

1 **Intra- and inter-rater reliability, agreement, and minimal detectable**
2 **change of the handheld dynamometer in individuals with symptomatic**
3 **hip osteoarthritis.**

4

5

6 Gilvan Ferreira Vaz^{1,2,¶*}, Felipe Florêncio Freire^{2¶}, Henrique Mansur Gonçalves^{3&},
7 Marcus Alexandre Brito de Aviz^{4&}, Wagner Rodrigues Martins^{5&}, João Luiz Quagliotti
8 Durigan^{5,6&}

9

10 ¹Rehabilitation Sciences Program, University of Brasília, Brasília, DF, Brazil

11 ²Department of Orthopaedics, Medicine Division, Hospital das Forças Armadas (HFA),
12 Brasília, Brazil

13 ³Orthopedic Department of the Santa Helena Hospital, Brasília, DF, Brazil

14 ⁴Anesthesiology Department of the Institute Hospital de Base, Brasília, DF, Brazil

15 ⁵Rehabilitation Sciences Program, University of Brasília, Brasília, DF, Brazil

16 ⁶Laboratory of Muscle and Tendon Plasticity, Rehabilitation Sciences Program,
17 University of Brasília, Brasília, DF, Brazil

18

19 * Corresponding author:

20 Email: gilvanvaz@gmail.com (GFV)

21

22

23 [¶]These authors contributed equally to this work.

24 [&]These authors also contributed equally to this work. NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

25 **Abstract**

26 **Introduction:** The handheld dynamometer has been validated to measure muscle strength
27 in different muscle groups. However, to date, it has not been tested in individuals who
28 experience pain induced by hip osteoarthritis. The current study aimed to evaluate the
29 intra- and inter-rater reliability, agreement, and minimal detectable change of the
30 Lafayette handheld dynamometer, model 1165, to assess the peak force (Pk) and average
31 peak force (Af) of hip muscles in individuals with symptomatic hip osteoarthritis.

32 **Methods:** Twenty participants with hip osteoarthritis (mean \pm SD age: 58.7 \pm 15.3 years;
33 body mass index: 28.8 \pm 4.2 kg/m²) and a pain intensity on the Visual Analogue Scale \geq 4
34 (8.05 \pm 1.2) were recruited to participate in this study. Pk and Af of hip flexors (seated
35 position), abductors and adductors (supine position), and extensors (prone position) were
36 collected in a single day by two independent raters, each one obtaining test and retest in
37 randomly ordered separate sessions. **Results** The intra-rater intraclass correlation
38 coefficient (ICC) was classified as good (>0.75) or excellent (≥ 0.90) for all muscle
39 groups, and all inter-rater ICCs were classified as excellent. Rater A had a lower standard
40 error of measurement compared to rater B, ranging from 0.15 to 0.58 kilogram-force
41 (Kgf) compared with 0.34 to 1.25 kg, respectively. However, the inter-rater comparison
42 showed a minimal detectable change $< 10\%$ for all Pk and Af measures (except Af for the
43 abductor muscle group). Finally, the inter-rater Bland-Altman analysis demonstrated
44 good agreement for abductors, adductors, and extensors. **Conclusion:** Despite pain and
45 dysfunction related to hip osteoarthritis, the handheld dynamometer was shown to be a
46 reliable tool to assess hip muscle strength, with good to excellent intra- and inter-rater
47 ICCs, satisfactory agreement, and small values for minimal detectable change.

48

49 **Keywords:** hip osteoarthritis, muscle strength dynamometer, reproducibility of results.

50 Introduction

51

52 Hip Osteoarthritis (OA) is an end-stage disease from various causes, resulting in
53 chronic hip pain, dysfunction, and stiffness. It is estimated that symptoms are present in
54 5 to 10% of adults older than 40 and 45 years, considering the Spanish and American
55 populations, respectively, with a higher prevalence with increasing age [1,2]. Chronic hip
56 pain is associated with muscle atrophy and weakness, as demonstrated in a meta-analysis
57 conducted with pooled data from thirteen articles. Collectively, the authors observed a
58 reduction in muscle strength in individuals with osteoarthritis that mainly affects hip
59 flexors (-22%) and hip extensors (-21%), or abductors (-31%) and adductors (-25%)
60 compared to healthy control groups [3].

61 Muscle weakness and atrophy seem to have a central role in the dysfunction
62 related to hip osteoarthritis, as demonstrated by imaging studies and isometric
63 dynamometry [3–5]. Both are deeply connected to the degree of radiographic OA
64 classification, and their progress should be avoided through participation in exercise
65 programs that include aerobic and strengthening exercises [6,7]. However, a reliable and
66 easy method seems to be necessary to measure the strength of hip muscles in the clinical
67 routine, in order to monitor disease and treatment progression in the rehabilitation
68 process.

69 Measurement of Peak Force (Pk) has been considered the gold standard method
70 for isokinetic test parameters to evaluate muscle function [8] and can be acquired with
71 fixed or portable dynamometry. Handheld dynamometers (HHD) have been suggested as
72 a practical, feasible, and simple tool to assess isometric lower limb muscle strength in the
73 clinical setting [9] compared to fixed laboratory-based dynamometry, such as isokinetic
74 dynamometers [10,11]. In addition, manual dynamometers require little training for

75 proficient application [10,12] and have lower costs than fixed laboratory-based
76 dynamometry [13,14].

77 Several studies have validated and recommended the use of different HHDs to
78 measure hip muscle strength, with good to excellent intraclass correlation coefficients
79 (ICC). Mentiplay et al (2015) demonstrated the validity of two HHDs compared with a
80 fixed laboratory-based dynamometer, with good to excellent reliability, particularly for
81 proximal muscle groups in the lower limbs of health subjects. Fulcher et al (2010) also
82 found good to excellent reliability when evaluating hip flexors and adductors of young
83 adult football players, in accordance with Florencio et al (2019), who tested young,
84 healthy adults with similar results when testing hip and knee strength. Other authors
85 assessed different hip muscles in various protocols and also recommended the use of
86 manual devices to acquire hip muscle strength in healthy subjects [9,15,16]. Only one
87 study tested the use of the HHD in older people (> 65 years old) and found good reliability
88 for hip and knee muscle strength measures, without specifying lower limb articular
89 disease [10]. There are no definitive findings about the reliability of HHD measurements
90 in participants with symptomatic hip OA.

91 Accordingly, it is crucial to determine if pain intensity related to hip OA could
92 affect the reliability of HHD in muscle strength evaluations of the hip, since comfort may
93 be a potential limitation for strength performance [12]. Furthermore, the standard error of
94 measurement (SEM) and minimal detectable change (MDC) need to be determined to
95 allow comparability for routine measurements in clinical settings of symptomatic hip OA
96 patients. The purpose of this methodological study was to analyze the reliability,
97 agreement, and minimal detectable change of an HHD in individuals with chronic hip
98 pain related to OA. We hypothesized that HHD could be a reliable tool to measure muscle
99 strength for hip muscles even if symptomatic hip OA is present. Our findings could help

100 clinicians and physical therapists to design more rational assessment strategies for
101 individuals with chronic hip pain related to OA, using a tool that requires little training,
102 is low cost, and minimizes the time needed by patients and clinicians.

103

104 **Methods**

105 **Study Design**

106 A methodological study with repeated measures was conducted to determine the
107 intra- and inter-rater reliability, agreement, and MDC for strength assessment of hip
108 muscles obtained with the HHD, testing subjects who experience chronic hip pain.
109 Participants were assessed in a single-day session. Data collection occurred between
110 August 2021 and March 2022 after approval by the local Ethics Committee (CAAE
111 40347320.1.1001.0025), following the Helsinki Declaration of 1975. All participants
112 signed an informed consent before data collection. The research was conducted at the
113 Hospital das Forças Armadas (Brasília, Brazil) and Instituto Hospital de Base (Brasília,
114 Brazil), following the guidelines for reporting reliability and agreement studies (GRRAS)
115 [17].

116 **Participants**

117 Twenty participants {40% female, mean age 58.7 (\pm 15.3) years, age range = 41-
118 79 years, body mass index = 28.8 (\pm 4.2) kg/m²} with symptomatic hip OA were enrolled
119 in the present study from the Orthopedic Department of two tertiary hospitals. Eligibility
120 and demographic data were obtained using an interview questionnaire formulated by the
121 authors. Study procedures were explained to potential participants, and they were

122 assigned to the study protocol if eligible and after giving written informed consent.
123 Participants were included if they presented hip OA radiographically classified as type II
124 (Definite osteophytes, possible joint space narrowing), III (moderate osteophytes, definite
125 joint space narrowing, some sclerosis, possible bone-end deformity), or IV (Large
126 osteophytes, marked joint space narrowing, severe sclerosis and definite bone ends
127 deformity) according to the Kellgren and Lawrence classification [18,19], performed by
128 rater B. All participants had previously been screened with x-ray images as part of their
129 usual care, and no additional image investigation was performed. Other causes of the
130 reported pain, lower limb and back, were also excluded as the primary source of pain, and
131 range of motion was tested to guarantee the hip as the source of the symptoms.

132 **Instruments**

133 Pain intensity was assessed using the Visual Analogue Scale (VAS), with faces
134 ranging from 0 to 10, presented to the participants at the eligibility interview and after
135 each protocol sequence of muscle strength assessment [20,21]. The Western Ontario and
136 McMaster Universities Index (WOMAC) was used as a Health-related Quality of Life
137 (HrQOL) questionnaire developed for patients with hip and knee OA as a self-reported
138 tridimensional scale. The questionnaire evaluates pain, function, and joint stiffness (five
139 questions for the subscale of pain, two questions for the subscale of stiffness, and
140 seventeen questions for the subscale of function). Answer options are presented on a 5-
141 point Likert scale. The total possible score ranges from 0 to 96; the fewer points scored,
142 the better the patient's HrQOL [22,23]. Lastly, to characterize the severity of hip OA, the
143 Harris Hip Score (HHS) was applied by one of the examiners (rater A) to evaluate four
144 domains: Pain (0-44 points), function (0-47 points), absence of deformity (0 or 4 points),

145 and mobility (0-5 points). Scores range from 0 to 100, with higher scores demonstrating
146 less compromised hip joints [24–26].

147 **Procedures**

148 The HHD Lafayette Manual Muscle Testing System Model-01165 (Lafayette
149 Instrument Company, Lafayette IN, USA) was used to assess hip muscle strength during
150 a three-second maximal effort, following the protocol sequence: hip flexors (seated
151 position), hip abductors, and adductors (supine, long-lever), and hip extensors (prone,
152 long-lever) performed on a regular examination table and collected on the same day by
153 both raters. This time frame was chosen considering the clinical context, that individuals
154 with symptomatic hip OA and older adults have difficulties in moving around, which
155 could affect adherence to a second day of evaluation. We assumed that patients would be
156 more interested in participating in the study protocol if measurements were taken on the
157 same day as their regular medical evaluation. In addition, our protocol aimed to mimic
158 the clinical routine of physicians and physiotherapists when evaluating their patients,
159 reproducing a more realistic scenario to be adopted in practice [27,28].

160 Participant and rater positions have been described elsewhere [11], with some
161 minor modifications. Hip flexors were evaluated with the participant on an examination
162 table, seated with both legs hanging off the table and arms positioned at the sides of the
163 body, and both knees and hips at 90°. The assessor was placed right in front of the affected
164 lower limb, holding the HHD with both hands at the anterior aspect of the thigh, 1 to 2
165 cm above the superior edge of the patella. Participants were instructed to push against the
166 HHD, trying to flex the hip with the maximal force for three seconds. Hip abductors were
167 tested in the supine position, hands crossed in front of the chest, hip and knee at 0°, with
168 the assessor standing by the side of the examination table and holding the HHD with both

169 hands above the lateral malleolus (long-lever), using their own body to stabilize it.
170 Similarly, the participant tried to abduct the affected hip against the HHD. Hip adductors
171 were evaluated with the participant in the same position, but now, with the HHD held
172 above the medial malleolus (long-lever) and the examiner placing their knee in the middle
173 of both participant's ankles. In this situation, the participant was encouraged to adduct
174 only the affected leg. Finally, the participant was instructed to lie in the prone position to
175 evaluate hip extensors, arms crossed under the forehead, hip and knee at 0°. The rater
176 stood immediately in front of the end of the table, holding the device with both hands,
177 elbows extended, at 3-4 cm above the posterior calcaneal tuberosity (long-lever),
178 followed by an attempt to extend the hip while maintaining knee at full extension. All
179 participants were advised not to flex the knee during hip extension.

180 Before every protocol sequence of muscle strength assessment, participants were
181 instructed to push against the HHD with their maximum force for three seconds and were
182 reminded that the test starts as they push the HHD and hear a single sound alarm and
183 finishes as they hear a double sound alarm. A submaximal strength test trial was
184 performed in the seated position with the non-affected limb to familiarize the participant
185 with how the device works and the sound alarm. One demonstration was also performed
186 in the supine and prone position to clarify how the test could be performed if required
187 [10,11]. None of the participants had any previous familiarity with this device.

188 Two independent raters performed data collection, both physicians (V.G.F.;
189 F.F.F) with no experience with the HHD. Raters were allowed to practice the
190 measurements protocol sequence for four months. Data were registered using REDCap
191 (Research Electronic Data Capture) electronic data capture tools hosted at Instituto
192 Hospital de Base [29,30]. Each rater repeated measurements twice on the same day. To
193 minimize any possible effect of cumulative pain resulting from test-retest, the order of

194 data collection was defined using a randomized sequence generated on the website
195 sealedenvelope.com (proportion of 1:1, in blocks of four). Participants were allowed to
196 rest between each protocol sequence until they felt comfortable to start the next round
197 [14]. The VAS for pain was measured after each sequence. Participants were given
198 continuous encouragement to push harder against the HHD to obtain maximal isometric
199 force during the 3 seconds of each test [11,14].

200 **Statistical Analysis**

201 Descriptive statistics were used to describe participants' sociodemographic
202 characteristics. The Shapiro-Wilk test was performed to confirm the normal distribution
203 of the data. The Paired t-test was used to compare VAS for pain intensity between intra-
204 and inter-rater measures. Assessment of intra- and inter-rater reliability regarding Pk and
205 Af measures was conducted using the ANOVA 2-way mixed model, with a Confidence
206 Interval of 95% (95%CI), to compare test-retest measures for intra-rater analysis, and the
207 mean of test-retest for inter-rater analysis. To categorize the reliability between repeated
208 measures, we assessed the intraclass correlation coefficient (ICC) and the correlation
209 between measures was classified as poor ($ICC < 0.5$), moderate ($0.5 \geq ICC < 0.75$), good
210 ($0.75 \geq ICC < 0.90$), and excellent ($ICC \geq 0.90$) [11,31]. To define the presence of bias
211 in the data and establish the Limit of agreement (LoA) between raters, mean values
212 considering the two measures were plotted with a 95% CI using the Bland-Altman (BA)
213 method [32,33]. Absolute reliability was evaluated by calculating the Standard Error of
214 Measurement (SEM) and percentage of values (SEM%), and Minimal Detectable Change
215 (MDC) and percentage of values (MDC %) for a 90% CI were calculated considering the
216 following equation: $SEM = \left(\sqrt{\frac{SS_{total}}{n-1}} \right) \times \sqrt{(1-ICC)}$ and $MDC = [z\ score(90\% CI)]$

217 $x \text{ SEM } x \sqrt{2}$ [34,35]. Statistical significance was assumed when $P < 0.05$. All statistical
218 analyses were performed using SPSS version 25 (IBM Corp., Chicago, IL, USA), and the
219 BA graphs were plotted by GraphPad Software (San Diego, CA, USA). The sample size
220 was calculated using an acceptable ICC of 0.70, an expected ICC of 0.90, and assuming
221 an α of 5% and power of 80%, with a drop-out rate of 10%, resulting in a minimal sample
222 of 20 participants [36].

223

224 **Results**

225 The demographic data of the sample are shown in table 1. Approximately 85% of
226 the participants presented a defined joint space reduction associated with sclerosis and
227 moderate to severe osteophytes (types III/IV), representing the whole spectrum of
228 substantial alterations in the x-ray related to osteoarthritis. Considering all daily activities
229 during the week before inclusion in the study, the pain intensity (VAS; mean \pm SD) was
230 8.05 ± 1.2 , 95%CI {7.47 – 8.62}, and together with an HHS score of 50.2 ± 20.1 and a
231 WOMAC score of 63.5 ± 14.0 , the data show considerable pain, dysfunction, and a
232 reduction in quality-of-life related to hip OA.

233

234 **Table 1- Characteristics of participants.**

Characteristic	Sample (n=20)
Age, mean (SD), y	58.7 ± 15.28
Sex (%)	
Male	12 (60)
Female	8 (40)

235	Radiographic disease severity (%)	
236	KL II	3 (15)
237	KL III	6 (30)
238	KL IV	11 (55)
239	Symptoms (%)	
240	6m – 1y	1 (5)
241	1y – 2y	9 (45)
242	2y – 5y	5 (25)
243	> 5y	5 (25)
244	Body mass index, mean	28.82 ± 4.23
245	(SD)	
246	VAS (0 -10), mean (SD)	8.05 ± 1.23
247		
248	HHS (0 -100), mean (SD)	50.2 ± 20.1
249		
250	WOMAC (0 -96), mean	63.5 ± 14.0
251	(SD)	
252	<hr/>	

253

254

255 Table 1 Legend: SD: standard deviation; y: year; m: month; KL: Kellgren and Lawrance
 256 classification; VAS: visual analogue scale; HHS: Harris Hip Score; WOMAC: Western
 257 Ontario and McMaster Universities.

258

259 A difference in VAS (mean±SD, 95%CI) was observed for pain intensity after test
 260 and retest for rater A (Test: 6.11±2.96, {4.68 – 7.36}; Retest: 6.74±2.94, {5.32 – 8.15};
 261 $p=0.01$) that was not observed for rater B (Test: 6.42±2.52, {5.20 – 7.73}; 6.74±2.74,
 262 {5.73 – 8.04}; $p=0.49$), or between raters when considering the mean VAS for the pain
 263 intensity after two measures (A test-retest: 6.55 ± 2.96, {5.12 – 7.98}; B test-retest: 6.73
 264 ± 2.39, {5.58 – 7.89}; $p= 0.54$). Although there was an existing difference in VAS for
 265 pain intensity after the rater A test compared to the retest, it did not seem to have a relevant

266 effect on intra-rater ICC, since rater A presented better ICC and lower SEM values than
267 rater B.

268 Table 2 shows the mean \pm SD values of test-retest Pk and Af, relative reliability
269 expressed as ICC, absolute reliability expressed as SEM, and MDC for the four major hip
270 muscle groups, comparing intra and inter-rater reliability.

271 **Table 2- Handheld dynamometer reliability analysis for hip muscle groups.**

Hip muscle group	Measure	Intra-rater A					Intra-rater B					Interrater				
		Test (mean±SD)	Retest (mean±SD)	ICC (95% CI)	SEM (SEM%)	MDC ₉₀ (MDC%)	Test (mean±SD)	Retest (mean±SD)	ICC (95% CI)	SEM (SEM%)	MDC ₉₀ (MDC%)	Rater A (mean±SD)	Rater B (mean±SD)	ICC (95% CI)	SEM (SEM%)	MDC ₉₀ (MDC%)
Flexors	Pk	13.11±6.00	13.32±6.20	0.931 ^a (0.822-0.974)	0.58 (4.36)	1.33 (10.10)	11.93±3.76	12.67±6.32	0.851 ^a (0.612-0.942)	1.04 (8.49)	2.42 (19.68)	13.22±5.90	12.30±4.85	0.966 ^a (0.912-0.987)	0.28 (2.22)	0.66 ^b (5.16)
	Af	10.83±5.08	10.86±5.04	0.939 ^a (0.841-0.976)	0.42 (3.91)	0.98 ^b (9.07)	9.54±2.95	10.24±4.90	0.761 ^a (0.380-0.908)	1.25 (12.68)	2.91 (29.40)	10.85±4.91	9.89±3.64	0.935 ^a (0.832-0.975)	0.42 (4.08)	0.98 ^b (9.46)
Abductors	Pk	7.52±4.09	8.15±4.13	0.974 ^a (0.932-0.990)	0.17 (2.12)	0.39 ^b (4.92)	7.87±4.84	8.21±4.49	0.927 ^a (0.812-0.972)	0.47 (5.85)	1.09 (13.52)	7.83±4.06	8.04±4.51	0.971 ^a (0.924-0.989)	0.18 (2.22)	0.41 ^b (5.15)
	Af	5.97±2.99	6.45±3.15	0.968 ^a (0.917-0.988)	0.15 (2.42)	0.35 ^b (5.60)	6.42±4.01	6.58±3.48	0.932 ^a (0.822-0.974)	0.35 (5.41)	0.82 (12.56)	6.21±3.02	6.50±3.63	0.913 ^a (0.774-0.967)	0.40 (6.28)	0.93 (14.57)
Adductors	Pk	9.31±5.05	10.91±5.69	0.975 ^a (0.935-0.990)	0.26 (2.60)	0.61 ^b (6.03)	9.16±4.94	10.87±6.13	0.930 ^a (0.818-0.973)	0.57 (5.69)	1.32 (13.20)	10.11±5.32	9.97±5.33	0.982 ^a (0.952-0.993)	0.14 (1.36)	0.32 ^b (3.15)
	Af	7.08±3.71	8.31±4.08	0.957 ^a (0.888-0.983)	0.30 (3.85)	0.69 ^b (8.94)	6.95±3.72	8.38±4.47	0.945 ^a (0.854-0.980)	0.43 (5.55)	0.99 (12.87)	7.69±3.82	7.67±3.97	0.983 ^a (0.955-0.993)	0.09 (1.22)	0.22 ^b (2.84)
Extensors	Pk	9.32±4.75	9.83±4.83	0.924 ^a (0.797-0.972)	0.51 (5.28)	1.17 (12.26)	8.64±4.38	9.97±4.88	0.940 ^a (0.839-0.977)	0.45 (4.83)	1.04 (11.21)	9.58±4.62	9.31±4.50	0.973 ^a (0.929-0.990)	0.17 (1.84)	0.40 ^b (4.26)
	Af	7.38±3.64	7.92±3.76	0.920 ^a (0.786-0.970)	0.42 (5.46)	0.97 (12.66)	6.57±3.11	7.77±3.57	0.942 ^a (0.844-0.978)	0.34 (4.76)	0.79 (11.03)	7.65±3.56	7.17±3.25	0.957 ^a (0.885-0.984)	0.22 (2.91)	0.50 ^b (6.75)

272

273 Table 2 Legend: Pk: Peak Force (Kgf); Af: Average Force (Kgf); SD: Standard Deviation; ICC: Intra-class Correlation Coefficient; 95% CI: 95%

274 Confidence Interval; SEM: Standard Error of Measurement; MDC₉₀: Minimal Detectable Change (90% CI); ^aGood/excellent ICC (≥0.75); ^bMDC₉₀

275 < 10%.

276 The HHD reliability analysis demonstrated a high to very high ICC for test-retest
277 reliability. All rater A measurements presented an excellent correlation in the test-retest
278 analysis, considering both peak force (Pk) and average peak force (Af), while rater B
279 presented a good ICC for flexors Pk (ICC= 0.851; 95%CI {0.612 - 0.942}) and flexors
280 Af (ICC= 0.761; 95%CI {0.380 - 0.908}), and excellent correlations for abductors,
281 adductors, and extensors.

282 The SEM ranged from 0.15 to 0.58Kgf (kilogram-force) for rater A and 0.34 to
283 1.25kgf for rater B, with rater A being more consistent in the test-retest measurements of
284 Pk and Af for flexor, abductor, and adductor hip muscles. In addition, rater A obtained
285 smaller values of MDC when considering all flexor, abductor, and adductor muscles for
286 Pk and Af measures. This difference between raters was more pronounced in the flexors
287 muscle group, which presented the highest mean values of strength for both raters in the
288 test-retest measurements.

289 Nevertheless, when we consider the mean of the two measures in the inter-rater
290 analysis of relative reliability, all ICCs for both Pk and Af were classified as excellent
291 (≥ 0.90) with good precision, expressed by the 95% CI; the smallest value was found for
292 Abductor Af (ICC = 0,913; 95%CI{0,774 - 0,967}) and the highest value for Adductor
293 Af (ICC = 0,983; 95%CI {0,955 - 0,993}). The absolute reliability found for Pk ranged
294 from 0.14 to 0.28kgf, and for Af, it ranged from 0.09 to 0.42kgf, with better consistency
295 for adductor, followed by extensor, abductor, and flexor muscle groups for both measures.
296 These results of MDC% (90%CI) were smaller than 10% for all Pk measures analyzed,
297 which may reflect a satisfactory parameter when comparing the mean of two measures
298 between different raters.

299 The Bland-Altman plot (Fig 1) shows the distribution of the differences in mean
300 values between raters (A-B) versus the mean of all measures. The differences were well

301 distributed for abductor, adductor, and extensor muscle groups, demonstrated by the low
302 bias for Pk and Af, with the lowest tendency of disagreement for hip adductors (Pk bias=
303 0.10 {LoA -2.69 to 2.90}, Fig1e and Af bias = 0.02 {LoA -1.97 to 2.03}, Fig1f), followed
304 by hip abductors (Pk bias = - 0.2 {LoA -3.05 to 2.63}, Fig1c and Af bias = -0.28 {LoA -
305 3.99 to 3.41}, Fig1d), and hip extensors (Pk bias= -0.51 {LoA -5.49 to 4.47}, Fig1g and
306 Af bias = 0.47 {LoA -2.23 to 3.19}, Fig1h). The regression line did not show a statistically
307 significant difference for systematic error for those muscle groups. On the other hand, hip
308 flexor bias demonstrated that differences in measures for rater A for Pk were, on average
309 0.91 Kgf higher than for rater B (Pk bias = 0.91 {LoA -2.93 to 4.73}, Fig1a); and the
310 differences in measures for Af were on average 0.95kgf higher than rater B (Af bias =
311 0.95 {LoA -3.22 to 5.13}, Fig1b). These higher values seem to be related to a tendency
312 of rater A to measure higher values, with increased mean flexor strength when compared
313 to rater B, with a significant deviation from zero for Pk ($p=0.01$) and Af ($p=0.01$) in the
314 positive direction.

315 **Fig 1: Bland-Altman plots comparing the average of all measures against the**
316 **differences between the average measures (rater A-B).** Each black dot represents the
317 average of all measures (Kgf) of one individual. Dashed red lines represent the Limit of
318 Agreement (LoA) of 95% and the continuous red line represents the bias. Flex: flexors;
319 ABD: abductors; AD: adductors; Ext: extensors; Pk: peak force; Af: Average peak force.

320

321 Discussion

322 To our knowledge, this is the first study to assess the use of an HHD in a clinical
323 population with symptomatic hip osteoarthritis, considering the degree of radiographic
324 impairment and pain related to the disease. Our study was designed to reproduce a clinical
325 situation where repeated strength measures could be collected easily in a viable routine

326 rather than a laboratory study design. We demonstrated that the Lafayette HHD is a
327 reliable instrument to evaluate hip muscle strength in this population, with good to
328 excellent intra- and inter-rater reliability, satisfactory consistency, and minimal
329 differences in the intra-rater and inter-rater analyses. Thus, clinicians can use the HHD to
330 evaluate disuse or treatment effects on muscle strength in symptomatic hip OA patients.

331 Previous studies demonstrated that considering the lower limb musculature, the
332 hip presented the strongest validity and reliability for measures of peak force, comparing
333 the same HHD and a fixed dynamometer. Excellent reliability was also found when
334 comparing the HHD applied by a rater or a belt system. Nevertheless, both these studies
335 evaluated healthy and active subjects, and the authors suggest caution with generalization
336 for the clinical population [11,37]. Only one study assessed the HHD reliability for lower
337 limb strength in older individuals (over 65 years old), including participants with hip and
338 knee OA, and demonstrating good intra- and inter-rater reliability for hip and knee muscle
339 strength assessments [10]. However, only ~60% of the participants included in that study
340 have hip or knee OA, and the descriptions of the pain and source of symptoms were poorly
341 characterized, which makes comparisons between our results and those of Arnold and
342 colleagues [10] difficult.

343 Interestingly, the present study demonstrated that the participants present good
344 tolerance for the time taken to perform the measurements (3 seconds), even when pain
345 was also perceived. Collectively, these data also corroborate previous results concerning
346 older adults [10], suggesting that even when the articular disease is present in the lower
347 limb, notably hip OA, the reliability of the HHD is satisfactory to recommend this
348 instrument as a tool for clinical assessment. We also provide adequate information about
349 the characteristics of the participants' hip OA, making it clear how much pain,
350 dysfunction, and reduction in quality of life could be associated with the disease, in order

351 to define more precisely the population of interest in this study. Despite the participants
352 experiencing pain when performing the test protocol, the HHD test demonstrated good to
353 excellent ICC, raising the question of the interference of patient discomfort as a potential
354 limitation to performing tests with enough reliability, as suggested in the literature [12].

355 Rater A had a better correlation between test-rest measures when compared to
356 rater B for all muscle group measurements for Pk and Af, notably in the flexors group.
357 These results may be explained by the difference in anthropometric measurements of the
358 raters and their presumed strength (1.80m and 85kg versus 1.69m and 68kg),
359 demonstrated previously in the literature as a factor that could influence HHD
360 measurements [1,15,38]. It is possible the use of a stabilization belt system, particularly
361 for hip flexors, could help solve this problem, given that it does not depend on the
362 examiner's strength [15,39]. However, there are conflicting data in the literature
363 regarding the advantage of belt stabilization for HHD, since this device does not provide
364 a stabilization belt [37]. Adaptations to stabilize the device and the lack of a proper
365 method of fixation could interfere with measurements and should be further tested and
366 validated before any recommendations are made.

367 The most reliable muscle strength measurement was found for hip adductors,
368 followed by extensors and adductors, demonstrated by excellent values of ICC and an
369 adequate 95% CI, ranging from good to excellent reliability values. An exception was
370 observed for intra-rater B reliability, who, despite showing good ICCs for Pk (ICC=
371 0.851, 95% CI {0.612 – 0.942}) and Af (ICC= 0.761, 95% CI {0.318 – 0.908}), presented
372 a wide range of 95% CI, that could be explained by the stronger participants who had
373 larger differences between test-retest for both raters. This result agrees with Kelln and
374 colleagues (2008), who demonstrated that stronger muscles present wider differences in
375 test-retest evaluations. Our data also suggest that the muscle strength assessment would

376 be more feasible in situations with muscle weakness [11,39,40], expressed by the low
377 SEM values in the inter-rater analysis.

378 The MDC (90% CI) calculated in the intra-rater analysis was smaller for rater A
379 than for rater B. When the mean of two measures was considered, values for both Pk and
380 Af were lower than 10% (except for Af in the abductors group), which is considered an
381 adequate parameter to express any real difference instead of a random error of
382 measurement, according to Prentice et al (2004, quoted in Chamorro et al, 2017). These
383 values suggest that the protocol of measurement with the HHD tested seems to be reliable
384 for clinical purposes since it can detect small variations that could be attributed to a real
385 clinical change. Although MDC has been considered worthwhile to screen patient
386 progression with good precision, future studies should consider economic evaluations of
387 screening strategies concerning HHD assessment, with many specific challenges to
388 overcome [41].

389 The Bland-Altman inter-rater analysis demonstrated small values of bias for
390 abductors, adductors, and extensors when considering the mean of the test and retest.
391 There was reasonable agreement with low bias for both variables, Pk and Af, for all
392 muscle groups evaluated, with a tendency to systematic error only for flexors when
393 comparing raters. However, the LoA demonstrated a large range, especially for flexors
394 and extensors. Future studies should evaluate the influence of experience and routine
395 practice on the LoA range when using this device.

396 Some limitations should be addressed in our study. We did not perform measures
397 on different days and in different positions, so the conclusions raised here should be
398 restricted to conditions that replicate this protocol and compared with caution when
399 considering studies performed in a different setting. With respect to raters, the experience
400 level of both raters was the same; the inclusion of raters with different levels of expertise

401 and practice with this instrument would reflect a more realistic scenario. Furthermore, the
402 sample size did not allow further analysis of the subgroup related to hip osteoarthritis
403 classification, and the relation between radiographic impairment and HHD reliability may
404 not be inferred in our results. Future studies are needed to evaluate the reliability of the
405 HHD in other clinical situations, such as knee osteoarthritis.

406

407 **Conclusion**

408 The HHD is a reliable method to evaluate hip muscle strength in individuals with
409 symptomatic hip OA, with good to excellent intra- and inter-rater reliability and low
410 values of SEM, even in the presence of pain related to the disease. The mean of two
411 measures provides values with satisfactory agreement and reliability between raters, with
412 adequate precision in an easily applied protocol. This study also provided values for the
413 MDC, which could help to define a threshold to quantify improvements or reductions in
414 hip muscle strength during treatment interventions or evaluation of disease progression
415 with a low-cost, portable, and useful tool that requires little training for routine patient
416 care assessment.

417

418 **Acknowledgments:** We thank all participants and collaborators involved in this study.

419

420 **References:**

421

422 [1] Jordan JM, Helmick CG, Renner JB, Luta G, Dragomir AD, Woodard J, et al.
423 Prevalence of hip symptoms and radiographic and symptomatic hip osteoarthritis

- 424 in African Americans and Caucasians: the Johnston County Osteoarthritis Project.
425 J Rheumatol 2009; 36:809–15. <https://doi.org/10.3899/jrheum.080677>.
- 426 [2] Blanco FJ, Silva-Díaz M, Quevedo Vila V, Seoane-Mato D, Pérez Ruiz F, Juan-
427 Mas A, et al. Prevalence of symptomatic osteoarthritis in Spain: EPISER2016
428 study. Reumatol Clin 2021; 17:461–70.
429 <https://doi.org/10.1016/j.reumae.2020.01.005>.
- 430 [3] Loureiro A, Mills PM, Barrett RS. Muscle weakness in hip osteoarthritis: a
431 systematic review. Arthritis Care Res (Hoboken) 2013; 65:340–52.
432 <https://doi.org/10.1002/acr.21806>.
- 433 [4] Loureiro A, Constantinou M, Diamond LE, Beck B, Barrett R. Individuals with
434 mild-to-moderate hip osteoarthritis have lower limb muscle strength and volume
435 deficits. BMC Musculoskelet Disord 2018; 19:303.
436 <https://doi.org/10.1186/s12891-018-2230-4>.
- 437 [5] Zacharias A, Pizzari T, English DJ, Kapakoulakis T, Green RA. Hip abductor
438 muscle volume in hip osteoarthritis and matched controls. Osteoarthritis Cartilage
439 2016; 24:1727–35. <https://doi.org/10.1016/j.joca.2016.05.002>.
- 440 [6] Kolasinski SL, Neogi T, Hochberg MC, Oatis C, Guyatt G, Block J, et al. 2019
441 American College of Rheumatology/Arthritis Foundation Guideline for the
442 Management of Osteoarthritis of the Hand, Hip, and Knee. Arthritis Care Res
443 (Hoboken) 2020; 72:149–62. <https://doi.org/10.1002/acr.24131>.
- 444 [7] Fransen M, McConnell S, Reichenbach S. Exercise for osteoarthritis of the hip
445 (Review). Cochrane Database of Systematic Reviews 2014.
446 <https://doi.org/10.1002/14651858.CD007912.pub2>.

- 447 [8] Kannus P. Isokinetic Evaluation of Muscular Performance. *Int J Sports Med* 1994;
448 15:S11–8. <https://doi.org/10.1055/s-2007-1021104>.
- 449 [9] Krause DA, Neuger MD, Lambert KA, Johnson AE, DeViny HA, Hollman JH.
450 Effects of examiner strength on reliability of hip-strength testing using a handheld
451 dynamometer. *J Sport Rehabil* 2014; 23:56–64. [https://doi.org/10.1123/jsr.2012-](https://doi.org/10.1123/jsr.2012-0070)
452 0070.
- 453 [10] Arnold CM, Warkentin KD, Chilibeck PD, Magnus CRA. The reliability and
454 validity of handheld dynamometry for the measurement of lower-extremity muscle
455 strength in older adults. *J Strength Cond Res* 2010; 24:815–24.
456 <https://doi.org/10.1519/JSC.0b013e3181aa36b8>.
- 457 [11] Mentiplay BF, Perraton LG, Bower KJ, Adair B, Pua Y-H, Williams GP, et al.
458 Assessment of Lower Limb Muscle Strength and Power Using Hand-Held and
459 Fixed Dynamometry: A Reliability and Validity Study. *PLoS One* 2015;
460 10:e0140822. <https://doi.org/10.1371/journal.pone.0140822>.
- 461 [12] Kelln BM, McKeon PO, Gontkof LM, Hertel J. Hand-held dynamometry:
462 reliability of lower extremity muscle testing in healthy, physically active, young
463 adults. *J Sport Rehabil* 2008; 17:160–70. <https://doi.org/10.1123/jsr.17.2.160>.
- 464 [13] Oliveira GDS, Ribeiro-Alvares JB de A, de Lima-E-Silva FX, Rodrigues R, Vaz
465 MA, Baroni BM. Reliability of a Clinical Test for Measuring Eccentric Knee
466 Flexor Strength Using a Handheld Dynamometer. *J Sport Rehabil* 2022; 31:115–
467 9. <https://doi.org/10.1123/jsr.2020-0014>.
- 468 [14] Sisto SA, Dyson-Hudson T. Dynamometry testing in spinal cord injury. *J Rehabil*
469 *Res Dev* 2007; 44:123–36. <https://doi.org/10.1682/jrrd.2005.11.0172>.

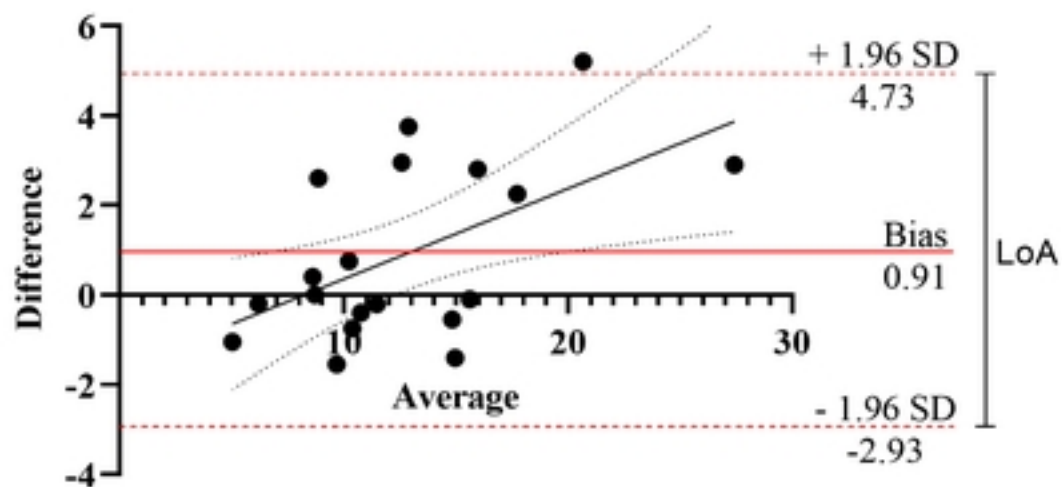
- 470 [15] Ieiri A, Tushima E, Ishida K, Inoue M, Kanno T, Masuda T. Reliability of
471 measurements of hip abduction strength obtained with a hand-held dynamometer.
472 *Physiother Theory Pract* 2015; 31:146–52.
473 <https://doi.org/10.3109/09593985.2014.960539>.
- 474 [16] Kim S-G, Lee Y-S. The intra- and inter-rater reliabilities of lower extremity muscle
475 strength assessment of healthy adults using a handheld dynamometer. *J Phys Ther*
476 *Sci* 2015; 27:1799–801. <https://doi.org/10.1589/jpts.27.1799>.
- 477 [17] Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, et al.
478 Guidelines for reporting reliability and agreement studies (GRRAS) were
479 proposed. *J Clin Epidemiol* 2011; 64:96–106.
480 <https://doi.org/10.1016/j.jclinepi.2010.03.002>.
- 481 [18] Kohn MD, Sassoon AA, Fernando ND. Classifications in Brief: Kellgren-
482 Lawrence Classification of Osteoarthritis. *Clin Orthop Relat Res* 2016; 474:1886–
483 93. <https://doi.org/10.1007/s11999-016-4732-4>.
- 484 [19] Kellgren JH, Lawrence JS. Radiological assessment of osteo-arthritis. *Ann*
485 *Rheum Dis* 1957; 16:494–502. <https://doi.org/10.1136/ard.16.4.494>.
- 486 [20] Leão MG de S, Martins Neta GP, Coutinho LI, da Silva TM, Ferreira YMC, Dias
487 WRV. Análise comparativa da dor em pacientes submetidos à artroplastia total do
488 joelho em relação aos níveis pressóricos do torniquete pneumático. *Rev Bras Ortop*
489 (Sao Paulo) 2016; 51:672–9. <https://doi.org/10.1016/j.rbo.2016.02.002>.
- 490 [21] Wong DL BCM. Pain in children: comparison of assessment scales. *Pediatr Nurs*
491 1988; 14:9–17.

- 492 [22] Woolacott NF, Corbett MS, Rice SJC. The use and reporting of WOMAC in the
493 assessment of the benefit of physical therapies for the pain of osteoarthritis of the
494 knee: findings from a systematic review of clinical trials. *Rheumatology (Oxford)*
495 2012; 51:1440–6. <https://doi.org/10.1093/rheumatology/kes043>.
- 496 [23] Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. Validation
497 study of WOMAC: a health status instrument for measuring clinically important
498 patient relevant outcomes to antirheumatic drug therapy in patients with
499 osteoarthritis of the hip or knee. *J Rheumatol* 1988; 15:1833–40.
- 500 [24] Lane NE, Hochberg MC, Nevitt MC, Simon LS, Nelson AE, Doherty M, et al.
501 OARSI Clinical Trials Recommendations: Design and conduct of clinical trials for
502 hip osteoarthritis. *Osteoarthritis Cartilage* 2015; 23:761–71.
503 <https://doi.org/10.1016/j.joca.2015.03.006>.
- 504 [25] Guimarães RP, Alves DPL, Silva GB, Bittar ST, Ono NK, Honda E, et al. Tradução
505 e adaptação transcultural do instrumento de avaliação do quadril “Harris Hip
506 Score.” *Acta Ortop Bras* 2010; 18:142–7. [https://doi.org/10.1590/S1413-](https://doi.org/10.1590/S1413-78522010000300005)
507 [78522010000300005](https://doi.org/10.1590/S1413-78522010000300005).
- 508 [26] Harris WH. Traumatic arthritis of the hip after dislocation and acetabular fractures:
509 treatment by mold arthroplasty. An end-result study using a new method of result
510 evaluation. *J Bone Joint Surg Am* 1969; 51:737–55.
- 511 [27] Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J,
512 et al. Quality criteria were proposed for measurement properties of health status
513 questionnaires. *J Clin Epidemiol* 2007; 60:34–42.
514 <https://doi.org/10.1016/j.jclinepi.2006.03.012>.

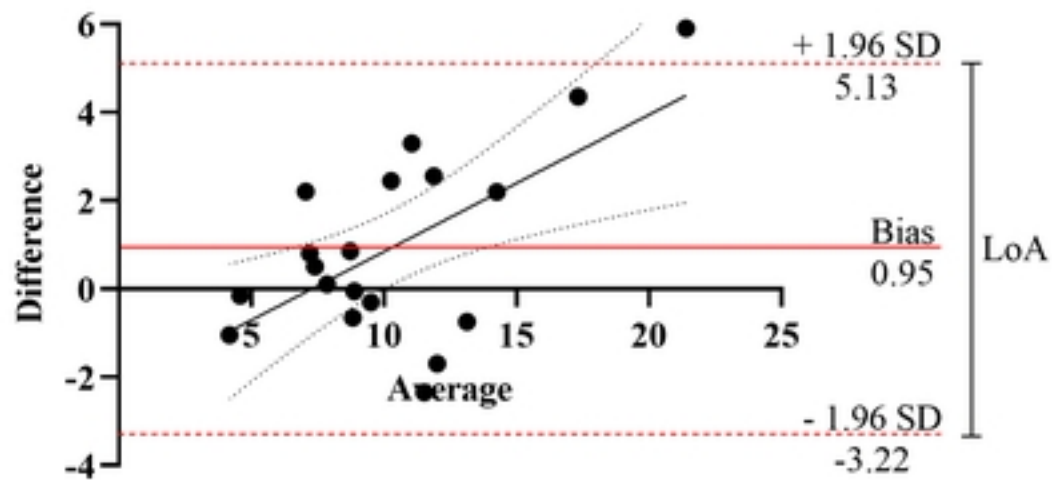
- 515 [28] Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, et al.
516 Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were
517 proposed. *J Clin Epidemiol* 2011; 64:96–106.
518 <https://doi.org/10.1016/j.jclinepi.2010.03.002>.
- 519 [29] Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O’Neal L, et al. The
520 REDCap consortium: Building an international community of software platform
521 partners. *J Biomed Inform* 2019;95. <https://doi.org/10.1016/j.jbi.2019.103208>.
- 522 [30] Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research
523 electronic data capture (REDCap) - a metadata-driven methodology and workflow
524 process for providing translational research informatics support. *J Biomed Inform*
525 2009; 42:377–81. <https://doi.org/10.1016/j.jbi.2008.08.010>.
- 526 [31] Portney, Leslie Gross MPW. *Foundations of Clinical Research: Applications to*
527 *Practice*. 3rd ed. Upper Saddle River, New Jersey: Pearson/Prentice Hall; 2009.
- 528 [32] Giavarina D. Understanding Bland Altman analysis. *Biochem Med (Zagreb)* 2015;
529 25:141–51. <https://doi.org/10.11613/BM.2015.015>.
- 530 [33] Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat*
531 *Methods Med Res* 1999; 8:135–60.
532 <https://doi.org/10.1177/096228029900800204>.
- 533 [34] Haley SM, Fragala-Pinkham MA. Interpreting change scores of tests and measures
534 used in physical therapy. *Phys Ther* 2006; 86:735–43.
- 535 [35] Cardoso JR, Beisheim EH, Horne JR, Sions JM. Test-Retest Reliability of
536 Dynamic Balance Performance-Based Measures Among Adults With a Unilateral

- 537 Lower-Limb Amputation. PM R 2019; 11:243–51.
538 <https://doi.org/10.1016/j.pmrj.2018.07.005>.
- 539 [36] Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability
540 studies. Stat Med 1998; 17:101–10. [https://doi.org/10.1002/\(SICI\)1097-
541 0258\(19980115\)17:1<101::AID-SIM727>3.0.CO;2-E](https://doi.org/10.1002/(SICI)1097-0258(19980115)17:1<101::AID-SIM727>3.0.CO;2-E).
- 542 [37] Florencio LL, Martins J, da Silva MRB, da Silva JR, Bellizzi GL, Bevilaqua-
543 Grossi D. Knee and hip strength measurements obtained by a hand-held
544 dynamometer stabilized by a belt and an examiner demonstrate parallel reliability
545 but not agreement. Phys Ther Sport 2019; 38:115–22.
546 <https://doi.org/10.1016/j.ptsp.2019.04.011>.
- 547 [38] Wikholm JB, Bohannon RW. Hand-held Dynamometer Measurements: Tester
548 Strength Makes a Difference. Journal of Orthopaedic & Sports Physical Therapy
549 1991; 13:191–8. <https://doi.org/10.2519/jospt.1991.13.4.191>.
- 550 [39] Bohannon RW. Hand-held dynamometry: A practicable alternative for obtaining
551 objective measures of muscle strength. Isokinet Exerc Sci 2012; 20:301–15.
552 <https://doi.org/10.3233/IES-2012-0476>.
- 553 [40] Brinkmann JR. Comparison of a hand-held and fixed dynamometer in measuring
554 strength of patients with neuromuscular disease. J Orthop Sports Phys Ther 1994;
555 19:100–4. <https://doi.org/10.2519/jospt.1994.19.2.100>.
- 556 [41] Irigorri N, Spackman E. Assessing the value of screening tools: reviewing the
557 challenges and opportunities of cost-effectiveness analysis. Public Health Rev
558 2018; 39:17. <https://doi.org/10.1186/s40985-018-0093-8>.

(a) Difference vs. average: Bland-Altman of Flex- Pk

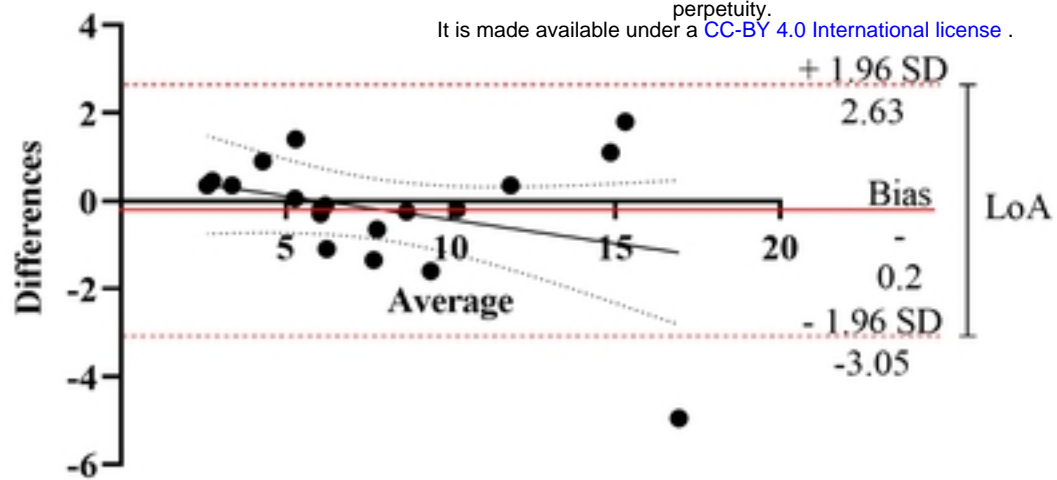


(b) Difference vs. average: Bland-Altman of Flex - Af

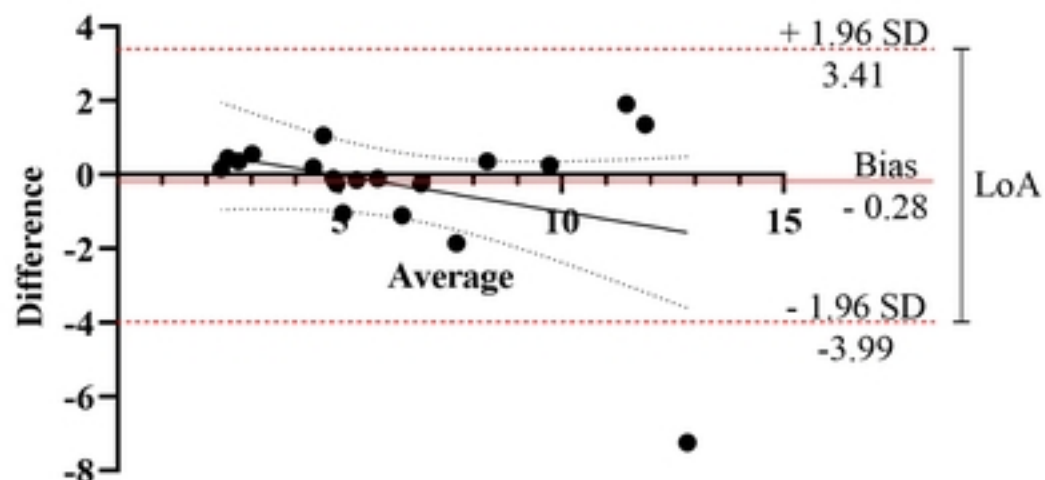


(c) Difference vs. average: Bland-Altman of ABD - Pk

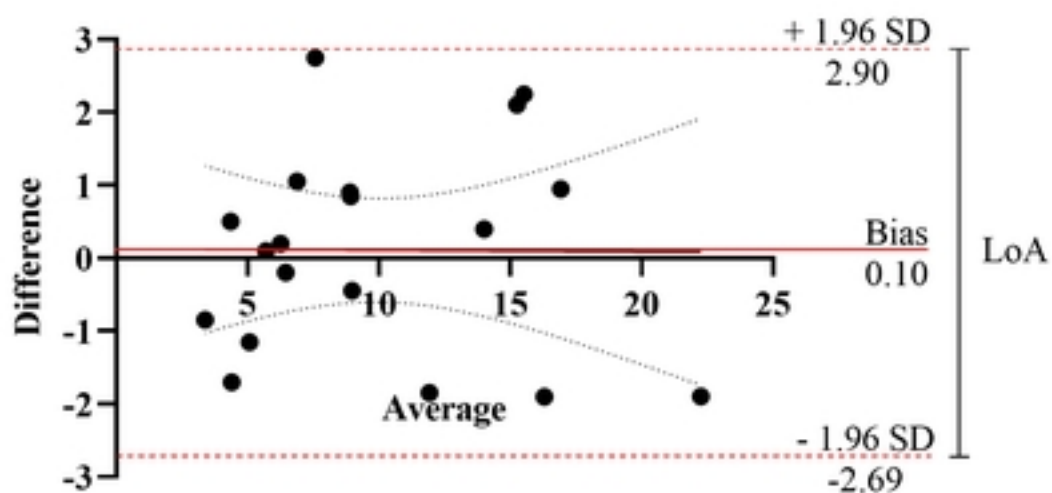
medRxiv preprint doi: <https://doi.org/10.1101/2022.11.10.22282186>; this version posted November 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY 4.0 International license](https://creativecommons.org/licenses/by/4.0/).



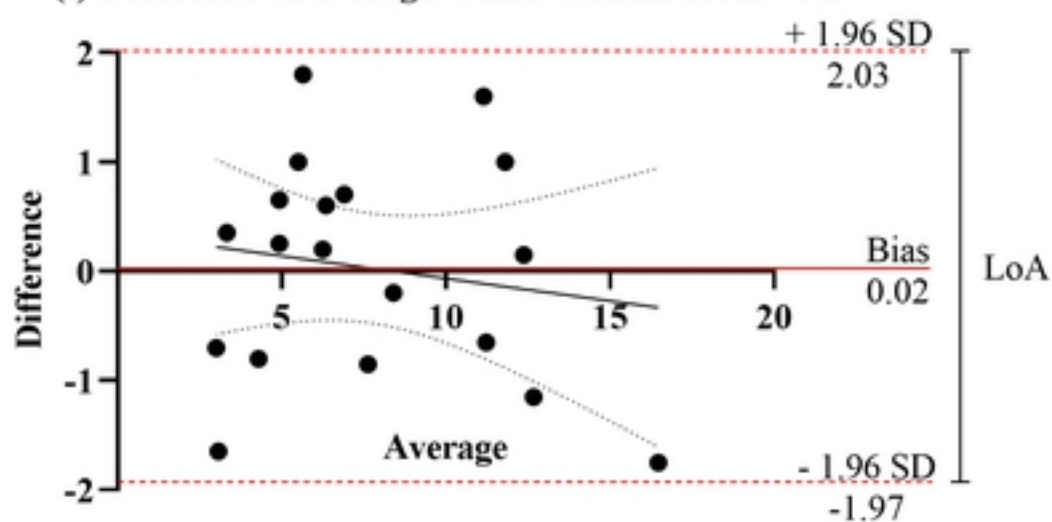
(d) Difference vs. average: Bland-Altman of ABD - Af



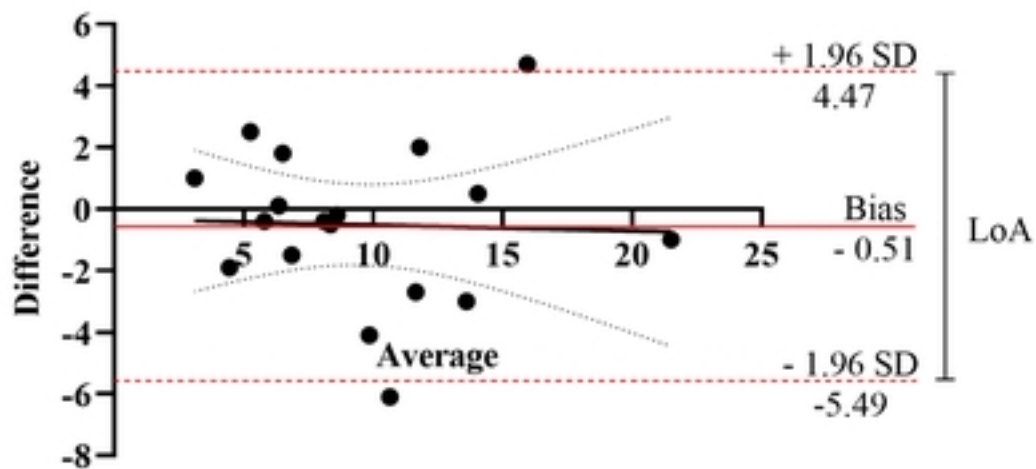
(e) Difference vs. average: Bland-Altman of AD - Pk



(f) Difference vs. average: Bland-Altman of AD - Af



(g) Difference vs. average: Bland-Altman of Ext- Pk



(h) Difference vs. average: Bland-Altman of Ext - Af

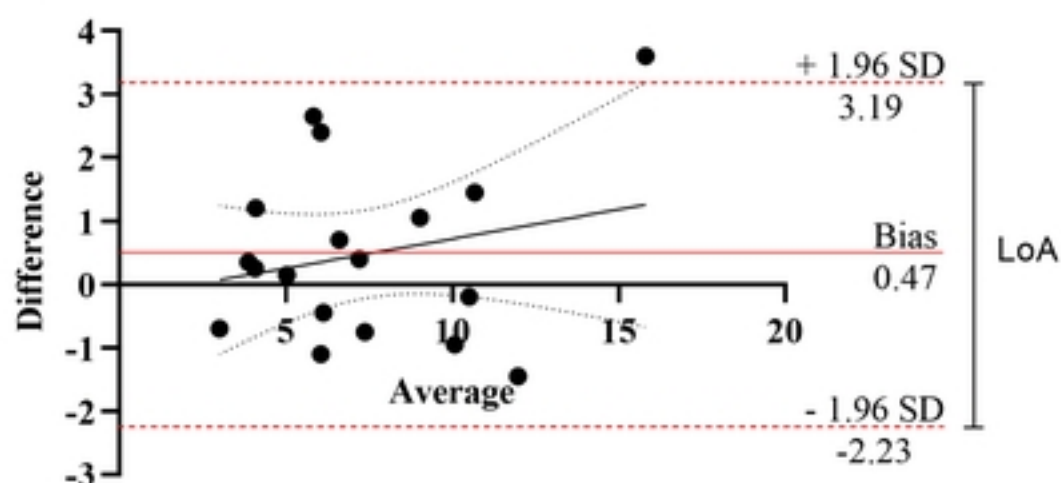


Figure 1