**Table S1**. Search syntax for all searched databases

|  |  |
| --- | --- |
| **Database**  | **Search syntax** |
| AUSPORT | caffeine AND (meta-an\* OR "systematic review") AND (exercise OR training OR muscle OR "physical performance") |
| EBSCOHost Research Databases (including: Academic Search Premier, CINAHL, ERIC, Health Source: Nursing/Academic Edition, MasterFILE Premier, PsycINFO, SPORTDiscus) | caffeine AND (meta-an\* OR "systematic review") AND (exercise OR training OR muscle OR "physical performance") |
| Cochrane Library | caffeine AND (meta-an\* OR "systematic review") AND (exercise OR training OR muscle OR "physical performance") |
| PubMed/MEDLINE | caffeine[tw] AND (meta-an\*[tw] OR "systematic review"[tw]) AND (exercise[tw] OR training[tw] OR muscle[tw] OR "physical performance"[tw]) |
| Scopus | Title-abs-key(caffeine AND (meta-an\* OR "systematic review") AND (exercise OR training OR muscle OR "physical performance")) |
| Web of Science (including Science Citation Index Expanded, Social Sciences Citation Index, and Arts & Humanities Citation Index) | TS=(caffeine AND (meta-an\* OR "systematic review") AND (exercise OR training OR muscle OR "physical performance")) |

**Table S2.** Excluded reviews with the reasons for their exclusion

|  |  |
| --- | --- |
| **Reference** | **Reasons for exclusion** |
| Astorino and Roberson (2010) | No meta-analysis performed.  |
| Brown et al. (2013) | Examined the effects of energy drinks in which both caffeine and taurine was ingested. |
| Doherty and Smith (2005) | Conducted using the same search process as the initial analysis from this group of authors. |
| Ganio et al. (2009) | No meta-analysis performed. |
| Glaister and Gissane (2018) | Focused on physiological responses during exercise and not on exercise performance.  |
| Souza et al. (2017) | Examined the effects of energy drinks in which both caffeine and taurine was ingested. |
| Zhang et al. (2015) | Focused on physiological responses during exercise and not on exercise performance. |

**Table S3.** Summary of subgroup analyses conducted in the included reviews

|  |  |  |
| --- | --- | --- |
| **Reference** | **Subgroup analyses focus** | **Subgroups analyses results\*** |
| Conger et al. (2011) | Timing of caffeine ingestion | Immediately before or during exercise: 0.26 (0.09, 0.42) – 9 studies≥60 min before and during exercise: 0.16 (–0.11, 0.42) – 4 studies30–90 min before exercise: 0.34 (0.16, 0.52) – 9 studies>90 min before exercise: 0.38 (–0.18, 0.95) – 1 study |
| Exercise mode | Cycling: 0.30 (0.18, 0.42) – 18 studiesRunning: 0.08 (–0.15, 0.32) – 3 studies |
| Performance test | Open endpoint: 0.40 (0.21, 0.60) – 7 studiesFixed endpoint: 0.20 (0.08, 0.33) – 14 studies |
| Sustained submaximal exercise bout before performance task | No: 0.29 (0.13, 0.46) – 10 studiesYes: 0.24 (0.08, 0.40) – 11 studies |
| Sex | Men: 0.23 (0.10, 0.37) – 16 studiesMen and women: 0.33 (0.09, 0.58) – 4 studiesWomen: 0.50 (–0.11, 0.11) – 1 study |
| Publication status | Unpublished studies: 0.13 (–0.08, 0.33) – 6 studiesPublished studies: 0.32 (0.19, 0.46) – 15 studies |
| Doherty and Smith (2004) | Exercise protocol | Time-to-exhaustion tests: 0.67 (0.52, 0.81) – 38 effect sizesTime trials: 0.13 (0.02, 0.25) – 27 effect sizesGraded exercise tests: 0.17 (-0.02, 0.36) – 11 effect sizes |
| Training status | Trained: 0.15 (–0.08, 0.38) – 19 effect sizesHighly trained: 0.20 (0.09, 0.31) – 7 effect sizes |
| Grgic and Pickering (2019) | Muscle group  | Knee extensors: 0.19 (0.10, 0.28) – 9 studiesOther muscle groups: 0.10 (-0.02, 0.21) – 8 studies  |
| Angular velocity | Velocity of 30◦s−1: 0.16 (-0.08, 0.39) – 6 studiesVelocity of 60◦s−1: 0.21 (0.07, 0.36) – 3 studiesVelocity of 180◦s−1: 0.23 (0.07, 0.38) – 3 studies  |
| Grgic et al. (2018) – muscular strength  | Muscle group location | Upper body: 0.21 (0.02, 0.39) – 7 studiesLower body: 0.15 (−0.05, 0.34) – 8 studies |
| Caffeine form  | Capsule form of caffeine: 0.27 (0.04, 0.50) – 6 studiesLiquid form of caffeine: 0.11 (−0.17, 0.39) – 3 studies |
| Sex | Males: 0.21 (0.02, 0.41) – 8 studiesFemales: 0.15 (−0.13, 0.43) – 3 studies |
| Training status | Trained participants: 0.18 (−0.02, 0.37) – 7 studiesUntrained participants: 0.27 (−0.09, 0.63) – 4 studies |
| Grgic et al. (2018) – power  | Caffeine form | Capsule form of caffeine: 0.14 (−0.06, 0.34) – 8 studiesLiquid form of caffeine: 0.24 (−0.06, 0.54) – 3 studies |
| Sex | Men: 0.16 (−0.02, 0,34) – 9 studiesWomen: 0.23 (−0.23, 0.69) – 3 studies |
| Training status | Athletes: 0.23 (0.03, 0.42) – 8 studiesNon-athletes: 0.03 (−0.33, 0.40) – 2 studies |
| Exercise test | Countermovement jump: 0.14 (−0.04, 0.32) – 8 studiesSargent test: 0.31 (−0.09, 0.70) – 2 studies |
| Polito et al. (2016) – muscular strength | Muscle group location | Upper-body: 0.08 (−0.09, 0.25) – 4 effect sizes |
| Muscle size | Large: 0.09 (−0.07, 0.25) – 5 effect sizes |
| Sex | Men: 0.09 (−0.07, 0.26) – 4 effect sizes |
| Caffeine form | Capsule: 0.09 (−0.07, 0.26) – 4 effect sizes |
| Caffeine dose | ≤ 4 mg/kg: 0.08 (–0.11, 0.28) – 2 effect sizes6 mg/kg: 0.10 (–0.15, 0.36) – 3 effect sizes |
| Timing of caffeine intake | 45 min: 0.08 (–0.11, 0.28) – 2 effect sizes60 min: 0.10 (–0.15, 0.36 ) – 3 effect sizes |
| Polito et al. (2016) – muscular endurance | Muscle group location | Upper-body: 0.32 (0.19, 0.44) – 24 effect sizesLower-body: 0.42 (0.25, 0.58) – 14 effect sizes |
| Muscle size | Large: 0.38 (0.28, 0.49) – 37 effect sizesSmall: 0.40 (0.11, 0.68) – 5 effect sizes |
| Sex | Men: 0.41 (0.31, 0.51) – 39 effect sizes |
| Caffeine form | Capsule: 0.40 (0.29, 0.51) – 35 effect sizesLiquid: 0.32 (0.10, 0.56) – 7 effect sizes |
| Caffeine dose | ≤ 4 mg/kg: 0.43 (0.20, 0.65) – 11 effect sizes5 mg/kg: 0.44 (0.20, 0.68) – 7 effect sizes6 mg/kg: 0.30 (0.14, 0.47) – 14 effect sizes> 6 mg/kg: 0.51 (0.28, 0.74) – 8 effect sizes |
| Timing of caffeine intake | 45 min: 0.23 (-0.04, 0.49) – 8 effect sizes60 min: 0.42 (0.31, 0.53) – 32 effect sizes90 min: 0.18 (-0.26, 0.63) – 2 effect sizes |
| Warren et al. (2010) – muscular strength\*\*  | Publication status | Published: 0.16 – 22 studiesUnpublished: 0.31 – 5 studies |
| Study design | Crossover: 0.20 – 25 studiesBetween-groups: 0.11 – 2 studies |
| Sex | Men: 0.21 – 19 studiesMen and women: 0.15 – 8 studies |
| Training status | Trained: 0.13 – 6 studiesUntrained: 0.21 – 21 studies |
| Caffeine form | Solid: 0.25 – 18 studiesLiquid: 0.05 – 8 studies |
| Muscle action type | Isokinetic: 0.21 – 6 studiesIsometric: 0.18 – 20 studies |
| Muscle size | Large: 0.31 – 18 studiesSmall: 0.05 – 12 studies |
| Muscle group location | Upper-body: 0.07 – 13 studiesLower-body: 0.29 – 18 studies |
| Muscle group | Knee extensors: 0.40 – 15 studiesKnee flexors: 0.04 – 4 studiesElbow flexors: 0.07 – studies |
| Warren et al. (2010) – muscular endurance\*\* | Publication status | Published: 0.27 – 19 studiesUnpublished: 0.31 – 4 studies |
| Study design | Crossover: 0.26 – 20 studies Between-groups: 0.50 – studies |
| Sex | Men: 0.21 – 15 studiesMen and women: 0.43 – 7 studies |
| Training status | Trained: 0.07 – 6 studiesUntrained: 0.37 – 15 studies |
| Caffeine form | Solid: 0.23 – 15 studies Liquid: 0.39 – 8 studies |
| Muscle action type | Isokinetic: 0.20 – 6 studiesIsometric: 0.36 – 12 studiesIsotonic: 0.16 – 5 studies |
| Exercise test | Open end point: 0.37 – 18 studiesFixed end point: -0.08 – 5 studies |
| Type of load | Constant: 0.33 – 18 studiesVariable: 0.09 – 5 studies |
| Muscle size | Large: 0.23 – 17 studiesSmall: 0.40 – 8 studies |
| Muscle group location | Upper-body: 0.37 – 12 studiesLower-body: 0.25 – 15 studies |
| Muscle group | Knee extensors: 0.33 – 11 studiesKnee flexors: -0.07 – 3 studiesElbow flexors: 0.31 – 4 studiesPectorals/shoulders/triceps: 0.31 – 4 studies Hip and knee extensors: 0.21 – 3 studies |
| \* Presented as mean (95% confidence interval)\*\* Warren et al. (2010) did not present 95% confidence intervals |

**Table S4.** Results of the Grading of Recommendations Assessment, Development and Evaluation (GRADE) assessment

|  |  |  |
| --- | --- | --- |
| **Reference** | **GRADE items** | **Quality of the evidence\*** |
| **Risk of bias** | **Inconsistency** | **Indirectness** | **Imprecision** | **Publication bias** |
| Christensen et al. (2017) | Not serious | Not serious | Serious indirectness (the majority of included studies were conducted in men and, therefore, these results cannot be generalised to women) | Not serious  | Strongly suspected (asymmetry of the funnel plot was not explored and the effect size of the largest study was smaller than the pooled estimate) | Low⊕⊕ΟΟ |
| Conger et al. (2011) | Carbohydrate vs. caffeine + carbohydrate: unclear (no quality assessment performed) | Carbohydrate vs. caffeine + carbohydrate: not serious | Carbohydrate vs. caffeine + carbohydrate: serious indirectness (the majority of included studies were conducted in men and, therefore, these results cannot be generalised to women) | Carbohydrate vs. caffeine + carbohydrate: not serious | Carbohydrate vs. caffeine + carbohydrate: undetected | Carbohydrate vs. caffeine + carbohydrate: low⊕⊕ΟΟ |
| Caffeine vs. placebo: unclear (no quality assessment performed) | Caffeine vs. placebo: not serious | Caffeine vs. placebo: serious indirectness (the majority of included studies were conducted in men and, therefore, these results cannot be generalised to women) | Caffeine vs. placebo: not serious | Caffeine vs. placebo: undetected | Caffeine vs. placebo: low⊕⊕ΟΟ |
| Doherty and Smith (2004) | Aerobic exercise: unclear (no quality assessment performed) | Aerobic exercise: not serious | Aerobic exercise: serious indirectness (the majority of included studies were conducted in men and, therefore, these results cannot be generalised to women) | Aerobic exercise: not serious | Aerobic exercise: undetected | Aerobic exercise: low⊕⊕ΟΟ |
| Graded exercise tests: unclear (no quality assessment performed) | Aerobic endurance as assessed by graded exercise tests: not serious | Aerobic endurance as assessed by graded exercise tests: serious indirectness (the majority of included studies were conducted in men and, therefore, these results cannot be generalised to women) | Aerobic endurance as assessed by graded exercise tests: serious limitation | Aerobic endurance as assessed by graded exercise tests: undetected | Aerobic endurance as assessed by graded exercise tests: very low⊕ΟΟΟ |
| Short-term high-intensity exercise: unclear (no quality assessment performed) | Short-term high-intensity exercise: not serious | Short-term high-intensity exercise: serious indirectness (the majority of included studies were conducted in men and, therefore, these results cannot be generalised to women) | Short-term high-intensity exercise: not serious | Short-term high-intensity exercise: undetected  | Short-term high-intensity exercise: low⊕⊕ΟΟ |
| Gonçalves Ribeiro et al. (2017) | Time-trial duration: serious limitation (the majority of included studies received “unclear risk of bias” on random sequence generation, allocation concealment, and on the blinding of outcome assessors) | Time-trial duration: not serious | Time-trial duration: serious indirectness (all of the included studies were conducted in men and, therefore, these results cannot be generalised to women) | Time-trial duration: not serious | Time-trial duration: undetected | Time-trial duration: low⊕⊕ΟΟ |
| Power: serious limitation (the majority of included studies received “unclear risk of bias” on random sequence generation, allocation concealment, and on the blinding of outcome assessors) | Power: not serious | Power: serious indirectness (all of the included studies were conducted in men and, therefore, these results cannot be generalised to women) | Power: serious limitation | Power: strongly suspected (asymmetry of the funnel plot was not explored and the effect size of the largest study was smaller than the pooled estimate) | Power: very low⊕ΟΟΟ |
| Running distance: serious limitation (the majority of included studies received “unclear risk of bias” on random sequence generation, allocation concealment, and on the blinding of outcome assessors) | Running distance: not serious | Running distance: serious indirectness (all of the included studies were conducted in men and, therefore, these results cannot be generalised to women) | Running distance: serious limitation | Running distance: strongly suspected (asymmetry of the funnel plot was not explored and the effect size of the largest study was smaller than the pooled estimate) | Running distance: very low⊕ΟΟΟ |
| Grgic (2018) | Peak power: not serious | Peak power: not serious | Peak power: serious indirectness (the majority of included studies were conducted in men and, therefore, these results cannot be generalised to women) | Peak power: not serious | Peak power: undetected  | Peak power: moderate⊕⊕⊕Ο |
| Mean power: not serious | Mean power: not serious | Mean power: serious indirectness (the majority of included studies were conducted in men and, therefore, these results cannot be generalised to women) | Mean power: not serious | Mean power: undetected | Mean power: moderate⊕⊕⊕Ο |
| Grgic and Pickering (2019) | Not serious | Not serious | Serious indirectness (the majority of included studies were conducted in men and, therefore, these results cannot be generalised to women) | Not serious | Undetected | Moderate⊕⊕⊕Ο |
| Grgic et al. (2018) | 1RM: not serious | 1RM: not serious | 1RM: serious indirectness (the majority of included studies were conducted in men and, therefore, these results cannot be generalised to women) | 1RM: not serious | 1RM: undetected | 1RM: moderate⊕⊕⊕Ο |
| Vertical jump: not serious | Vertical jump: not serious | Vertical jump: serious indirectness (the majority of included studies were conducted in men and, therefore, these results cannot be generalised to women) | Vertical jump: not serious | Vertical jump: undetected  | Vertical jump: moderate⊕⊕⊕Ο |
| Polito et al. (2016) | 1RM: not serious | 1RM: not serious | 1RM: serious indirectness (the majority of included studies were conducted in men and, therefore, these results cannot be generalised to women) | 1RM: serious limitation | 1RM: undetected | 1RM: low⊕⊕ΟΟ |
| Muscular endurance: not serious | Muscular endurance: not serious | Muscular endurance: serious indirectness (the majority of included studies were conducted in men and, therefore, these results cannot be generalised to women) | Muscular endurance: not serious | Muscular endurance: undetected | Muscular endurance: moderate⊕⊕⊕Ο |
| Shen et al. (2019) | Not serious  | Not serious  | Serious indirectness (the majority of included studies were conducted in men and, therefore, these results cannot be generalised to women) | Not serious | Undetected  | Moderate⊕⊕⊕Ο |
| Southward et al. (2018) | Time-trial time: not serious | Time-trial time: not serious | Time-trial duration: serious indirectness (the majority of included studies were conducted in men and, therefore, these results cannot be generalised to women) | Time-trial duration: not serious | Time-trial duration: undetected | Time-trial duration: moderate⊕⊕⊕Ο |
| Time-trial power: not serious | Time-trial power: not serious | Time-trial power: serious indirectness (the majority of included studies were conducted in men and, therefore, these results cannot be generalised to women) | Time-trial power: not serious | Time-trial power: undetected | Time-trial power: moderate⊕⊕⊕Ο |
| Warren et al. (2010) | MVC: not serious | MVC: not serious | MVC: serious indirectness (the majority of included studies were conducted in men and, therefore, these results cannot be generalised to women) | MVC: not serious | MVC: undetected | MVC: moderate⊕⊕⊕Ο |
| Muscular endurance: not serious | Muscular endurance: not serious | Muscular endurance: serious indirectness (the majority of included studies were conducted in men and, therefore, these results cannot be generalised to women) | Muscular endurance: not serious | Muscular endurance: undetected | Muscular endurance: moderate⊕⊕⊕Ο |
| *1RM*: one repetition maximum test; *MVC*: maximal voluntary contraction; \* classification based on the GRADE Handbook as: ⊕⊕⊕⊕ = high quality⊕⊕⊕Ο = moderate quality⊕⊕ΟΟ = low quality⊕ΟΟΟ = very low quality |

**Table S5**. Summary of the included meta-analyses based on the quality of the review, quality of evidence, and the 95% prediction interval categories

|  |  |
| --- | --- |
| **Quality of evidence and prediction interval (PI) categories** | **Quality of the review** |
| Moderate | High |
| Quality of evidence: “very low”95% PI includes zero | *14% of the included meta-analyses** Aerobic endurance as assessed by graded exercise tests in Doherty and Smith (2004)
* Aerobic endurance as assessed by running distance in Gonçalves Ribeiro et al. (2017)
* Anaerobic power in Gonçalves Ribeiro et al. (2017)
 | / |
| Quality of evidence: “very low”95% PI does not include zero | / | / |
| Quality of evidence: “low”95% PI includes zero | *19% of the included meta-analyses** Aerobic endurance in the carbohydrate vs. caffeine + carbohydrate comparison in Conger et al. (2011)
* Aerobic endurance in the caffeine vs. placebo comparison in Conger et al. (2011)
* Muscle strength in Polito et al. (2016)
* Short-term high-intensity exercise in Doherty and Smith (2004)
 | / |
| Quality of evidence: “low”95% PI does not include zero | *14% of the included meta-analyses** Aerobic endurance as assessed by time trial duration in Gonçalves Ribeiro et al. (2017)
* Aerobic endurance in Doherty and Smith (2004)
* Exercise speed in Christensen et al. (2017)
 | / |
| Quality of evidence: “moderate”95% PI includes zero | *24% of the included meta-analyses** Muscle endurance in Warren et al. (2010)
* Muscle strength in Grgic and Pickering (2019)
* Muscle strength in Warren et al. (2010)
* Peak anaerobic power in Grgic (2018)
* Vertical jump in Grgic et al. (2018)
 | / |
| Quality of evidence: “moderate”95% PI does not include zero | *14% of the included meta-analyses** Time-trial time in Southward et al. (2018)
* Time-trial power in Southward et al. (2018)
* Mean anaerobic power in Grgic (2018)
 | *14% of the included meta-analyses** Muscle strength in Grgic et al. (2018)
* Muscle endurance in Polito et al. (2016)
* Aerobic endurance in Shen et al. (2019)
 |
| Note: Quality of systematic review was assessed using the AMSTAR 2 checklist (none of the reviews were categorised as “low” quality); Quality of evidence was assessed using the GRADE criteria (none of the meta-analyses provided “high” quality of evidence) |