FitMáx©-Questionnaire. *Int J Gen Med*. 2022;15:3727–3737.
53. Weemaes ATR, Meijer R, Beelen M et al. Monitoring aerobic capacity in cancer survivors using self-reported questionnaires: criterion validity and responsiveness. *J Patient Rep Outcomes*. 2023;7:73.



54. Lönnfors S, Vermeire S, Greco M, Hommes D, Bell C, Avedano L. IBD and health-related quality of life – discovering the true impact. *J Crohns Colitis*. 2014;8:1281–1286.
55. Hood L. Systems biology and p4 medicine: past, present, and future. *Rambam Maimonides Med J*. 2013;4:e0012.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.