TITLE:
Assessing physical fitness during pregnancy: validity and reliability of fitness tests, and relationship with maternal and neonatal health-related outcomes. A systematic review.

AUTHORS:

Romero-Gallardo, L.¹-²; Roldan-Reoyo, O*³-⁴; Castro-Piñero, J*⁵-⁶.; May, L⁷-⁸; Ocón, O.⁹-¹⁰, Aparicio, V.A¹¹-²; Soriano-Maldonado, A¹²-¹³.

¹PA HELP “Physical Activity for Health Promotion, CTS 1018” Research Group, Department of Physical Education and Sport, Faculty of Sport Sciences, University of Granada, Granada, Spain

²Sport and Health University Research Centre, University of Granada, Spain.

³School of Sport Science, College of Engineering, Swansea University, Wales.

⁴Applied Sports Technology Exercise and Medicine Research Centre, Swansea University, Wales.

⁵GALENO Research Group, Department of Physical Education, Faculty of Education Sciences, University of Cádiz, Puerto Real, Spain.

⁶Biomedical Research and Innovation Institute of Cádiz (INIBICA), Cádiz, Spain.

⁷Department of Foundational Science and Research, Department of Kinesiology, East Carolina University, NC, US.

⁸Department of Obstetrics & Gynecology, East Carolina University, NC, US.

⁹Gynaecology and Obstetrics Unit, ‘San Cecilio’ University Hospital, Granada, Spain.

¹⁰Biohealth Research Institute in Granada (ibs.GRANADA), Spain

¹¹Department of Physiology, Institute of Nutrition and Food Technology and Biomedical Research Centre, University of Granada, Granada, Spain.

¹²Department of Education, Faculty of Education Sciences, University of Almería, Almería, Spain

¹³SPORT Research Group (CTS-1024), CERNEP Research Center, University of Almería, Almería, Spain

* Both authors have contributed equally.

CORRESPONDENCE AUTHOR: lidiaromerogallardo@gmail.com

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
ORCID NUMBERS

Romero-Gallardo, L.: ORCID: 0000-0001-6341-855X

Roldan-Reoyo, O: ORCID: 0000-0003-2625-4060

Castro-Piñero, J: ORCID: 0000-0002-7353-0382

Aparicio, V.A: ORCID: 0000-0002-2867-378X

May, L: ORCID: 0000-0002-8231-2280

Ocón, O: ORCID: 0000-0002-8113-3651

Soriano-Maldonado, A: ORCID: 0000-0002-4626-420X
Abstract

Background: Physical fitness (PF) is a powerful marker of health throughout the lifespan. In pregnant women, higher PF is associated with better maternal and fetal health, better delivery outcomes and earlier postpartum recovery. The assessment of PF during pregnancy requires special considerations to preserve fetal and maternal health; thus, providing a compilation of the most frequently used fitness tests, and assessing their validity, reliability, and association with maternal and neonatal health-related outcomes is of scientific and clinical interest.

Objectives: To systematically review studies evaluating one or more components of PF in pregnant women, to answer two research questions: 1) What fitness tests have been previously employed in pregnant women? and 2) What is the validity and reliability of these tests and their relationship with health-related outcomes? Data Sources: PubMed and Web of Science. Methods: Two independent reviewers systematically examined the articles in each database. The information from the included articles was summarized by a single researcher.

Results: A total of 149 articles containing a sum of 191 fitness tests were included. Among the 191 fitness tests, 99 (i.e., 52%) assessed cardiorespiratory fitness through 75 different protocols, 28 (15%) assessed muscular fitness through 16 different protocols, 14 (7%) assessed flexibility through 13 different protocols, 45 (24%) assessed balance through 40 different protocols, 2 assessed speed with the same protocol and 3 were multidimensional tests using one protocol. A total of 19 articles with 23 tests (13%) assessed either validity (n=4), reliability (n=6), or the relationship of PF with maternal and neonatal health-related outcomes (n=16). Conclusion: PF during pregnancy has been assessed through a wide variety of protocols, mostly lacking validity and reliability data.

PROSPERO Registration Number: CRD42018117554

Key Words: Physical fitness, pregnant women, exercise test, reliability of results, test-retest, field test, laboratory test.
Key points:

• Physical Fitness during pregnancy has been assessed through a wide variety of tests (n=191) and different protocols (n=149).
• We found that cardiorespiratory fitness has been assessed with 75 different protocols (many of them made ad hoc), muscular fitness with 16 different protocols, flexibility with 13 different protocols, and balance with 40 different protocols.
• Most of the protocols lacked validity and reliability data, which limits the confidence on the association of fitness with maternal and neonatal health-related outcomes, which is still scarce.
• We advocate for an expert consensus to be developed in the following years to achieve the goal of appropriate and effective PF assessment during pregnancy.

Running Heading:

Assessing physical fitness during pregnancy: validity, reliability and relationship with maternal and neonatal health-related outcomes.

1. Introduction

Physical Fitness has been defined as the ability to carry out daily tasks with vigor and alertness, without undue fatigue and with ample energy to enjoy leisure-time pursuits and meet unforeseen emergencies [1,2]. Physical fitness (PF) is considered a powerful marker of health that is associated with a lower risk of cardiovascular events, cancer and all-cause mortality in all ages [3–7]. In pregnant women, some studies have recently underlined the potential impact of PF on maternal and fetal health [8–15]. Low PF levels are associated with low infant birth weight [8], increased risk of gestational diabetes mellitus [9,10], poor postpartum recovery [11] and worse delivery outcomes [12,13]. Moreover, the anatomical, biomechanical, physiological, and psychological changes during pregnancy might compromise PF levels [16–18]. Consequently, it is of clinical and public health interest to assess PF during pregnancy, and to understand what the best available tests to assess PF during this critical period of women’s life are.

Two categories of PF components have been defined: 1) Health-related components (cardiorespiratory fitness (CRF), muscular fitness, muscular endurance, and flexibility) 2) Skill-related components (ability, coordination, balance, power, reaction time and speed)
These PF components can be assessed subjectively through questionnaires [15], objectively and accurately through laboratory tests, and efficiently, economically and easily through field-based tests. During pregnancy, a wide variety of fitness tests have been used to assess PF; although a compilation of these tests has not been published to date. Compiling all fitness tests performed in pregnant women would help practitioners to select the most useful test according to their purpose. It is also important to note that, although laboratory tests are generally the gold standard for assessing PF, these tests are not generally accessible to everyone because they need sophisticated and expensive equipment, and it is not possible to evaluate a relatively large sample in a short period of time. As an alternative, a number of field test exist that provide an opportunity to assess PF in a more accessible way [2]. However, there is no consensus on which fitness tests should be used to assess PF in pregnant women, and the validity and reliability of many of the tests used to assess PF during pregnancy is unknown [19].

Since the assessment of PF in pregnant women requires special considerations to preserve fetal and maternal health [18,20,21], understanding which fitness tests are valid, reliable, and associated with maternal and neonatal health-related outcomes, would provide a framework for improving PF assessment during pregnancy and also for improving exercise prescription in this population.

The aims of this systematic review were to 1) describe which fitness tests have been used to evaluate PF in pregnant women and, 2) to evaluate the validity and reliability of the fitness tests, and their relationship with maternal and neonatal health-related outcomes.

2. Methods

2.1. Registration and Elaboration

This systematic review was prospectively registered at PROSPERO (CRD42018117554; available at http://www.t.ly/fS6a). In addition, the review followed the PRISMA explanation and elaboration [22] and the PRISMA Checklist [23] is included as supplementary material. See electronic supplementary material 1 (ESM 1).

2.2. Search Strategy
Articles were searched by two independent reviewers from two major databases, MEDLINE (PubMed) and the Web of Science (WOS) from inception until January 2021. The complete search strategy and further details are presented in ESM 2 (tables S1 and S2).

2.3. Inclusion Criteria

The inclusion criteria were: 1) Healthy pregnant women (no restriction regarding gestational week), 2) At least one component of PF assessed either through field-based or laboratory tests, 3) Access to full text, 4) Only one original article from the same study/project using the same test were included, and 5) Text in English or Spanish.

2.4. Quality assessment of the articles

To assess the quality of the articles included for aim 2, we used three quality scores that are comprehensively described in ESM 3 and ESM 4.

2.5. Process and data extraction

After checking title and abstract, only the studies meeting all inclusion criteria were introduced in a reference manager software (Mendeley). In the event of disagreement between the two independent reviewers concerning the inclusion/exclusion of an article, a consensus was reached (there was no need of a third person). The snowball strategy was also used. Information including reference, age, sample size and fitness test description are summarized in ESM 5-Table S6.

3. Results

A comprehensive PRISMA flow diagram is presented in Figure 1.

3.1. Overall results, quality assessment and gestational week.

We identified 2617 studies, of which 149 were included (Figure 1). These articles contained 191 fitness tests, using 149 different protocols that were included for aim 1. A comprehensive scheme of the fitness tests and the different protocols performed to date, divided by PF component, is presented in figure 2. A summary of the number and percentage of articles that assessed PF during pregnancy and protocols used for its assessment, divided by PF components, is presented in table 1.

Regarding aim 1, 99 tests (including 75 different protocols) were used to assess cardiorespiratory fitness, 28 (including 16 different protocols) to assess muscular fitness, 14
(including 13 different protocols) to assess flexibility, 45 tests (including 40 different protocols) to assess balance, 2 tests using the same protocol to assess speed, and 3 tests using the same protocol were multidimensional. No results were found for other PF components such as agility or coordination.

Regarding aim 2, a total of 19 articles (13% of the total number of articles included) assessed at least validity (n=3), reliability (n=4) of fitness tests, or the relationship of PF with maternal and neonatal health-related outcomes (n=16). Of these 16 articles, 11 were classified as very low quality [13,28–36] and 5 were classified as low quality [8,33,37–39] (table 2). Of the 3 articles [40–42] that assessed validity, 2 articles were classified as low quality [40,41] and 1 as high quality [42]. Of the 4 articles that assessed reliability criteria, 3 were considered high quality [40,43,44] and 1 low quality [45] (table 2).

The gestational week at PF assessment ranged from 8 to 41 across articles. Some articles assessed PF at different time points throughout pregnancy; therefore, we divided pregnancy into two stages. Early pregnancy (i.e., from week 0 to week 20 of gestation) and late pregnancy (i.e., from week 21 to week 40). Using this approach, 11 articles (7%) were performed in early pregnancy; 57 articles (38%) were performed in late pregnancy; 55 articles (37%) were performed several times (i.e., range 2 to 5 times) throughout pregnancy; 7 (5%) articles specified a range of weeks that included early to late pregnancy; 14 articles (9%) reported only the trimester without specifying gestational week; 4 articles (3%) provided no information, and 1 article (1%) assessed PF on the day of labor.

3.2. Cardiorespiratory fitness

3.2.1. Tests used

We identified 99 tests assessing CRF, of which 61 (62%) were performed on cycle-ergometer, 25 (25%) on treadmill, 10 (10%) on track and there were 3 (3%) step protocols (figure 2). Of the 99 tests, a total of 75 corresponded to different protocols. For instance, there were 56 different protocols using cycle-ergometer, distributed as follows: only one article used the Arstila test [29]; one used the Bruce Protocol at 75% HRmax [46]; one applied the Modified Bruce ramp protocol at anaerobic threshold [47]; two employed the Modified Balke protocol at 70% HRmax [48,49]; one used YMCA protocol [50]; and the remaining articles (n=55) used ad-hoc tests (i.e. specifically designed for the purpose of the investigation); 11 of which
[28,49,51–57] used steady-state tests and 44 [29,30,62–71,37,72–81,47,82–91,49,92–96,50,58–61] used incremental tests. When analyzing the type of test based on intensity, we found that 13 articles used maximal tests [47,63,93,94,97,69,70,72,73,75,84,85,90], 37 submaximal tests [29,30,59,61,62,65–68,71,76,78,37,79–83,87–89,91,95,50,96,98,99,51,53,54,56–58] and 3 used mixed tests [49,60,64] containing submaximal and maximal stages within the same protocol.

There were 25 treadmill tests that used 14 different protocols, distributed as follows: the Modified Balke protocol was used in 10 articles [8,42,63,100–106]; the Modified Bruce protocol in 2 articles [13,107]; and the traditional Balke protocol -twice in the same article-[108]; the traditional Bruce protocol [109], the Cornell protocol [40], the SWET protocol and the Ebbelling single-stage protocol [18] were each used in 1 article. There were other 7 ad hoc tests of which 2 were steady-state [54,110], and 5 were incremental tests [58,111–114]. According to intensity, 3 were maximal tests [110,112,113] and 4 submaximal tests [54,58,111,114].

Of the 10 tests on track, 6 articles performed the 6-minute walk test protocol [115–120], and 4 were ad hoc tests (i.e. 1 maximal and 4 were submaximal). In regards to the 3 step tests, 1 Canadian Home Fitness test [121] was used and 2 ad hoc incremental submaximal tests were used [31,122].

3.2.2. Validity and reliability

We identified 2 articles examining validity [40,42]. Yeo et al. [40] aimed to validate a portable metabolic testing system (VO2000) on healthy sedentary pregnant women. The VO2000 consistently overestimated VO2 measurements, compared to the same manufacturer’s reference system, by 4.4±3.6 standard deviation (SD) ml/kg/min although the Pearson correlation was significant (r=0.48; p=0.01). When VO2000 was used twice, the mean difference was statistically significant (1.0±1.8 ml/kg/min; t(45)=3.9, p<0.001). Mottola et al. [42] provided a prediction equation for VO2peak in pregnant women between 16 and 22 weeks of gestation, using a modified Balke protocol. The results of this equation revealed an adjusted R2 of 0.71 (p value not reported). When the authors used this equation to predict VO2peak in a cross-validation group (n=39), they found a predicted value of 23.38±4.03 mL×kg⁻¹×min⁻¹, while the actual value was 23.54 ± 5.9 mL×kg⁻¹×min⁻¹ (p value not reported).

3.2.3. Relationship with maternal and neonatal health-related outcomes
A total of 6 articles analyzed the association of CRF with maternal and neonatal health-related outcomes. Pomerance et al. [28] observed that VO$_{2\text{max}}$ was inversely associated with the length of labor in multiparas ($r=-0.65; p=0.001$) and pre-pregnancy weight ($r=-0.63; p=0.001$). However, VO$_{2\text{max}}$ was neither correlated with newborn weight, length, or head circumference, nor with the one-minute Apgar score (all $p>0.05$). In the same line, Wong & McKenzie [30], observed that fit mothers showed lower HR at submaximal exercise intensity ($p<0.05$) and the second stage of labor was shorter (no statistics reported) compared to unfit pregnant mothers. However, there was no difference between fit and unfit in the length of gestation or weight gained (no statistics reported). In the same article, the authors showed neither positive nor negative effect of maternal fitness on newborn weight or Apgar scores.

In addition, Erkkola et al. [29] found that newborn from fit pregnant women had higher pH than fetuses of less physically fit women ($p<0.01$). In this article, participants with low physical performance were more likely to have asphyxiated neonates than neonates of physically fit women ($p<0.05$). In the same line, Baena-García et al. [13] observed that maternal CRF at the 16th gestational week was related to higher arterial umbilical cord PO$_2$ ($r=0.267, p<0.05$), and women who had caesarean sections had significantly lower CRF compared with the women who had vaginal births ($p<0.001$).

On the other hand, Bisson et al. [8] studied the association of CRF in early pregnancy with physical activity before and during early pregnancy. The authors found that a higher VO$_{2\text{peak}}$ in early pregnancy was positively associated with physical activity spent at sports and exercise before and during early pregnancy ($p<0.001$).

3.3. Muscular fitness
3.3.1. Tests used

A total of 28 tests (i.e., 14% of all included articles) that included 16 different protocols assessed muscular fitness, of which 10 performed maximal hand-grip strength tests [8,12,128,13,32,33,123–127], 3 performed endurance hand-grip test, 2 for 3-min [129,130] and 1 for 2-min period [131] (figure 2). In 2 of the articles conducting an endurance hand-grip test [129,131], a hand grip sphygmomanometer was used instead of dynamometry. On the other hand, 1 used hand-held dynamometer fixed to a chair to assess quadriceps strength [132] and 1 used toe-grip dynamometer [132]. Moreover, 2 ad hoc isometric tests were used to assess maximal voluntary hip extension and back flexors endurance in the same article.
Finally, 13 dynamic endurance tests were found, 9 were listed as ad hoc tests [12,126,134] and other 3 (30-sec Chair Stand Test, 5 Times Sit to Stand test, Trendelenburg’s test) were classified as others dynamic tests [13,44,126].

3.3.2. Validity and reliability

Only 2 muscular fitness tests assessed reliability [44,45]. Yenisehir et al. [44] analyzed reliability and validity of Five Times Sit-to-Stand. Inter-rater reliability was excellent for subjects with and without pelvic girdle pain (ICC = 0.999, 95% CI = 0.999–1.000; ICC = 0.999, 95% CI = 0.999–0.999, respectively). Test-retest reliability was also very high for subjects with and without pelvic girdle pain (PGP) (ICC = 0.986, 95% CI = 0.959–0.995; ICC = 0.828, 95% CI = 0.632–0.920, respectively).

On the one hand, Gutke et al.[45] analyzed the reliability for an ad hoc test. This test consisted in a maximal voluntary isometric hip extension with a fixed sensor holding a sling around the thigh and pulling for 5 seconds during 3 reps with 5-10 seconds of rest (r=0.82 for the right leg and r=0.88 for the left leg; ICC=0.87 for the right leg and 0.85 for the left leg; with p value no reported).

3.3.3. Relationship with maternal and neonatal health-related outcomes

Bisson et al.[8] observed that hand-grip strength was positively associated with infant birth weight (r=0.34, p=0.0068) even after adjustment for confounders (r=0.27, p=0.0480). Zelazniewicz et al. [33] observed that hand-grip strength was associated with offspring birth weight when controlled for the newborn sex and gestational age at delivery (F(2.182)=3.15; p=0.04). Baena-García et al. [13] found greater hand-grip strength weakly associated with greater neonatal birth weight (r = 0.191, p<0.05). Wickboldt [32] found that hand-grip strength was moderately correlated with pain scores, where the mean hand-grip strength during contractions had the highest correlation coefficient (r=0.67; p<0.001) compared with peak handgrip strength (r=0.56; p<0.001) and the area under the curve of handgrip force (r=0.55; p<0.001).

3.4. Flexibility

3.4.1. Tests used

Our search identified 14 (7%) tests that assessed flexibility using 13 different protocols, including the side bending test [135], the sit-and-reach test [12], the back-scratch test (twice)
[13,125], the Motion Analysis (i.e. including 3 different tests such as the seated and standing forward flexion, seated and standing side to side flexion and seated axial rotation [136]) and optoelectrical system (i.e. performing 4 different tests) [137]. Goniometry was used in two different articles to measure isquiosural flexibility, [138] wrist flexion-extension and medial lateral deviation [139]. Only one article used an *ad hoc* machine to test passive abduction of the left fourth finger [34].

### 3.4.2. Validity and reliability

Lindgren et al.[34] designed an *ad hoc* machine to test passive abduction of the left fourth finger and its relationship with low-back pain during pregnancy and early postpartum. Abduction angle was measured at three different times throughout the pregnancy and once in the postnatal period. Reliability of the abduction angle was analyzed by the intra-individual coefficient of variance. The coefficients of variance between the first and second measurement was 0.077, between the second and third 0.070 and between the third and fourth 0.071.

### 3.4.3. Relationship with maternal and neonatal health-related outcomes

Only 2 flexibility tests evaluated associations with maternal and neonatal health-related outcomes. Lindgren et al. [34] found that women with greater passive abduction angle of the left fourth finger was associated with highest back pain incidence (OR=1.09; CI=1.01-1.17; *p*=0.022) and the highest number of previous pregnancies (OR=3.24; CI=1.57-6.68; *p*=0.002). Baena-Garcia et al. [13] found increased flexibility associated with a more alkaline arterial pH (*r* = 0.220, *p*<0.05), higher arterial PO2 (*r* = 0.237, *p*<0.05) and lower arterial PCO2 (*r* = -0.331, *p*<0.01) in the umbilical cord blood.

### 3.5. Balance

#### 3.5.1. Tests used

We identified 45 (24%) articles assessing balance of which 19 analyzed static balance and 26 dynamic balance trough 40 different protocols. With regard to *static balance*, 18 were laboratory tests of which 12 assessed balance through stabilometry tests on force platform [35,140,149,150,141–148], one on pressures platform [143] and another on Equitest® platform [151]. Four articles did not mention the type of platform used (40,150,176,177). Regarding protocols, all articles conducted the tests with participants standing with bipedal
support. However, standing position varied between articles. Ten articles maintained a standing posture with feet separated [35,132,145,147–153], 1 with feet together [146], 2 used mixed protocols [38,142], 1 with medial malleoli separated [154] and 4 did not mention the standing posture [140,141,143,144]. Moreover, 3 articles used protocols with eyes open [141,148,150] exclusively, 8 articles used mixed protocols with eyes open and closed, 1 used visual target and visual tasks [144] and 6 did not specify whether participants kept their eyes closed or opened. Only 1 article used a field test, the one-legged standing protocol [125]. On the other hand, 1 test was field-test without platform.

In relation to the 26 articles measuring dynamic balance, 9 of them assessed balance using platforms. Each of these articles used different testing tool such as balance master platform [155], pressure platform [143], force platform [156], Equitest® platform [36] and movable platform, which was used in two articles [157,158]. Two of these articles were walking protocols [143,156], 1 with translational perturbations [159], one was standing with one knee flexed and arms across the chest [157,158]. Another 15 articles used 3-D camera motion capture systems using 13 different protocols. Twelve of the 15 articles were walking protocols [160,161,170–172,162–169] and 2 used a stand to sit motion protocol [173,174]. Moreover, 1 article used a triaxial accelerometer[175]; another article assessed balance through recording (without specification of camera type) [176], and another using instrumented insoles [177]. All three articles used walking protocols.

3.5.2. Validity and reliability

No validity or reliability assessments were performed regarding balance tests.

3.5.3. Relationship with maternal and neonatal health-related outcomes

3 articles associated balance with neonatal and maternal health-related outcomes. Ozturk [38] observed that static balance decreased and fall risk increased in pregnant women with lower back pain (49.90±24.47 vs 28.47±19.60; p<0.0001). In relation to exercise, McCrory et al. [159] showed that exercise may play a role in fall prevention in pregnant women (p=0.005); they also found that dynamic balance is altered in pregnant women who have fallen compared with non-fallers and non-pregnant women (p<0.001). Nagai et al. [35] studied the relationship between anxiety and balance. They concluded that when anxiety increases during pregnancy, the standing posture is destabilized (r=0.559, p=0.020), which may increase the chance of falling.

3.6. Speed
3.6.1. Test used

The only protocol that was used to assess speed during pregnancy was the ten-meters timed walk test (10mTWT). However, the same test was identified in two different articles. [41,43]. In the 10mTWT, the participants commenced standing at a chair. When told to start, subjects walk as fast as possible along 14 meters marked with white blank tapes placed at 0 m, 2 m, 12 m and 14 m. The time (100th of a second) required to walk between the 2 m and 12 m markers was recorded and converted into speed in meters per second (m/sec).

3.6.2. Validity and reliability

Validity and reliability for 10mTWT was studied by Evensen et al. in two different articles [41,43]. In 2015, Evensen al. [43] analyzed the test-retest reliability of 10mTWT showing an intraclass correlation coefficient (ICC) of (0.74). Intertester reliability was determined in the first 13 participants with strong correlation (ICC =0.94). In 2016 [41] the same authors analyzed the convergent validity of 10mTWT by comparing performances with scores achieved on the Active Straight Leg Raise (ASLR) test and observed moderate positive correlations between 10mTWT and ASLR (r=0.65, p=0.003).

3.6.3. Relationship with maternal and neonatal health-related outcomes

This systematic review did not find any articles that analyzed the association of speed with maternal and neonatal health-related outcomes.

3.7. Agility and coordination

No articles of agility and coordination were identified.

3.8. Multidimensional

3.8.1. Test used

Our search identified a walking multidimensional test that was used in three studies [39,41,43]. In Timed Up and Go Test (TUG) the participant begins seated in a chair with their arms on armrests and their toes against a start line. The purpose is to cross the front white line at three meters away, turn around, and walk back to the chair and sit down as fast as possible. The performance is measured in time (100th of a second).

3.8.2. Validity and reliability
Validity and reliability for TUG was analyzed by Evensen et al. in two different studies [41,43]. The TUG showed good test-retest reliability (ICC=0.88) and intertester reliability (ICC=0.95). Regarding reliability, strong correlations were found between the TUG and Active Straight Leg Raise ($r= 0.73, p= 0.001$).

3.8.3. **Relationship with maternal and neonatal health-related outcomes**

The time on TUG among pregnant women with pelvic girdle pain was significantly higher (mean (95% CI) 6.9 (6.5, 7.3) seconds) than for asymptomatic pregnant (5.8 (5.5, 6.0), $p < 0.001$) and non-pregnant (5.5 (5.4, 5.6), $p < 0.001$) women.

4. **Discussion**

4.1. **Summary of the evidence**

This systematic review revealed that PF has been assessed through a wide variety of tests during pregnancy. However, very little is known on the validity and reliability of the tests performed. These findings have important research and clinical implications. First, until a specific battery of fitness assessments for pregnant women is developed and validated, the confidence in PF data during pregnancy is limited and potentially unreliable and may prove harmful if unreliable values are used for exercise recommendation during pregnancy. Second, the large variety of tests used makes it challenging to compare results from different studies. Third, provided the lack of rigorous information on validity and reliability of PF tests, it is also difficult to evaluate the association of PF with maternal and neonatal health-related outcomes, which is of wide clinical and public health interest. However, some studies have attempted to present associations of PF with maternal and neonatal health-related outcomes, which undoubtedly needs to be replicated once a PF test battery is released. Before that, exhaustive research must be performed in validating such battery of tests.

4.1.1. **Cardiorespiratory Fitness**

This systematic review identified that cycle-ergometer has been the equipment most frequently used to assess CRF followed by treadmill and field tests; although step tests have also been conducted. There is a large disparity of protocols and wide variety of ad hoc tests used, which makes comparing results between studies difficult. However, the Modified Balke treadmill Protocol validated by Mottola et al. [42] for pregnant women has been the most frequently used test. There have been more incremental tests used for CRF tests during pregnancy compared to steady-state tests and more submaximal compared to maximal tests. There is no consensus regarding test termination criteria for submaximal tests, which undoubtedly needs further research. Some articles used relative intensity using physiological
variables such as %HR\textsubscript{max} or %VO\textsubscript{2max}, and other used absolute intensity, such as specific HR (beats per minute). Among the studies that used %HR\textsubscript{max} as a test termination criterion, there was a variety of percentages such as 70% [48,58,65,96], 75% [46,61,91,107] or 85% [13,40,81]. Among the studies that used %VO\textsubscript{2max}, there were different percentages such as 40% [54], 50% [53,59], 60% [52,54], or 70% [178]. Among the studies that used absolute HR as a test termination criterion, the HR for finalizing the tests were set either at 125 [86], 150 [30,56,66,87], 155 [122], 160 [89], or 170 [77,78,82,83] beats per minute. Some studies even used the rate perceived exertion as complementary criteria [50,71,78] or peak aerobic power [67]. These complementary criteria have been recommended and studied in pregnant women by authors like Hesse et al. [109] since the physical and emotional changes during pregnancy limit performance. It must be noted that the same equation was not used to estimate HR\textsubscript{max}. Some articles used the traditional 220-age formula [61,65,81,91,107] while others used the Karvonen [40] or Tanaka [96] formulas. Some articles did not specify how HR\textsubscript{max} was estimated [46,48,58]. This heterogeneity could be due to the physiological complexity of the pregnant woman, in terms of cardiac changes and response to exercise and the lack of scientific information in this regard. Moreover, the gestational week could be a determinant for physiological response since Bijl et al. [96] observed a slower hemodynamic recovery and an increased ventilatory response to exercise in early pregnancy compared to non-pregnant women. With regards to the maximal tests, different terms have been used like maximal criteria such as volitional fatigue [64,69,70,72,73,93,94,109,178], exhaustion [63], anaerobic threshold [47,75,112,113], and point of symptom limitation [84,85,95].

This lack of consensus has many drawbacks that need to be resolved in view of the need to accurately assess CRF during pregnancy. We advocate for an expert consensus to be developed in the following years to achieve the goal of appropriate and effective CRF assessment during pregnancy. In particular, it seems essential to develop a treadmill and a cycle ergometer submaximal test that reveals sufficient validity to confidently estimate VO\textsubscript{2max} throughout gestation.

4.1.2. Muscular fitness

Muscular fitness tests included muscular strength, endurance and power [2]. The studies included in this systematic review show that muscular strength was the most frequently assessed component of muscular fitness, since only 6 studies [12,13,44,126,134,179] assessed endurance and none of them assessed power in pregnant women. In most studies, muscular
strength was evaluated through handgrip maximal strength using a dynamometer. However, 2 studies used a handgrip sphygmmomanometer test [129,131]. Some of the handgrip tests were performed in standing position [8,123], while others used sitting position [125] or supine position [127], and others did not reveal the position used for the assessment [33,124,126,138]. Some tests were completed 3 times [126], others twice [8,33,124], and others only once [125,127,138]. This clearly reveals a large methodological variability that might influence the results and make comparing results between studies difficult. Another limitation is the fact that the main strength outcome was handgrip strength. While handgrip strength is a good marker of health [180], it is unclear whether handgrip responds to changes following exercise interventions. Therefore, validating other muscular strength tests, including lower limb strength tests, is needed in order for researchers and practitioners to confidently assess muscular strength during pregnancy.

There were no validity studies and the reliability was assessed only in one maximal isometric hip extension test [45]. This test has limitations since the pregnant abdomen must be on a bed and, as acknowledged by the authors, it cannot be performed during the third trimester. It must be noted that higher handgrip strength was associated with higher birth weight [8,33]. Moreover, increased hand-grip strength was produced during uterine contraction [32]. The advantage of using handgrip is that it represents an inexpensive, rapid, and easy-to-use assessment with minimal training needed to appropriately administer. However, assessing the performance of pregnant athletes with this test seems clearly insufficient. More quality in tests employed is necessary since the association of muscular strength with maternal and neonatal health-related outcomes is of clinical importance. Moreover, other studies are needed to understand the extent to which preserving strength throughout pregnancy and postpartum relates to clinical outcomes.

4.1.3. Flexibility

Although there were 7 studies assessing flexibility, none of them used the same protocol. Once again, this reflects a lack of agreement when assessing the same component of PF. Moreover, Lindgren et al. [34] found that a higher flexibility showed a higher low back pain. Despite the limitation of a finger laxity test, we considered these findings an interesting association that warrants further article since passive stretching is one of the most common practical prescriptions for exercise professionals instead of mobility and breathing exercises. On the other hand, the results of Baena-García et al. [13] are very relevant to foetal health
since the flexibility was associated with a better pH, PO₂ and PCO₂ in umbilical cord blood. Hence, more research about flexibility tests, their outcomes and their prescription are needed.

4.1.4. Balance

We identified that balance was the second PF component most frequently evaluated during pregnancy, following CRF. This makes sense since the center of gravity changes during pregnancy as a result of expansion of the uterus and the risk of falls increases. However, there is a great heterogeneity between the protocols employed in different studies. For static balance, the protocol most frequently used was stabilometry on force platform with bipedal support and eyes open and eyes closed within the same test [35,38,154,181,182,140–144,146,147,151]. For dynamic balance, there was a greater heterogeneity across protocols both in the platform used and in the movements over the platforms. Regarding the assessment tool, the 3-D camera was the device most frequently used [151,160–162,164,173]. Likewise, we observed differences between the number of platform pieces, trials and Hz utilized. Some protocols were performed on 2 piece-platforms [141,147,154], others on 1 piece-platforms [35,140,142,145,146,148,152] and others did not specify the type of platform [143,144,151]. Although the number of trials and the frequency of recording (i.e. Hz) are important parameters that should be carefully described about the protocols, only 5 (out of 13) articles described the number of trials [140,145,147,152,183] and 1 described frequency of recording [141]. The usefulness of these tests are restricted to the research area and all of them use expensive technological tools; therefore, it is difficult to extrapolate these tests to fitness centers or clinical settings. We could prevent falls during pregnancy if we could assess balance easily. For this reason, it is necessary to develop an inexpensive and easy-to-use balance field-test.

4.2. Validity, reliability and maternal and neonatal health-related outcomes

Unfortunately, studies that examine validity and reliability of PF tests are scarce. The physical fitness component most frequently studied was CRF, however we have only found two studies that analyzed the validity of the CRF tests and no studies examined the reliability of these tests. On treadmill platform, Mottola et al. [42], validated a special equation for modified Balke protocol that has been used by numerous other authors. In contrast, Yeo tried to validate a portable metabolic testing system (mod. VO2000) but it overestimated VO₂ measurements for pregnant women compared to non-pregnant women and males.
Regarding muscular fitness, hand grip test was most commonly used; this test was used as the gold standard for muscular fitness during pregnancy. Only Gutke et al. [45], studied the reliability of a test for hip extension. However, the $p$ value was not reported, and the position adopted in the test could be uncomfortable for pregnant participants. Finally, the studies evaluating validity and reliability of speed tests and multidimensional components of PF have been researched by Evensen et al. [41,43]. They demonstrated that TUG and 10mTWT are reliable and valid tests for pregnant women.

The validity and reliability of balance (without tests), agility and/or coordination tests has not been investigated to date.

We suggest that specific tests to be performed in pregnant women are needed and their validity and reliability must be assessed to understand the extent to which one might rely on such measures when prescribing exercise, or making clinical recommendations.

Regarding maternal and neonatal health-related outcomes, we can conclude that more research is also necessary. Nevertheless, from this review we can highlight some interesting associations with different fitness components. A better CRF was associated with a shorter labor [30,98] and a lower risk of cesarean section [13]. However, no association was found regarding other fetal outcomes such as Apgar scores or the newborn anthropometrics [30,98]. By contrast, muscular strength was associated with optimum infant birth weight [8,13,33]. Other neonatal outcomes like fetal umbilical cord pH were positively associated with maternal CRF [29]. On the other hand, better balance scores are associated with decreased fall risk [35,184,185]. These results are very useful for exercise professionals, as it implies that protocols during pregnancy must be implemented with balance exercises. Finally, Evensen et al.[39] found that pelvic girdle pain could be a limiting factor to assess physical fitness in pregnant women since the time of TUG was significantly higher in women with pain than in asymptomatic pregnant and non-pregnant women.

None of the studies reviewed in this article have described adverse events during PF assessment. Moreover, official bodies such as the American College of Obstetrician and Gynaecology or the Canadian Society of Exercise Physiology have highlighted the benefits of an adequate PF assessment, and assert the need of consensus in the PF assessment during pregnancy [186].

4.3. Limitations and Strengths
A limitation of this article is that, although PubMed and WOS are among the most relevant databases in the medical literature, the possibility that a small number of studies have been overlooked cannot be discarded. Nevertheless, these two databases are the biggest databases in sports medicine and sports sciences and, therefore, include the vast majority of studies.

A strength of this systematic review is the fact that, to the best of our knowledge, this is the first article to comprehensively analyze PF assessments, the validity and reliability of fitness tests, and their relationship with maternal and neonatal health-related outcomes during pregnancy. The results from this systematic review provide an overall picture of how PF is being assessed in this population, what type of tests are being performed, their specific characteristics, whether these tests have been tested for validity and/or reliability; and whether PF is associated with maternal and neonatal health-related outcomes. All this information is of wide and undoubted clinical interest.

5. Conclusions

The main finding of this systematic review is that PF has been assessed through a wide variety of protocols, mostly lacking validity and reliability data, and that no consensus exists on the most suitable fitness tests to be performed during pregnancy. In addition, information regarding the association of PF with maternal and neonatal health-related outcomes is scarce and should further evaluated as well. Provided the need to assess PF during pregnancy and the importance not only to understand the physical state of the pregnant women but also to precisely prescribe exercise in this population, extensive research is needed to design and validate a battery of fitness tests to be used for the safe and effective assessment of PF during pregnancy. We advocate for an expert consensus to be developed in the following years to achieve the goal of appropriate and effective PF assessment during pregnancy.

AUTHORS CONTRIBUTIONS

Conceptualization: Romero-Gallardo, L.; Aparicio, V.; Castro-Piñero, J.; Soriano-Maldonado, A.


Formal analysis and investigation: Romero-Gallardo, L.; Roldán-Reoyo, O.; Soriano-Maldonado, A.
Writing - original draft preparation: Romero-Gallardo, L.

Writing - review and editing: Romero-Gallardo, L.; Roldan-Reoyo, O; Castro-Piñero, J.; Aparicio, A.; May, L; Ocón, O., Soriano-Maldonado, A.

Resources: Aparicio, V., Castro-Piñero, J.; Soriano-Maldonado, A.

Supervision: Aparicio, V.; Ocón O.; Castro-Piñedo, J.; Soriano-Maldonado, A.

FUNDING

This study has been partially funded by the University of Granada, Plan Propio de Investigación 2016, Excellence actions: Units of Excellence: Unit of Excellence on Exercise and Health (UCEES), and by the Junta de Andalucía, Consejería de Conocimiento, Investigación y Universidades and European Regional Development Fund (ERDF), ref. SOMM17/6107/UGR. Current research activities of Dr. Alberto Soriano-Maldonado are supported by a grant from the Spanish Ministry of Science, Innovation and Universities (ref. RTI2018-093302-A-I00). This study is of the Doctoral Thesis of Lidia Romero Gallardo, within the Biomedicine Doctoral Program at the University of Granada.

CONFLICT OF INTEREST:

Authors declare that they have no conflict of interest.
REFERENCES


50. Sussman D, Saini BS, Schneiderman JE, Spitzer R, Seed M, Lye SJ, et al. Uterine...


120. Amola M, Pawara S, Kalra S. Effect of Inspiratory Muscle Training and


152. Opala-Berdzik A, Bacik B, Markiewicz A, Cieślińska-Świder J, Swider D, Sobota G,


170. Gimunova M, Zvonar M, Sebera M, Turcinek P, Kolarova K. Special footwear designed for pregnant women and its effect on kinematic gait parameters during pregnancy


Appendix

TABLES:

Table 1. Number (%) of articles that assessed the different components of physical fitness during pregnancy and protocols used for its assessment.

Table 2. Overview of studies that assessed the validity and/or reliability of fitness tests during pregnancy or the association of physical fitness with health-related outcomes (HrO) in pregnant women.

FIGURES:

Figure 1. Flow chart of the literature search and paper selection process.

Figure 2. Scheme of the fitness tests and the different protocols divided by PF component.

Electronic Supplementary Material (ESM)

ESM 1: PRISMA Checklist
ESM 2: Electronic Supplementary Material 2
- Table S1. Search strategy used and number of articles found in Pubmed.
- Table S2. Search strategy used and number of articles found in Web of Science.

ESM 3: Comprehensive description of the three quality assessment scores used in the present systematic review.
ESM 4: Electronic Supplementary Material 4
- Table S3. Quality assessment criteria to evaluate validity and reliability studies.
- Table S4. Quality assessment criteria to evaluate reliability studies.
- Table S5. Quality assessment criteria to evaluate health-related outcomes studies.

ESM 5: Electronic Supplementary Material 5
- Table S6. Overview of studies included in the systematic review and description of physical fitness tests.
<table>
<thead>
<tr>
<th></th>
<th>TOTAL</th>
<th>CRF</th>
<th>MF</th>
<th>Flexibility</th>
<th>Balance</th>
<th>Speed</th>
<th>Multidimensional</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fitness tests</strong></td>
<td>191</td>
<td>99 (52)</td>
<td>28 (15)</td>
<td>14 (7)</td>
<td>45 (24)</td>
<td>2 (1)</td>
<td>3 (1)</td>
</tr>
<tr>
<td><strong>Protocols</strong></td>
<td>149</td>
<td>75 (50)</td>
<td>16 (11)</td>
<td>13 (9)</td>
<td>40 (29)</td>
<td>1 (0.5)</td>
<td>1 (0.5)</td>
</tr>
</tbody>
</table>

CRF: Cardio-respiratory fitness; MF: Muscular Fitness
<table>
<thead>
<tr>
<th>Reference (authors, year)</th>
<th>Validity</th>
<th>Reliability</th>
<th>HrO</th>
<th>Capacity evaluated, short test description and health-related outcomes or statistical results</th>
<th>Quality Score</th>
</tr>
</thead>
</table>
| Pomerance et al., (1974) [98] | No       | No          | Yes | Ad hoc, continuous test on CE.  
  1: PFS was inversely associated with pre-pregnancy weight ($r$=-0.63; $p=0.001$).  
  2: PFS was significantly associated with the length of labor in multiparas ($r$=-0.65; $p=0.05$).  
  3: PFS was not correlated with pre-pregnancy weight; infant birth weight, new-born length, new-born head circumference or one-minute Apgar score (all $p>0.05$). | 2 |
| Erkkola et al., (1976) [29] | No       | No          | Yes | Ad hoc, incremental submaximal test on CE.  
  1: Fetal pH in fit pregnant women is higher than the pH in less physically fit pregnant women ($p<0.01$).  
  2: Pregnant women with low physical performance are more liable to have asphyxiated babies than the physically fit pregnant women ($p<0.05$). | 2 |
  1: HR at submaximal exercise was lower in fit pregnant women than unfit pregnant women ($p<0.05$).  
  2: Second stage of labor was shorter in fit pregnant women; (no statistics reported).  
  3: There were no differences between fit and unfit participants in the length of gestation, weight gained vs pre-pregnancy measures. There were neither positive nor negative effects of maternal fitness on the new-born (Apgar test, birth weight); (no statistics reported). | 1 |
| Thorell et al., (2015) [37] | No       | No          | Yes | Ad hoc, incremental maximal test on CE.  
  Absolute VO_{2peak} was inversely correlated with duration of gestation among women with miscarriage ($r$=-0.52; $p=0.02$) and positively with duration of gestation among women with spontaneous onset of labor ($r=0.12; p=0.01$). | 3 |
<table>
<thead>
<tr>
<th>Study</th>
<th>Active</th>
<th>Inactive</th>
<th>CRF</th>
<th>Method</th>
<th>Vascular Fitness</th>
<th>Muscular Fitness</th>
</tr>
</thead>
</table>
1: Active women have better aerobic fitness (VO$_{2\text{max}}$) as compared to inactive women (34.9 vs 30.3 mL/kg/min; $p=0.01$). |
| Yeo et al., (2005) [40]       | Yes    | Yes      | No  | Cornell Protocol on treadmill platform.  
Validity: Bland-Altman plots. The mean difference was 4.4 ± 3.6 mL/kg/min. Data indicated that VO2000 over-estimates VO2 by an average of 4.4 mL/kg/min compared to CPX/D. Pearson correlation coefficient between the average and difference of paired measurements was close but not significant (r=0.48; $p>0.01$). Reliability: Paired t test ($t(45) = 3.9$, $p < .001$). Linear regression: y = 0.96X-1.6; 95% CI for the slope: 0.94 to 1.1; R2= 0.91, $p<0.001$ |
| Baena-García et al., (2020) [13] | No     | No       | Yes | Modified Bruce protocol until 85% HR$_{\text{max}}$  
1) Maternal CRF at the 16th gestational week was related to higher arterial umbilical cord PO2 (r=0.267, $p<0.05$).  
2) The women who had caesarean sections had lower CRF (p<0.001) at the 16th gestational week compared with the women who had vaginal births. |
| Gutke et al., (2008) [45]     | No     | Yes      | No  | Maximal voluntary isometric hip extension  
Reliability: Spearman’s rho (r) and Intercorrelation coefficient (ICC). Right leg: r=0.82; ICC=0.87. Left leg: r=0.88; ICC=0.85 (both $p$ value no reported). |
| Bisson et al., (2013) [8]     | No     | No       | Yes | Hand Grip Strength was positively associated with infant birth weight (r=0.27; $p=0.048$). |
Wickboldt et al., (2015) [32]  
No  No  Yes  
Hand Grip Strength was moderately correlated with pain scores on the Numerical Rate Scale. Mean handgrip force during contractions had the highest correlation coefficient ($r=0.67; p<0.001$) compared with peak handgrip force ($r=0.56; p<0.001$) and area under the curve of handgrip force ($r=0.55; p<0.001$).

Zelazniewicz et al., (2018) [33]  
No  No  Yes  
Hand Grip Strength in pregnancy was positively associated with offspring birth weight when controlled for a child's sex and week at delivery ($F_{(2,182)} = 3.15, p=0.04$). Women with greater hand grip strength in each trimester were more likely to give birth to a boy ($p<0.05$).

Baena-García et al., (2020) [13]  
No  No  Yes  
Hand Grip Strength was associated with greater neonatal birth weight ($r = 0.191, p<0.05$)

Yenisehir et al., (2020) [44]  
No  Yes  No  
Five Times Sit to Stand Test  
Reliability: Inter-rater reliability of 5TSS was excellent for subjects with and without PGP (ICC ¼ 0.999, 95% CI ¼ 0.999–1.000; ICC ¼ 0.999, 95% CI ¼ 0.999–0.999, respectively). Test-retest reliability of 5TSS was also very high for subjects with and without PGP (ICC ¼ 0.986, 95% CI ¼ 0.959–0.995; ICC ¼ 0.828, 95% CI ¼ 0.632–0.920, respectively).

Flexibility  
Ad hoc passive abduction of the left fourth finger.  
The highest back pain incidence was showed by women with the greatest passive abduction angle of the left fourth finger (OR=1.09; CI=1.01-1.17; $p=0.022$) and the highest number of previous pregnancies (OR=3.24; CI=1.57-6.68; $p=0.002$).  
Reliability: Intra-individual coefficient of variance. 1) Between the first and second measurement = 0.077; 2) Between the second and third = 0.070 and between the third
and fourth = 0.071.

<table>
<thead>
<tr>
<th>Study</th>
<th>Balance</th>
<th>Speed</th>
<th>Effect</th>
<th>p-value</th>
<th>Speed 1</th>
<th>Speed 2</th>
<th>Speed 3</th>
<th>Speed 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baena-García et al., (2020) [13]</td>
<td>No No Yes</td>
<td>Back-Scratch</td>
<td>Maternal flexibility was associated with a more alkaline arterial pH ($r = 0.220$, $p&lt;0.05$), higher arterial PO2 ($r = 0.237$, $p&lt;0.05$) and lower arterial PCO2 ($r = -0.331$, $p&lt;0.01$) in umbilical cord blood.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nagai et al., (2009) [35]</td>
<td>No No Yes</td>
<td>Static balance on force platform.</td>
<td>Increases in anxiety during pregnancy is associated with a destabilized standing posture ($r= 0.559$, $p=0.020$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ozturk et al., (2016) [38]</td>
<td>No No Yes</td>
<td>Static balance on non-specific platform in different positions.</td>
<td>Postural equilibrium decreases and fall risk increases in pregnant patients with lower back pain ($49.90 \pm 24.47$ vs $28.47 \pm 19.60$; $p&lt;0.0001$).</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>McCrory et al., (2010) [159]</td>
<td>No No Yes</td>
<td>Dynamic balance on Equitest posture platform.</td>
<td>Dynamic balance is altered in pregnant women who have fallen compared with non-fallers and controls. ($p&lt;0.001$). Exercise may play a role in fall prevention in pregnant women.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evensen et al., (2015) [43]</td>
<td>No Yes No</td>
<td>Ten-meters Timed walk Test Reliability: ICC from a one-way random effects model and reporting the 95% confidence interval (CI). Coefficients for test-retest reliability for 10mTWT: (ICC=0.74; 95% CI=0.42–0.90; SEM=0.17 m s$^{-1}$; MDC$<em>{95} = 0.47$ m s$^{-1}$) Coefficients for intertester reliability 10mTWT: (ICC=0.94; 95% CI=0.82–0.98; SEM=0.09 m s$^{-1}$; MDC$</em>{95} = 0.25$ m s$^{-1}$).</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evensen et al., (2016) [41]</td>
<td>Yes No No</td>
<td>Ten-meters Timed walk Test Validity: Spearman correlation coefficient. Between the 10mTWT and ASLR ($r = -0.65$, $p=0.003$). Between the 10 mTWT and PGQ ($r = -0.25$ to $-0.56$).</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Multidimensional**

<table>
<thead>
<tr>
<th>Study Authors</th>
<th>Test Type</th>
<th>Relevance</th>
<th>Reliability</th>
<th>Validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evensen et al., (2015) [43]</td>
<td><strong>Timed Up and Go Test (TUG)</strong></td>
<td>No</td>
<td>ICC from a one-way random effects model and reporting the 95% confidence interval (CI). Coefficients for test-retest reliability TUG: (ICC=0.88; 95% CI=0.70–0.95; SEM=0.42 seconds; MDC95=1.16 seconds.) Coefficients for intertester reliability TUG: (ICC= 0.95; 95% CI=0.84–0.98; SEM=0.36 m/s; MDC95= 1.00 m/s).</td>
<td></td>
</tr>
<tr>
<td>Evensen et al., (2016) [41]</td>
<td><strong>Timed Up and Go Test (TUG)</strong></td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Christensen et al., (2019) [39]</td>
<td><strong>Timed Up and Go Test (TUG)</strong></td>
<td>No</td>
<td>The time on TUG among pregnant women with PGP was significantly higher (mean (95% CI) 6.9 (6.5, 7.3) seconds) than for asymptomatic pregnant (5.8 (5.5, 6.0), p &lt; 0.001) and non-pregnant (5.5 (5.4, 5.6), p &lt; 0.001) women.</td>
<td></td>
</tr>
</tbody>
</table>

HrO: Health related outcome; CRF: Cardio-respiratory fitness; MS: Muscular Strength; PFS: Physical Fitness Score; HR: Heart Rate; PGP: Pelvic girdle pain; ICC: intraclass correlation coefficient (relative reliability); 95% CI: 95% confidence interval; SEM: standard error of measurement (absolute reliability); MDC95, minimal detectable change at 95% CI; OR: odds ratio; TUG: Timed Up and Go test; 10mTWT: Ten-meter Timed Walk Test.
**PRISMA 2009 Flow Diagram**

**Figure 1.** Flow chart of the literature search and paper selection process.

1. **Identification**
   - Records identified through PubMed database searching (n = 930)
   - Records identified through WoS database searching (n = 1687)
   - Papers identified in the search (n = 2617)
   - Excluded duplicated articles (n = 565)
   - Records excluded after checking title and abstract (n = 1862)

2. **Screening**
   - Records screened (n = 2052)
   - Full-text articles excluded, with reasons (n = 40)
     1. Physical fitness not assessed or insufficiently described (n = 20)
     2. Not original article (n = 11)
     3. Fitness not assessed during pregnancy (n = 4)
     4. Identical data derived from the same research project (n = 5)

3. **Eligibility**
   - Full-text articles assessed for eligibility (n = 189)
   - Articles included to address aim 1: Physical fitness have been used to evaluate fitness (n = 149)
   - Articles included to address aim 2: validity, reliability and health-related outcomes (n = 19*)
   - Physical fitness tests included in 149 articles: (n = 191)
   - Physical fitness tests included in 19 articles: (n = 23)


*For more information, visit www.prisma-statement.org.*