**Supplementary online resource 2** Full search strategy as undertaken in MEDLINE

| | | |
|---|---|---|
| ☐ | S8 | S1 AND S7 |
| ☐ | S7 | S4 OR S6 |
| ☐ | S6 | S2 AND S3 AND S5 |
| ☐ | S5 | interval OR intermittent |
| ☐ | S4 | HIIT OR HIT OR HIIE |
| ☐ | S3 | high intensity OR high-intensity OR vigorous OR maximal OR sprint |
| ☐ | S2 | ( (ZW "exercise") or (ZW "exercise & training") or (ZW "exercise ( ex )") or (ZW "exercise (mesh d015444)") or (ZW "exercise / physical activity") or (ZW "exercise [mesh]") or (ZW "exercise activity") or (ZW "exercise adaptation") or (ZW "exercise adaptations") ) OR workout or workouts OR intervention |
| ☐ | S1 | sex OR gender OR ( females or women or female or woman ) AND ( males or men or male or man ) |

**Supplementary online resource 3** Newcastle - Ottawa quality assessment scale for case control studies [1], adapted for assessing bias for sex differences in outcomes to HIIT

**Note:** A study can be awarded a maximum of one star for each numbered item within the Selection and Exposure categories. A maximum of two stars can be given for Comparability.

| Scale item | Detailed explanation | Star rating |
|---|---|---|
| **Selection** | | |
| **1) Are the outcome measures adequate?** ***('outcome measures' vs 'case definition')*** | | |
| a) yes, with independent validation | *The outcome measures used have good validity; objective/gold standard where possible; preferably researcher blinded if the study is controlled; likely to have accurately measured the outcome of interest.* | ★ |
| b) yes, eg record linkage or based on self-reports | *Potentially valid outcome measures or unclear; estimated (not measured directly) or only somewhat likely to have accurately measured the outcome of interest* | No star |
| c) no description | *No description or inadequate measures used* | No star |
| **2) Representativeness of the cases** | | |
| a) consecutive or obviously representative series of cases | *Low risk of selection bias such as a large sample size, not self-selected into the intervention or whole target groups included; randomised controlled trial design* | ★ |
| b) potential for selection biases | *Moderate risk of selection bias (most likely); participants self-selected into research study; non-controlled study design* | No star |
| **3) Selection of controls** ***(male group, in comparison to females)*** | | |
| a) community controls | *Male participants recruited from the same community or population as female participants. Same recruitment strategies used.* | ★ |
| b) hospital controls | *Male participants recruited from a different community or population as female participants or different recruitment strategies used.* | No star |
| c) no description | *No description of the community or population that male or female participants were recruited.* | No star |
| **4) Definition of controls** ***– same inclusion criteria used for selection of male and female participants (i.e. baseline fitness (relative) or training level, pathology or risk factors between groups)*** | | |
| a) no history of disease (endpoint) | *Same inclusion criteria used and no significant differences in baseline fitness (relative) or training level, pathology or risk factors between groups are present* | ★ |
| b) no description of source | *Differing inclusion criteria used or significant differences in baseline fitness (relative) or training level, pathology or risk factors between groups are present* | No star |
| **Comparability (sex groups)** | | |
| **1) Comparability of** ***males and females*** **on the basis of the design or analysis** | | |
| a) study controls for *age* | *Male and female groups are of a similar mean age at baseline* | ★ |
| b) study controls (accounts) for *menstrual cycle* | *Study design or analysis accounts for menstrual cycle (i.e. female participants are tested at the same time during their cycle or other means of accounting for the cycle are used)* | ★ |
| **Exposure (intervention)** | | |
| **1) Ascertainment of exposure** ***- Description of intervention – has the intervention been applied equitably for both genders? (i.e. for HIIT example, intensities used were relative to individual maximal performance)*** | | |
| a) secure record (eg surgical records) | *Participants completed an equitable intervention with good adherence and researcher validation (lab-based or supervised exercise sessions with non-adherers excluded)* | ★ |
| b) structured interview where blind to case/control status | *Participants completed an equitable intervention with good adherence but self-reported adherence (non-supervised exercise sessions with non-adherers excluded)* | ★ |
| c) interview not blinded to case/control status - | *Equitable intervention with non-adherers NOT excluded* | No star |
| d) written self-report or medical record only | *Non-equitable intervention – intervention is not relatively similar for males and females* | No star |

| | | |
|---|---|---|
| e) no description | *Inadequate description of intervention* | No star |
| **2) Same method of ascertainment for *males and females*** | | |
| a) yes | *Adherence was tracked using the same method for both male and female participants.* | ★ |
| b) no | *Adherence was tracked using the differing methods for both male and female participants.* | No star |
| **3) Non-response rate *(withdrawals from the intervention and non-adherers)*** | | |
| a) same rate for *males and females* | *Withdrawal and non-adherence rates were similar for male and female participants* | ★ |
| b) non respondents described | *Withdrawal and non-adherence rates were not similar for male and female participants, but differences described in limitations or accounted for in analysis* | No star |
| c) rate different and no designation | *Withdrawal and non-adherence rates were not similar for male and female participants, and differences were NOT described in limitations or accounted for in analysis* | No star |

***Maximum possible score being ten stars with the higher the number of stars the high level of quality (lower risk of bias) for each individual study.***

**References**

[1] Wells G, Shea B, O'Connell D, Robertson J, Peterson J, Losos M, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomized studies in meta- analysis.

**Supplementary online resource 4** PRISMA 2020 checklist

| Section and Topic | Item # | Checklist item | Location where item is reported |
|---|---|---|---|
| **TITLE** | | | |
| Title | 1 | Identify the report as a systematic review. | P1 |
| **ABSTRACT** | | | |
| Abstract | 2 | See the PRISMA 2020 for Abstracts checklist. | P3-4 |
| **INTRODUCTION** | | | |
| Rationale | 3 | Describe the rationale for the review in the context of existing knowledge. | P5-7 |
| Objectives | 4 | Provide an explicit statement of the objective(s) or question(s) the review addresses. | P7 |
| **METHODS** | | | |
| Eligibility criteria | 5 | Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses. | P8<br>Section 2.2 |
| Information sources | 6 | Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted. | P7<br>Section 2.2 |
| Search strategy | 7 | Present the full search strategies for all databases, registers and websites, including any filters and limits used. | P7<br>Section 2.2<br>Supplementary Online Resource 2 |
| Selection process | 8 | Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process. | P8<br>Section 2.2 |
| Data collection process | 9 | Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process. | P9<br>Section 2.4 |
| Data items | 10a | List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect. | P9-10<br>Section 2.4 |
| | 10b | List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information. | P9-10<br>Section 2.4 |
| Study risk of bias assessment | 11 | Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process. | P9<br>Section 2.3<br>Supplementary Online Resource 3 |
| Effect measures | 12 | Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results. | P9-10<br>Section 2.4 |
| Synthesis | 13a | Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention | P9-10 |

| Section and Topic | Item # | Checklist item | Location where item is reported |
|---|---|---|---|
| methods | | characteristics and comparing against the planned groups for each synthesis (item #5)). | Section 2.4 |
| | 13b | Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions. | P9-10<br>Section 2.4 |
| | 13c | Describe any methods used to tabulate or visually display results of individual studies and syntheses. | P9-11<br>Sections 2.4-2.6 |
| | 13d | Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used. | P9-10<br>Sections 2.4-2.6 |
| | 13e | Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression). | P10<br>Section 2.5 |
| | 13f | Describe any sensitivity analyses conducted to assess robustness of the synthesized results. | P10<br>Section 2.5 |
| Reporting bias assessment | 14 | Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases). | P11<br>Section 2.7 |
| Certainty assessment | 15 | Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome. | P10-P11<br>Sections 2.5-2.7 |
| **RESULTS** | | | |
| Study selection | 16a | Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram. | P11<br>Section 3.1<br>PRISMA diagram<br>Figure 1 |
| | 16b | Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded. | P11-12<br>Sections 3.1 and 3.4.1<br>Supplementary Online Resource 6 |
| Study characteristics | 17 | Cite each included study and present its characteristics. | Table 1<br>Table 2 |
| Risk of bias in studies | 18 | Present assessments of risk of bias for each included study. | Supplementary Online Resource 5 |
| Results of individual studies | 19 | For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots. | P12-16<br>Sections 3.4-3.6<br>Tables 5-8 |
| Results of syntheses | 20a | For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies. | P12<br>3.2 |
| | 20b | Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect. | Figures 2-7 |

| Section and Topic | Item # | Checklist item | Location where item is reported |
|---|---|---|---|
| | | | Tables 3-4<br>Supplementary Online Resource 7 |
| | 20c | Present results of all investigations of possible causes of heterogeneity among study results. | P12-15<br>Sections 3.4-3.6<br>Tables 3-4 |
| | 20d | Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results. | P12-15<br>Sections 3.4-3.6 |
| Reporting biases | 21 | Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed. | P16-17<br>Section 3.7<br>Figure 8a, b, and c |
| Certainty of evidence | 22 | Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed. | P16-17<br>Section 3.7<br>Figure 8a, b, and c |
| **DISCUSSION** | | | |
| Discussion | 23a | Provide a general interpretation of the results in the context of other evidence. | P17-21<br>Section 4.0 |
| | 23b | Discuss any limitations of the evidence included in the review. | P20-21<br>Section 4.0 |
| | 23c | Discuss any limitations of the review processes used. | P20-21<br>Section 4.0 |
| | 23d | Discuss implications of the results for practice, policy, and future research. | P17-21<br>Section 4.0 |
| **OTHER INFORMATION** | | | |
| Registration and protocol | 24a | Provide registration information for the review, including register name and registration number, or state that the review was not registered. | P7<br>Section 2.2 |
| | 24b | Indicate where the review protocol can be accessed, or state that a protocol was not prepared. | P7<br>Section 2.2 |
| | 24c | Describe and explain any amendments to information provided at registration or in the protocol. | P7-11<br>Sections 2.2-2.5 |
| Support | 25 | Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review. | P2 |
| Competing interests | 26 | Declare any competing interests of review authors. | P2 |
| Availability of data, code and other materials | 27 | Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review. | Tables 1-8<br>Figures 1-8<br>Supplementary Online Resource 1- |

| Section and Topic | Item # | Checklist item | Location where item is reported |
|---|---|---|---|
| | | | 7 |

**Supplementary online resource 5** Breakdown of scoring for individual items of the Newcastle-Ottawa Scale for included studies and categorisation according to the Agency for Healthcare Research and Quality (AHRQ) thresholds*

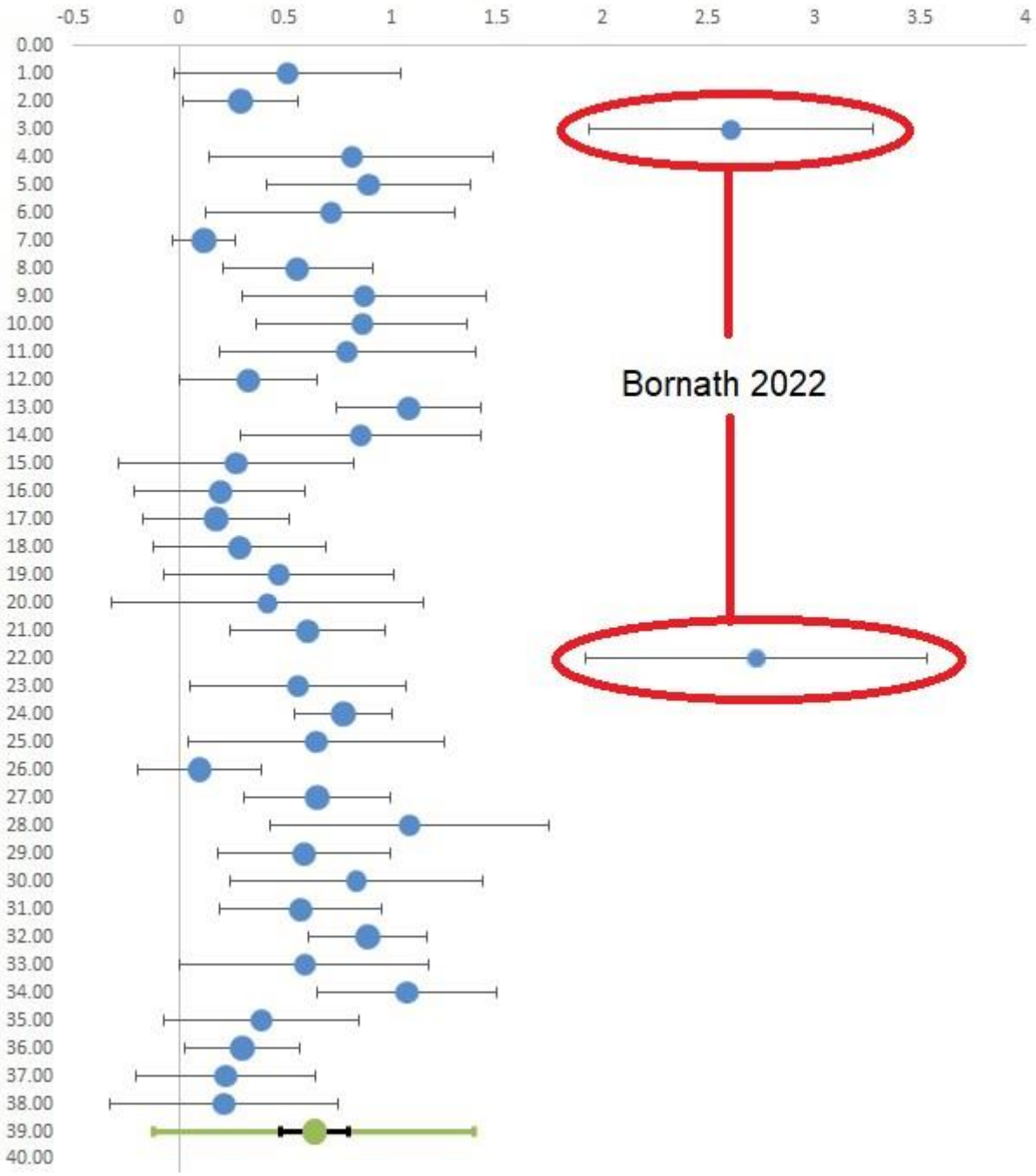
| **Domain** | **Selection** | | | | **Comparability (sex groups)** | | **Exposure (intervention)** | | | **Total score** | **Category** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Scoring item** | 1) Are the outcome measures adequate? | 2) Representativeness of the cases | 3) Selection of males compared to females | 4) Definition of comparison group | 5) Comparability of males and females on the basis of the design or analysis | | 6) Ascertainment of exposure - Description of intervention | 7) Same method for males and females | 8) Withdrawals and non-adherers | | |
| **Reference** | | | | | a) study controls for age | b) study controls (accounts) for menstrual cycle | | | | | |
| Astorino 2011 [76] + Astorino 2012 [66] | ✓ | | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | 7 | Good |
| Bagley 2016 [77] + Bagley 2021 [78] | ✓ | | | ✓ | ✓ | | ✓ | ✓ | ✓ | 6 | Fair |
| Bornath 2022 [79] | ✓ | | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | 7 | Good |
| Bostad 2021 [80] | ✓ | | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | 7 | Good |
| Chrøis 2020 [81] + Søgaard 2018 [93] | ✓ | | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | 7 | Good |
| Cicioni-Kolsky 2013 [98] | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | 8 | Good |
| Dalzill 2014 [82] | | ✓ | ✓ | ✓ | | | ✓ | ✓ | ✓ | 6 | Poor |
| Esbjörnsson Liljedahl 1996 [97] | ✓ | | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | 7 | Good |
| Fisher 2017 [96] + Hoffmann 2021 [73] | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 8 | Good |
| Gillen 2014 [83] | ✓ | | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | 7 | Good |
| Hiam 2021 [84] | ✓ | | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | 7 | Good |
| Hirsch 2021 [72] | ✓ | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 7 | Fair |
| Lepretre 2009 [85] | ✓ | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 7 | Fair |
| Liu 2021 [86] | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | 8 | Good |
| Marterer 2020 [67] | ✓ | | ✓ | | ✓ | | ✓ | ✓ | ✓ | 6 | Fair |
| Menz 2015 [87] | ✓ | | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | 7 | Good |
| Metcalfe 2012 [88] | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 8 | Good |
| Metcalfe 2016 [68] | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 8 | Good |
| Molina-Hidalgo 2020 [69] | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | 8 | Good |
| Mucci 2004 [70] | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 8 | Good |
| Phillips 2017 [89] | ✓ | ✓ | ✓ | | | | ✓ | ✓ | ✓ | 6 | Poor |
| Sawashita 2009 [90] | | | | ✓ | ✓ | | ✓ | ✓ | ✓ | 5 | Poor |
| Scalzo 2014 [91] | ✓ | | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | 7 | Good |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Schmitz 2020 [95] | | | ✓ | ✓ | ✓ | | ✓ | ✓ | | 5 | Fair |
| Schmitz 2019 [75] | ✓ | | ✓ | | ✓ | | ✓ | ✓ | ✓ | 6 | Fair |
| Schubert 2017a [71] + Schubert 2017b [92] | ✓ | | ✓ | | | ✓ | ✓ | ✓ | ✓ | 6 | Fair |
| Støren 2017 [94] | ✓ | | ✓ | | ✓ | | ✓ | ✓ | ✓ | 6 | Fair |
| Weber 2002 [74] | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 8 | Good |
| **IRR (% similarity)** | 92.59 | 74.07 | 66.67 | 81.48 | 85.19 | 92.59 | 88.89 | 100 | 66.67 | | |
| **IRR Mean (±SD) of individual items** | 83.13 (±11.87) | | | | | | | | | | |
| | **Total score, mean (±SD)** | | | | | | | | | 6.89 (0.93) | |

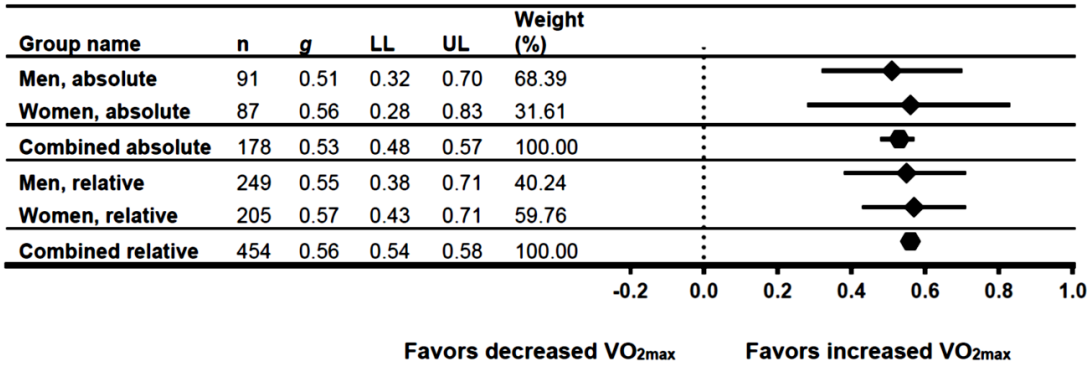
*Domain scores were used to categorize studies into good, fair, and poor quality using the following thresholds outlined by the Agency for Healthcare Research and Quality (AHRQ).

- Good quality: must have scored three or four stars in the selection domain, as well as one or two stars in the comparability domain, and two or three stars in the exposure (intervention) domain.
- Fair quality: two stars in the selection domain, one or two stars in the comparability domain, and two or three stars in the exposure (intervention) domain.
- Poor-quality studies if they scored zero or one star in the selection domain, zero stars in the comparability domain, or zero or one star in the exposure (intervention) domain.

**Supplementary online resource 6** Forest plot showing Bornath 2022 study as an outlier

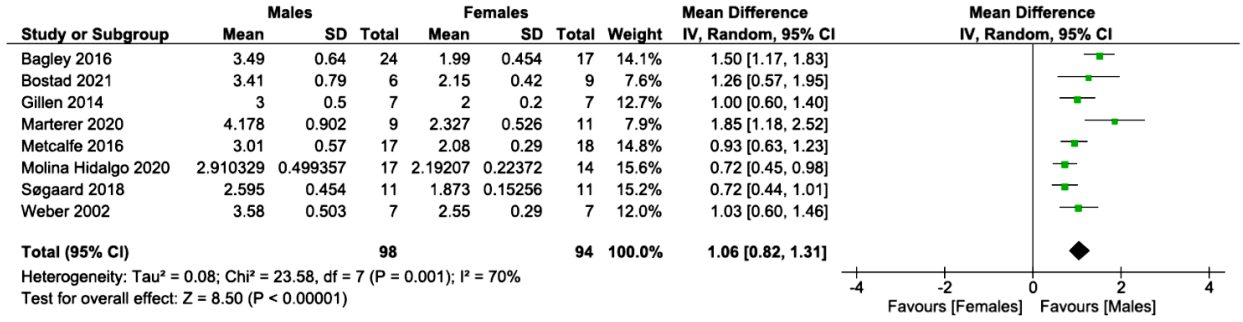


**Supplementary online resource 7** Forest plot of pooled absolute and relative $VO_{2max}$ outcomes

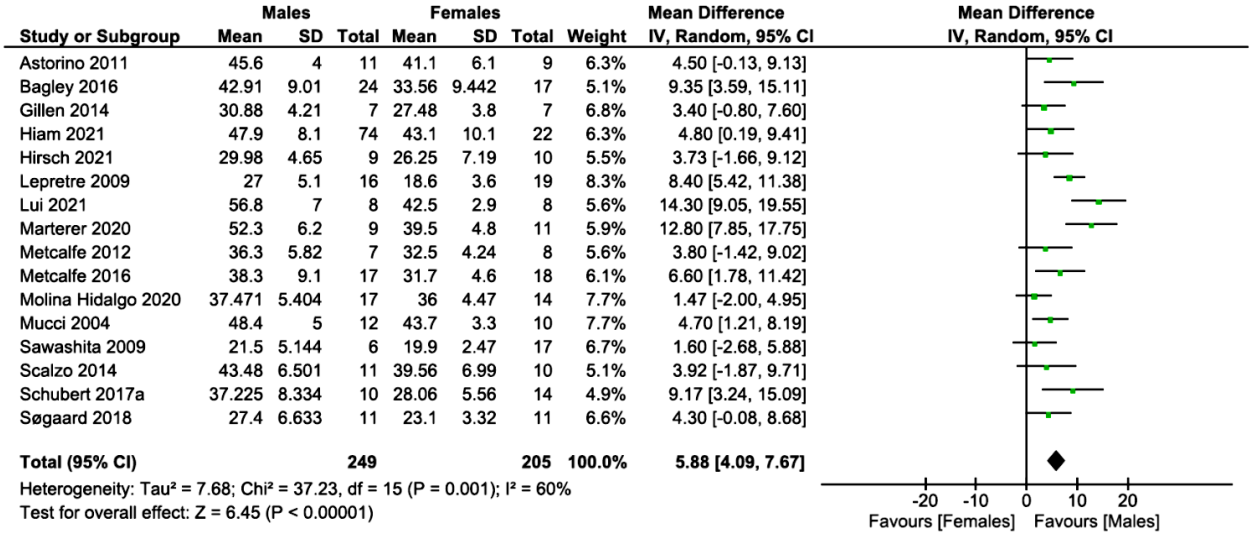


**Supplementary online resource 8**

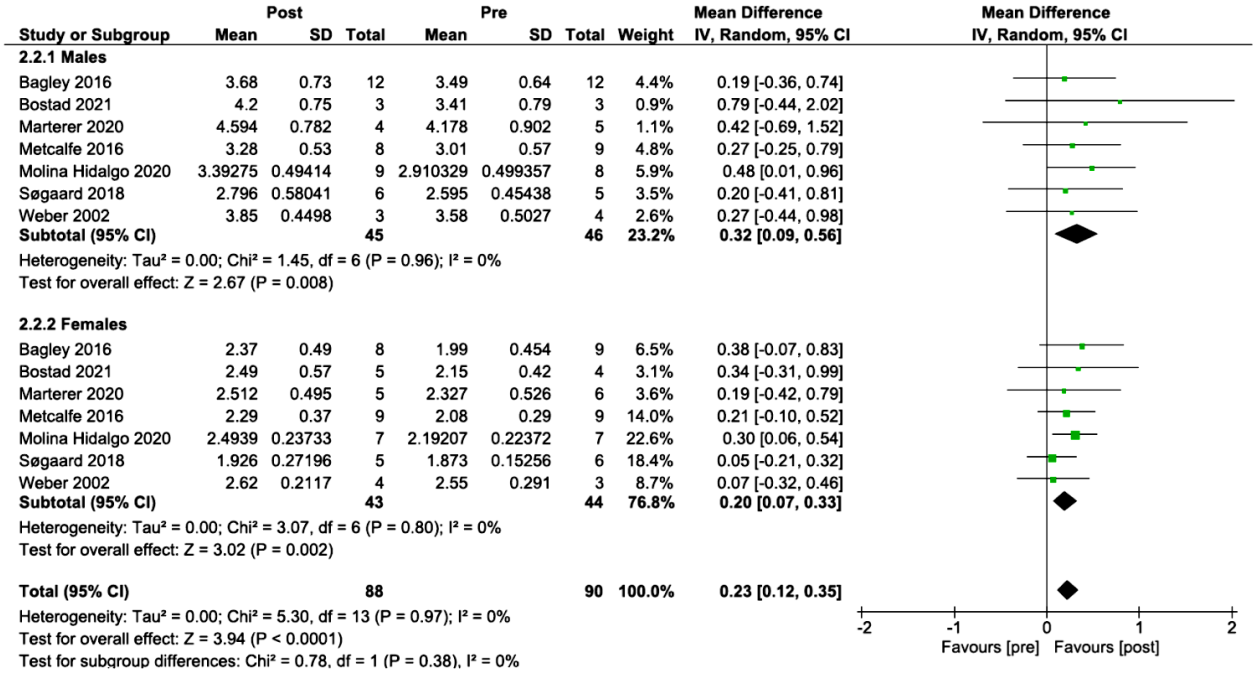
a) Forest plot of pooled baseline absolute $VO_{2max}$, males versus females, raw mean difference




b) Forest plot of pooled baseline relative $VO_{2max}$, males versus females, raw mean difference



c) Forest plot of pooled pre-post absolute $VO_{2max}$, sub-grouped as males and females, raw mean difference



d) Forest plot of pooled pre-post relative $VO_{2max}$, sub-grouped as males and females, raw mean difference