Assessment of physical fitness during pregnancy: validity and reliability of fitness tests, and relationship with maternal and neonatal health – a systematic review

Lidia Romero-Gallardo 1,2, Olga Roldan Reoyo 3,4, Jose Castro-Piñero 5,6, Linda E May 7,8, Olga Ocón-Hernández 9,10, Michelle F Mottola 11, Virginia A Aparicio 12,13,14, Alberto Soriano-Maldonado 13,14

ABSTRACT
Objectives To systematically review studies evaluating one or more components of physical fitness (PF) in pregnant women, to answer two research questions: (1) What tests have been employed to assess PF in pregnant women? and (2) What is the validity and reliability of these tests and their relationship with maternal and neonatal health?

Design A systematic review.

Data sources PubMed and Web of Science.

Eligibility criteria Original English or Spanish full-text articles in a group of healthy pregnant women which at least one component of PF was assessed (field based or laboratory tests).

Results A total of 149 articles containing a sum of 191 fitness tests were included. Among the 191 fitness tests, 99 (ie, 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 (n=16).

Conclusion Physical fitness has been assessed through a wide variety of protocols, mostly lacking validity and reliability data, and no consensus exists on the most suitable fitness tests to be performed during pregnancy.

PROSPERO registration number CRD42018117554.

BACKGROUND
Physical fitness (PF) has been defined as the ability to carry out daily tasks with vigour and alertness, without undue fatigue and with ample energy to enjoy leisure-time pursuits and meet unforeseen emergencies.1,2 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 individuals, some studies have recently highlighted 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 recovery11 and...
worse delivery outcomes.\textsuperscript{12, 13} Moreover, the anatomical, biomechanical, physiological and psychological changes during the pregnancy might compromise PF levels.\textsuperscript{16–18} Consequently, it is of clinical and public health interest to assess PF during the pregnancy, and to understand which available tests are best to assess PF during this critical period of life.

Two categories of PF components have been defined as follows: (1) health-related components (cardiorespiratory fitness (CRF), muscular fitness, muscular endurance and flexibility) and (2) skill-related components (ability, coordination, balance, power, reaction time and speed).\textsuperscript{12} These PF components can be assessed subjectively through questionnaires,\textsuperscript{15} objectively and accurately through laboratory tests and efficiently, economically and easily through field-based tests. During the 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 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 tests exist that provide an opportunity to assess PF in a more accessible way.\textsuperscript{2} However, there is no consensus on which fitness tests should be used to assess PF in pregnant individuals, and the validity and reliability of many of the tests used to assess PF during the pregnancy are unknown.\textsuperscript{19}

Since the assessment of PF in pregnancy requires special consideration to preserve fetal and maternal health,\textsuperscript{18, 20–21} understanding which fitness tests are valid, reliable, and associated with maternal and neonatal health 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 individuals; and (2) to evaluate the validity and reliability of the fitness tests, and their relationship with maternal and neonatal health.

METHODS
Registration and review guidelines and checklist
This systematic review was prospectively registered at PROSPERO (CRD42018117554; available at http://www.crd.york.ac.uk/PROSPERO). In addition, the review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines\textsuperscript{22} and the PRISMA checklist\textsuperscript{23} is included as online supplemental material 1, table S1. (1).

Search strategy
Articles were searched by two independent reviewers from two major databases, MEDLINE (PubMed) and the Web of Science (WOS) from inception to January 2021. For the search strategy undertaken in PubMed Medical Subject Heading, (MeSH) terms were used. All terms were combined using the connector OR for similar criteria the connector ‘AND’ was used to combine population group (ie, pregnant women), to delimit date of publication (‘0001/01/01’ (PDat): ‘2021/01/15’ (PDat)), to include full text papers, and to include studies performed in humans.

A similar search strategy and term combination was undertaken in the WoS (online supplemental material 2, table S2), although MeSH terms and its appropriate terms connection were not used as they are exclusive for PubMed. The complete search strategy and further details are presented in online supplemental material 2, tables S1 and S2.

Inclusion criteria
The inclusion criteria were as follows: (1) healthy pregnant individuals (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.

Quality assessment of the articles
To assess quality of the articles included in aim 2, three quality scores were applied. To assess validity and reliability, authors adapted two quality scores ad hoc previously used in two different systematic reviews following the same goal as the present review, however, undertaken in different populations.\textsuperscript{24, 25} To assess the association of PF with health-related outcomes the Effective Public Health Practice Project was used.\textsuperscript{26} All procedures are comprehensively described in online supplemental material 3, tables S3–S5.

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 summarised in online supplemental material 5, table S6.

RESULTS
A comprehensive PRISMA flow diagram is presented in figure 1.

Overall results, quality assessment and gestational week
The search 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 summary of the number of articles that
assessed PF during the pregnancy and the protocols used for its assessment is presented in figure 2. This has been organised based on each of the different PF components assessed in those articles. Moreover, a comprehensive diagram of the fitness tests and the different protocols performed to date, organised by PF component, is presented in figure 3.

Regarding aim 1, 99 tests (including 75 different protocols) were used to assess CRF, 8 12 13 18 27–108 (including 16 different protocols) to assess muscular fitness, 8 12 13 61 86 109–122 (including 13 different protocols) to assess flexibility, 12 13 110 114 123–127 45 tests (including 40 different protocols) to assess balance, 110 116 128–167 2 tests using the same protocol to assess speed 168 169 and 3 tests using the same protocol were multidimensional. 168–170 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) and reliability (n=4) of fitness tests. These articles are summarised in table 1. Of the three articles that assessed validity, two articles were classified as low quality 74, 169 and one as high quality. 75 Of the four articles that assessed reliability criteria, three were considered high quality 74, 113 168 and one low quality. The relationship of PF with maternal and neonatal health outcomes (n=16 tests) are summarised in table 2. Of these 16 tests, 11 were classified as very low quality 13, 57 68 95 108 111 126 157 158 and 5 were classified as low quality. 8, 63, 115 128 170

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 (ie, from week 0 to week 20 of gestation) and late pregnancy (ie, 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 (ie, range 2–5 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 labour.

**Aim 1: fitness tests used to evaluate PF in pregnant women**

**Cardiorespiratory fitness**

We identified 99 tests assessing CRF, of which 61 (62%) were performed on a cycle ergometer, 25 (25%) on a treadmill, 10 (10%) on a track and 3 (3%) used step...
protocols (figure 3). Of the 99 tests, a total of 75 corresponded to different protocols. For instance, there were 56 different protocols using a cycle ergometer, distributed as follows: only one article used the Arstila test\(^6\); one used the Bruce Protocol at 75% HR\(_{max}\)\(^2\); one applied the Modified Bruce ramp protocol at anaerobic threshold\(^1\); two employed the Modified Balke protocol at 70% HR\(_{max}\)\(^3\ 4\); one used a YMCA protocol\(^6\). The remaining of articles (n=55) used ad hoc tests (ie, specifically designed for the purpose of the investigation); 11 of which\(^3\ 32\ 37\ 38\ 41\ 45\ 57\ 64\ 79\ 107\) used steady-state tests and 44\(^28\ 30\ 33\ 35\ 36\ 39\ 40\ 42\ 43\ 44\ 46\ 47\ 49\ 50\ 51\ 52\ 54\ 55\ 56\ 59\ 60\ 62\ 63\ 65\ 66\ 68\ 69\ 72\ 73\ 75\ 77\ 90\ 92\ 100\ 102\ 108\ 171\ 172\) used incremental tests. When analysing the type of test based on intensity, we found that 13 tests were maximal\(^3\ 43\ 44\ 47\ 49\ 59\ 60\ 66\ 70\ 103\ 105\ 171\ 173\) submaximal tests\(^2\ 3\ 13\ 35\ 36\ 39\ 40\ 43\ 44\ 46\ 56\ 59\ 63\ 65\ 90\ 100\ 102\ 106\ 108\ 172\) and 3 used mixed tests\(^2\ 3\ 38\ 41\) 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\(^2\ 8\ 11\ 21\ 32\ 37\ 38\ 41\ 45\ 57\ 64\ 79\ 107\) and the traditional Balke protocol used twice in the same article\(^7\). The remaining of articles (n=44) used ad hoc tests (ie, specifically designed for the purpose of the investigation); 11 of which\(^3\ 32\ 37\ 38\ 41\ 45\ 57\ 64\ 79\ 107\) used steady-state tests and 44\(^28\ 30\ 33\ 35\ 36\ 39\ 40\ 42\ 43\ 44\ 46\ 47\ 49\ 50\ 51\ 52\ 54\ 55\ 56\ 59\ 60\ 62\ 63\ 65\ 66\ 68\ 69\ 72\ 73\ 75\ 77\ 90\ 92\ 100\ 102\ 106\ 108\ 172\) used incremental tests. When analysing the type of test based on intensity, we found that 13 tests were maximal\(^3\ 43\ 44\ 47\ 49\ 59\ 60\ 66\ 70\ 103\ 105\ 171\ 173\) submaximal tests\(^2\ 3\ 13\ 35\ 36\ 39\ 40\ 43\ 44\ 46\ 56\ 59\ 63\ 65\ 90\ 100\ 102\ 106\ 108\ 172\) and 3 used mixed tests\(^2\ 3\ 38\ 41\) containing submaximal and maximal stages within the same protocol.

Muscular fitness
A total of 28 tests (ie, 14% of all included articles) that included 16 different protocols assessed muscular fitness, of which 10 performed maximal hand-grip strength tests,\(^\)\(^8\ 12\ 13\ 15\ 36\ 109\ 115\) performed endurance hand-grip tests, 2 for 3 min\(^1\ 112\ 120\) and 1 for a 2 min period\(^1\ 119\) (figure 3). In two of the articles conducting an endurance hand-grip test,\(^1\ 118\ 119\) a hand-grip sphygmomanometer was used instead of dynamometry. On the other hand, one used a hand-held dynamometer fixed to a chair to assess quadriceps strength\(^1\) and one used a toe-grip dynamometer.\(^1\) Moreover, two ad hoc isometric tests were used to assess maximal voluntary hip extension and back flexors endurance in the same article.\(^1\) Finally, 13 dynamic endurance tests were found, 9 were listed as ad hoc tests\(^1\ 112\ 122\) and another 3 (30 s Chair Stand Test, 5 Times Sit to Stand test, Trendelenburg’s test) were classified as ‘other’ dynamic tests\(^1\ 112\ 117\).

Flexibility
Our search identified 14 (7%) tests that assessed flexibility using 13 different protocols, including the side bending test,\(^1\) the sit-and-reach test,\(^1\) the back-scratch test (twice),\(^1\ 13\ 16\) the motion analysis (ie, including three different tests such as the seated and standing forward flexion, seated and standing side to side flexion and seated axial rotation)\(^1\) and an optoelectrical system (ie, performing four different tests).\(^1\) Goniometry
was used in two different articles to measure hamstring flexibility,\(^{114}\) wrist flexion-extension and medial lateral deviation.\(^{124}\) Only one article used an ad hoc machine to test passive abduction of the left fourth finger.\(^{126}\)

### Balance

We identified 45 (24%) articles assessing balance of which 19 analysed static balance and 26 used dynamic balance with 40 different protocols. With regard to static balance, 18 were laboratory tests of which 12 assessed balance through stabilometry tests on a force platform,\(^{129}\)\(^{131}\)\(^{132}\)\(^{133}\)\(^{149}\)\(^{158}\)\(^{160}\)\(^{162}\)\(^{164}\)\(^{167}\) one on a pressure platform\(^{163}\) and another on an Equitest platform.\(^{165}\) Four articles did not mention the type of platform used.\(^{117}\)\(^{132}\)\(^{133}\)\(^{175}\) 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,\(^{116}\)\(^{131}\)\(^{132}\)\(^{147}\)\(^{158}\)\(^{159}\)\(^{162}\)\(^{165}\)–\(^{167}\)

### Table 1: Overview of studies that assessed the validity and/or reliability of fitness tests during pregnancy

<table>
<thead>
<tr>
<th>Reference (authors, year)</th>
<th>Validity</th>
<th>Reliability</th>
<th>Capacity evaluated, short test description and maternal and neonatal health outcomes or statistical results</th>
<th>Quality score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardio-respiratory fitness</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yeo et al (2005)(^{74})</td>
<td>Yes</td>
<td>Yes</td>
<td>Cornell Protocol on treadmill platform. Validity: Bland-Altman plots. The mean difference was 4.4±3.6mL/kg/min. Data indicated that VO2000 overestimates VO2 by an average of 4.4mL/kg/min compared with CPX/D. Pearson correlation coefficient between the average and difference of paired measurements was close but not significant (r=0.48; p&lt;0.01). Reliability: Paired t test (t (45)=3.9, p&lt;0.001). Linear regression: y=−0.96 x−1.6; 95% CI for the slope: 0.94 to 1.1; R²=0.91, p&lt;0.001</td>
<td>4–8</td>
</tr>
<tr>
<td>Mottola et al (2006)(^{75})</td>
<td>Yes</td>
<td>No</td>
<td>Modified Balke protocol on treadmill platform. Validity: Pearson Correlation: R²=0.72, R² adjusted=0.71 and SEE=2.7 (The prediction equation was compared with cross validation (n=39; p=0.78).</td>
<td>5</td>
</tr>
<tr>
<td>Muscular fitness</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gutke et al (2006)(^{21})</td>
<td>No</td>
<td>Yes</td>
<td>Maximal voluntary isometric hip extension Reliability: Spearman’s 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).</td>
<td>3</td>
</tr>
<tr>
<td>Yenisehir et al (2020)(^{17})</td>
<td>No</td>
<td>Yes</td>
<td>Five Times Sit to Stand Test (STSS) Reliability: Inter-rater reliability of STSS was excellent for subjects with and without pregnancy-related pelvic girdle pain (ICC ¼ 0.999, 95% CI ¼ 0.999 to 1.000; ICC ¼ 0.999, 95% CI ¼ 0.999 to 0.999, respectively). Test–retest reliability of STSS was also very high for subjects with and without PGP (ICC ¼ 0.986, 95% CI ¼ 0.959 to 0.995; ICC ¼ 0.828, 95% CI ¼ 0.632 to 0.920, respectively).</td>
<td>5–7</td>
</tr>
<tr>
<td>Flexibility</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lindgren and Kristiansson (2014)(^{126})</td>
<td>No</td>
<td>Yes</td>
<td>Ad hoc passive abduction of the left fourth finger. Reliability: Intraindividual 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.</td>
<td></td>
</tr>
<tr>
<td>Speed</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evensen et al (2015)(^{168})</td>
<td>No</td>
<td>Yes</td>
<td>Ten metres Timed walk Test Reliability: ICC from a one-way random effects model and reporting the 95% CI. Coefficients for test–retest reliability for 10mTWT: (ICC=0.74; 95% CI=0.42 to 0.90; SEM=0.17 m/s; MDC95=0.47 m/s) Coefficients for intertester reliability 10mTWT: (ICC=0.94; 95% CI=0.82 to 0.98; SEM=0.09 m/s; MDC95=0.25 m/s).</td>
<td>8</td>
</tr>
<tr>
<td>Evensen et al (2016)(^{168})</td>
<td>Yes</td>
<td>No</td>
<td>Ten metres Timed Walk Test Validity: Spearman correlation coefficient. Between the 10mTWT and ASLR (r=−0.65, p=0.003). Between the 10mTWT and PGQ (r=−0.25 to −0.56).</td>
<td>3</td>
</tr>
<tr>
<td>Multidimensional</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evensen et al (2015)(^{168})</td>
<td>No</td>
<td>Yes</td>
<td>Timed Up and Go Test (TUG) Reliability: ICC from a one-way random effects model and reporting the 95% CI. Coefficients for test–retest reliability TUG: (ICC=0.88; 95% CI 0.70 to 0.95; SEM=0.22 s; MDC95=1.16 s) Coefficients for intertester reliability TUG: (ICC=0.95; 95% CI 0.84 to 0.98; SEM=0.36 m/s; MDC95=1.00 m/s).</td>
<td>8</td>
</tr>
<tr>
<td>Evensen et al (2016)(^{169})</td>
<td>Yes</td>
<td>No</td>
<td>TUG Validity: Spearman correlation coefficient. Between the TUG and ASLR (r=0.73, p=0.001). Between the TUG and ASLR (r=0.73, p=0.001). Between the TUG and PGQ (r=0.41 to 0.52).</td>
<td>3</td>
</tr>
</tbody>
</table>

ASLR, Active Straight Leg Raised; MDC, minimal detectable change; 10mTWT, Ten-metre Timed Walk Test; PGP, Pelvic girdle pain; SEM, SE of measurement.
one with feet together,\textsuperscript{129} two used mixed protocols,\textsuperscript{128, 160} one with medial malleoli separated\textsuperscript{130} and four did not mention the standing posture.\textsuperscript{138, 149, 163, 164} Moreover, three articles used protocols with eyes open\textsuperscript{132, 149, 162} exclusively, eight articles used mixed protocols with eyes open and closed, one used visual target and visual tasks\textsuperscript{164} and six did not specify whether participants kept their eyes closed or opened. Only one article used a field test, the one-legged standing protocol.\textsuperscript{116} On the other hand, one test was a field-test without a platform.

In relation to the 26 articles measuring dynamic balance, 9 assessed balance using platforms. Each of these articles used a different testing tool such as a balance master platform,\textsuperscript{133} pressure platform,\textsuperscript{163} force platform,\textsuperscript{135} Equitest platform\textsuperscript{134} and a movable platform, which was used in two articles.\textsuperscript{136, 157} Two of these articles were walking protocols,\textsuperscript{133, 163} one with translational perturbations,\textsuperscript{157} one was standing with one knee flexed and arms across the chest.\textsuperscript{136, 157} Another 15 articles used three-dimensional (3-D) camera motion capture systems.
using 13 different protocols. Twelve of the 15 articles were walking protocols and 2 used a stand to sit motion protocol. Moreover, one article used a triaxial accelerometer; another article assessed balance through recording (without specification of camera type) and another using instrumented insoles. All three of these articles used walking protocols.

**Speed**

The only protocol that was used to assess speed during pregnancy was the 10 m timed walk test (10mTWT). However, the same test was identified in two different articles. In the 10mTWT, the participants commenced standing at a chair. When told to start, subjects walked as fast as possible along 14 m marked with white tape 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 metres per second (m/sec).

**Agility and coordination**

No articles of agility and coordination were identified.

**Multidimensional**

Our search identified a walking multidimensional test that was used in three studies. In the Timed Up and Go Test (TUG), the participant began seated in a chair with their arms on armrests and their feet against a start line. The purpose was to cross the front white line at 3 m 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).

**Aim 2: evaluation of the validity and reliability of the fitness tests, and their relationship with maternal and neonatal health**

Articles assessing validity and reliability are summarised in table 1. Articles assessing PF and its relationship with maternal and neonatal health outcomes are presented in table 2 and follows a similar format as Sallis et al.

**Cardiorespiratory fitness**

We identified two articles examining validity. Yeo et al. aimed to validate a portable metabolic testing system (VO2000) on healthy sedentary pregnant individuals. The VO2000 consistently overestimated VO2 measurements, compared with the same manufacturer’s reference system, by 4.4±3.6 SD mL/kg/min although the Pearson correlation was significant (r=0.48; p=0.01). When the 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. provided a prediction equation for VO2peak in pregnant individuals between 16 and 22 weeks of gestation, using a modified Balke protocol. The results of this equation revealed an adjusted R² of 0.71 and differences between actual and predicted VO2 of 2.7 mL/kg/min. 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/min, while the actual value was 23.54±5.9 mL/kg/min (p=0.78).

A total of six articles analysed the association of CRF with maternal and neonatal health outcomes. Pomerance et al. observed that VO2max was inversely associated with the length of labour in multiparas (r=−0.65; p=0.001) and prepregnancy weight (r=−0.63; p=0.001). However, VO2max was not correlated with newborn weight, length or head circumference, or with the 1 min Apgar scores (all p>0.05). In the same line, Wong and McKenzie observed that fit mothers showed lower HR at submaximal exercise intensity (p<0.05) and the second stage of labour was shorter (no statistics reported) compared with 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 effects of maternal fitness on newborn weight or Apgar scores.

In addition, Erkola and Rauramo found that newborns from fit pregnant individuals 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. observed that maternal CRF at the 16th gestational week was related to higher arterial umbilical cord PO2 (r=0.267, p<0.05), and those who had caesarean sections had significantly lower CRF compared with those who had vaginal births (p=0.001).

Moreover, Bisson et al. studied the association of CRF in early pregnancy with physical activity before and during early pregnancy. The authors found that a higher VO2peak in early pregnancy was positively associated with physical activity spent at sports and exercise before and during early pregnancy (p<0.001).

**Muscular fitness**

Only two muscular fitness tests assessed reliability. Yenişehr et al. analysed reliability and validity of Five Times Sit-to-Stand. Inter-rater reliability was excellent for subjects with and without pelvic girdle pain (PGP) (intra-class correlation coefficient, ICC=0.999, 95% CI 0.999 to 1.000; ICC=0.999, 95% CI 0.999 to 0.999, respectively). Test–retest reliability was also very high for subjects with and without PGP (ICC=0.986, 95% CI 0.959 to 0.995; ICC=0.828, 95% CI 0.632 to 0.920, respectively).

Gutke et al. analysed the reliability for an ad hoc test. This test consisted of a maximal voluntary isometric hip extension with a fixed sensor holding a sling around the thigh and pulling for 5 s during 3 reps with 5–10 s 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 not reported).

Bisson et al. 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). Želaźniwick and Pawłowski et al. 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.23 found greater hand-grip strength weakly associated with greater neonatal birth weight (r=0.191, p<0.05). Wickboldt et al.11 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 hand-grip strength (r=0.56; p<0.001) and the area under the curve of hand-grip force (r=0.55; p<0.001).

Flexibility
Lindgren and Kristiansson126 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 analysed by the intraclass correlation 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.

Only two flexibility tests evaluated associations with maternal and neonatal health outcomes. Lindgren and Kristiansson126 found that women with greater passive abduction angle of the left fourth finger was associated with the highest back pain incidence (OR 1.09; 95% CI 1.01 to 1.17; p=0.022) and the highest number of previous pregnancies (OR 3.24; 95% CI 1.57 to 6.68; p=0.002). Baena-García et al.13 found an 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.

Balance
No validity or reliability assessments were performed regarding balance tests.

Three articles associated balance with neonatal and maternal health-related outcomes. Öztürk et al.126 observed that static balance decreased and fall risk increased in pregnant individuals with lower back pain (49.90±24.47 vs 28.47±19.60; p<0.0001). In relation to exercise, McCrory et al.157 showed that exercise may play a role in fall prevention in pregnancy (p<0.005) and they also found that dynamic balance is altered in pregnant individuals who have fallen compared with non-fallers and non-pregnant individuals (p<0.001). Nagai et al.58 studied the relationship between anxiety and balance. They concluded that when anxiety increases during pregnancy, the standing posture is destabilised (r=0.559, p=0.020), which may increase the chance of falling.

Speed
Validity and reliability for 10mTWT was studied by Evensen et al. in two different articles.168 169 In 2015, Evensen et al.68 analysed the test–retest reliability of 10mTWT showing an ICC of (0.74). Intertester reliability was determined in the first 13 participants with strong correlation (ICC=0.94). In 2016,169 the same authors analysed 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).

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

Agility and coordination
No articles were identified.

Multidimensional
Validity and reliability for TUG was analysed by Evensen et al.168 169 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 ASLR (r=0.73, p=0.001).

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

DISCUSSION
Summary of the evidence
This systematic review revealed that PF has been assessed through a wide variety of tests during pregnancy. However, little is known on the validity and reliability of the tests performed, and the large variety of tests makes it challenging to compare results from different studies. Until a battery of specific fitness tests for pregnant women is developed and validated, the confidence of PF data obtained during pregnancy is limited and should be interpreted with caution. Consequently, the appropriateness of using this PF data to prescribe exercise during pregnancy could be questioned and is a matter that requires special attention. In this context, it is also difficult to evaluate the association of PF with maternal and neonatal health which, in fact, is of wide clinical and public health interest. However, some studies observed associations of PF with maternal and neonatal health outcomes, which needs to be replicated once a PF test battery is released. We strongly suggest that extensive research must be performed to validate such battery of PF tests.

Cardiorespiratory fitness
This systematic review identified that a 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.85 for pregnant women has been the most

Protected by copyright.
frequently used test. There have been more incremental tests used for CRF during the pregnancy compared with steady-state tests and more submaximal compared with 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 %HRmax or %VO2max and others used absolute intensity, such as specific HR (beats per minute). Among the studies that used %HRmax as a test termination criterion, there was a variety of percentages such as 70%, 34 35 90 100 75% 27 29 69 97 or 85%, 16 54 74. Among the studies that used %VO2max, there were different percentages such as 40%, 38 50%, 53 55 56 60% 32 38 or 70%. Among the studies that used absolute HR as a test termination criterion, the HR for finalising the tests were set either at 125, 61 150, 36 45 62 108 155, 91 160 65 or 170 90 53 55 56 beats per minute. Some studies even used the rate of perceived exertion as complementary criteria 46 50 106 or peak aerobic power. 39 These complementary criteria have been recommended and studied in pregnant women by authors like Hesse et al 33 since the physical and emotional changes during pregnancy limit performance. It must be noted that the same equation was not used to estimate HRmax. Some articles used the traditional 220-age formula 29 35 54 69 97 while others used the Karvonen 74 or Tanaka 100 formulas. Some articles did not specify how HRmax was estimated. 27 34 90 This heterogeneity could be due to the physiological complexity of pregnancy, 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 responses since Bijl et al 86 observed a slower haemodynamic recovery and an increased ventilatory response to exercise in early pregnancy compared with non-pregnant women. With regard to maximal tests, different terms have been used for maximal criteria such as volitional fatigue, 30 33 44 47 48 96 103 105 exhaustion, 31 anaerobic threshold, 73 80 104 171 and point of symptom limitation. 39 60 102

This lack of consensus has many drawbacks that should be resolved in view of the need to accurately assess CRF during the 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 the pregnancy. In particular, it seems essential to develop a treadmill and a cycle ergometer submaximal test that reveals sufficient validity to confidently estimate VO2max throughout gestation.

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 six studies 2 12 13 112 117 122 179 assessed endurance and none of them assessed power in pregnancy. In most studies, muscular strength was evaluated through hand-grip maximal strength using a dynamometer. However, two studies used a hand-grip sphygmomanometer test. 118 119 Some of the hand-grip tests were performed in a standing position, 8 106 while others used a sitting position 110 or supine position 113 and others did not reveal the position used for the assessment. 86 112 114 115 Some tests were completed three times, 112 others twice 8 86 115 and others only once. 110 113 114 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 hand-grip strength. While hand-grip strength is a good marker of health, 186 it is unclear whether hand-grip responds to changes following exercise interventions. Therefore, validating other muscular strength tests, including lower limb strength tests, is needed for researchers and practitioners to confidently assess muscular strength during the pregnancy.

There were no validity studies and the reliability was assessed only in one maximal isometric hip extension test. 123 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 hand-grip strength was associated with higher birth weight. 8 115 Moreover, increased hand-grip strength was produced during uterine contraction. 111 The advantage of using hand-grip 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 outcomes is of clinical importance. Moreover, other studies are needed to understand the extent to which preserving strength throughout pregnancy and post partum relates to clinical outcomes.

Flexibility

Although there were seven 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 and Kristiansson 126 found that higher flexibility showed higher low back pain. Despite the limitation of a finger laxity test, we considered these findings an interesting association that warrants further investigation 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 88 are very relevant to fetal health since 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.

Balance

We identified that balance was the second PF component most frequently evaluated during pregnancy, following CRF. This makes sense since the centre of gravity changes during pregnancy as a result of expansion of the
uterus and the risk of falls increases. However, there is high heterogeneity between the protocols employed in different studies. For static balance, the protocol most frequently used was stabilometry on a force platform with bipedal support and eyes open and eyes closed within the same test. 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.\(^{139-142}\)\(^{144} \)\(^{165}\) Likewise, we observed differences between the number of platform pieces, trials and Hz used. Some protocols were performed on two piece platforms,\(^ {130} \)\(^ {131} \)\(^ {138} \)\(^ {158} \)\(^ {160} \)\(^ {166} \)\(^ {167}\) others on one piece platforms\(^ {129} \)\(^ {132} \)\(^ {158} \)\(^ {160} \)\(^ {166} \)\(^ {167}\) and others did not specify the type of platform.\(^ {165-167} \) Although the number of trials and the frequency of recording (ie, Hz) are important protocol parameters that should be carefully documented, only 5 (out of 13) articles described the number of trials\(^ {133} \)\(^ {138} \)\(^ {166} \)\(^ {175}\) and 1 described frequency of recording.\(^ {149}\) The usefulness of these tests is restricted to the research area and all of them use expensive technological tools; therefore, it is difficult to extrapolate these tests to fitness centres or clinical settings. Falls during pregnancy could be prevented if balance was easily assessed. For this reason, it is necessary to develop an inexpensive and easy-to-use balance field test.

Validity and reliability of PF tests, and association with maternal and neonatal health

Unfortunately, studies that examine validity and reliability of PF tests are scarce. The PF component most frequently studied was CRF. However, we only found two studies that analysed the validity of the CRF tests, and no studies examined the reliability of these tests. On a treadmill platform, Mottola et al.\(^ {25}\) validated a special equation for modified Balke protocol that has been used by numerous other authors. In contrast, Yeo et al.\(^ {24}\) aimed to validate a portable metabolic testing system (mod. VO2000) but it overestimated VO\(_2\) measurements for pregnant individuals compared with non-pregnant females and males.

Regarding muscular fitness, the hand-grip test was most commonly used; this test was used as the gold standard for muscular fitness during pregnancy. Only Gutke et al.\(^ {21}\) 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 and multidimensional tests of PF have been researched by Evensen et al.\(^ {168-169}\) They demonstrated that TUG and 10mTWT are reliable and valid tests for use during the pregnancy.

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 pregnancy 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 the association of PF with maternal and neonatal health outcomes, we 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 labour\(^ {57} \)\(^ {108}\) and a lower risk of caesarean section.\(^ {13}\) However, no association was found regarding other fetal outcomes such as Apgar scores or the newborn anthropometrics.\(^ {57} \)\(^ {108}\) By contrast, muscular strength was associated with optimum infant birth weight.\(^ {8} \)\(^ {15} \)\(^ {115}\) Other neonatal outcomes like fetal umbilical cord pH were positively associated with maternal CRF.\(^ {68}\) On the other hand, better balance scores were associated with lower risk of falls,\(^ {128} \)\(^ {158} \)\(^ {181}\) which is of particular interest for exercise professionals, who might include balance as a component of exercise programs for pregnant women.

Finally, Evensen et al.\(^ {169}\) found that PGP could be a limiting factor to assess PF in pregnant individuals since the time of TUG was significantly higher in those with pain than in asymptomatic pregnant and non-pregnant individuals.

None of the studies reviewed in this article have described adverse events during PF assessment. Moreover, official bodies such as the American College of Obstetricians and Gynecologists, the Canadian Society of Exercise Physiology and the Society of Obstetricians and Gynaecologists of Canada have highlighted the benefits of an adequate PF assessment, and assert the need of consensus in PF assessment during the pregnancy.\(^ {182}\) Consequently, the findings from this study have important research and clinical implications.

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 analyse PF assessments, the validity and reliability of fitness tests, and their relationship with maternal and neonatal health outcomes during the 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 outcomes. All this information is of wide and undoubted clinical interest.

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, the available evidence regarding the association of PF with maternal and neonatal health outcomes is scarce and is a matter of further investigation. Provided the need to assess PF during the pregnancy and the importance not only to understand the physical state of the pregnant individual 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 panel to develop a battery of PF tests to assess the different PF components during pregnancy.

Author affiliations
1Department of Physical Education and Sport, Universidad de Granada, Granada, Spain
2Sport and Health University Research Centre, Universidad de Granada, Granada, Spain
3Applied Sports Technology Exercise and Medicine Research Centre, Swansea University, Swansea, UK
4Sport Science Department, Swansea University, Swansea, UK
5GALENO Research Group, Department of Physical Education, Faculty of Education, Universidad de Granada, Granada, Spain
6The Biomedical Research and Innovation Institute of Cadiz (INIBIC), Cadiz, España
7Kinesiology, East Carolina University College of Health and Human Performance, Greenville, North Carolina, USA
8Department of Obstetrics & Gynecology, Brody School of Medicine, East Carolina University, Greenville, North Carolina, USA
9Gynaecology and Obstetrics Unit, ‘San Cecilio’ University Hospital, Universidad de Granada, Granada, Spain
10The Biosanitary Research Institute of Granada.ibs, Granada, Spain
11R. Samuel McLaughlin Foundation- Exercise and Pregnancy Lab, School of Kinesiology, University of Western Ontario, London, Ontario, Canada
12Department of Physiology, Institute of Nutrition and Food Technology and Biomedical Research Centre, Universidad de Granada, Granada, Spain
13Department of Education, Faculty of Education Sciences, University of Almería, Almería, Spain
14SPORT Research Group (CTS-11R. Samuel McLaughlin Foundation- Exercise and Pregnancy Lab, School of Education, Sciences, Universidad de Cadiz, Cadiz, Spain
15The Biomedical Research and Innovation Institute of Cadiz (INIBIC), Cadiz, España

Twitter Lidia Romero-Gallardo @lidiaromero_owa and Alberto Soriano-Maldonado @AlbertoSoriano_


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 in 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. This study is included in the Development Fund (ERDF), ref. SOMM17/6107/UGR. This study is included in the Consejería de Conocimiento, Investigación and Universidades and European Regional Development Fund (ERDF), ref. SOMM17/6107/UGR. This study is included in the Plan Propio de Investigación 2016, Excellence actions: Units of Excellence: Unit of Excellence in 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. This study is included in the Consejería de Conocimiento, Investigación and Universidades and European Regional Development Fund (ERDF), ref. SOMM17/6107/UGR. This study is included in the

Competing interests Yes.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

Author note Please note that this work was posted as a pre-print prior to submission. The pre-print version can be accessed here https://doi.org/10.1101/2021.06.26.21259584.

ORCID ids Lidia Romero-Gallardo http://orcid.org/0000-0001-6341-855X Olga Roldan Reoyo http://orcid.org/0000-0003-2625-4060 Jose Castro-Piñero http://orcid.org/0000-0002-7353-0382 Linda E May http://orcid.org/0000-0002-8231-2280 Olga Ocón-Hernández http://orcid.org/0000-0002-8113-3651 Michelle F Mottola http://orcid.org/0000-0002-8707-4656 Virginia A Aparicio http://orcid.org/0000-0002-2867-378X Alberto Soriano-Maldonado http://orcid.org/0000-0002-4626-420X

REFERENCES


