CPQ Orthopaedics (2021) 5:3
Systematic Review

A Systematic Review of Return to Sport Physical Performance Tests of the Foot and Ankle

Michael Jeanfavre[1*], Amy Humphrey[2] & Matthew Klein[3]

[1]*Department of Physical Therapy, Azusa Pacific University, USA, OP Ortho & Sports Medicine Rehab, Stanford Health Care, USA*

[2]*Department of Physical Therapy, Azusa Pacific University, USA, Doctor of Physical Therapy Program, Messiah University, USA*

[3]*Department of Physical Therapy, Azusa Pacific University, USA, Physical Therapy Assistant Program, Stanbridge University, USA*

[*]**Correspondence to:** Dr. Michael Jeanfavre, Department of Physical Therapy, Azusa Pacific University, USA, OP Ortho & Sports Medicine Rehab, Stanford Health Care, USA.

Copyright

Received: 31 March 2021

Published: 16 April 2021

Keywords: *Physical Performance Tests; Return to Sport; Foot and Ankle Injury; Systematic Review*

Abstract

Objective

This systematic review sought to update a prior review and identify the current physical performance tests (PPT) that are used to determine return to sport (RTS) readiness in competitive athletes following musculoskeletal foot and ankle pathology.

Methods

The PubMed, CINHAL, and SPORTDiscus databases were systematically searched for articles

using keywords related to PPT, RTS, and foot and ankle injuries. The quality of the included studies was evaluated using the risk of bias (RoB) tools proposed by the CLARITY from the McMaster University of Health Sciences.

Results

Twenty-five articles (9 RCTs, 16 Non-RCTs) with 1372 subjects were included for the final analysis. Twelve PPTs were classified as muscle performance, postural control, or functional tests. The Star Excursion Balance Test, Side Hop Test, and Square Hop Test showed good reliability, agreement, and responsiveness when assessing athletes with foot and ankle pathologies. The quality assessment determined that <45% of RCTs failed to adequately blind participants, personnel, and/or assessors. While <40% of Non-RCTs clearly stated assessment exposure, the absence of the outcome of interest, and assessment of the presence of prognostic factors.

Conclusions

Though additional studies regarding RTS PPTs for the foot and ankle have been published, there remains a need for higher quality studies evaluating the psychometric properties of the PPTs in injured athlete populations.

Abbreviations

ANT = anterior
AOFAS = American Orthopedic Foot and Ankle Society Ankle-Hindfoot Score
BAPS = biomechanical ankle platform system
BESS = Balance Error Scoring System Test
CAI = chronic ankle instability
CAIT = Cumberland Ankle Instability Tool
CKC = closed kinetic chain
CLARITY = Clinical Advances Through Research and Information Translation
COSMIN = Consensus-based Standards for the selection of health Measurement Instruments
F = females
FAAM = Foot and Ankle Ability Measure
FAI = functional ankle instability
FAOS = Foot and Ankle Outcome Score
FTB = Functional test battery
HHD = hand-held dynamometer
Hx = history
IAC = International Ankle Consortium
ICCs = intraclass correlations
IdFAI = Identification of Functional Ankle Instability questionnaire

IKDC = international knee documentation committee subjective evaluation form
JPS = joint positional sense
KT = kinesio-tape
LAS = lateral ankle sprain
LE-YBT = Lower Extremity Y-Balance Test
LSI = limb symmetry index
M = males
MDC = minimal detectable change
MIC = miniml important difference
MMT = manual muscle testing
mo. = month
OKC = open kinetic chain
PICOT = Population, Intervention, Comparison, Outcome, Time
PL = posterior lateral
PM = posterior-medial
PPT = Physical Performance Test
RCT = randomized control trial
RoB = risk of bias assessment
RTS = Return to sport
SEBT = star excursion balance test
SEM = standard error of measure
SLB = single leg balance
sec = seconds
SL = single limb
TCJ = talocrural joint
Wk = week(s)
YBT = lower quarter y-balance test
y/a = years of age
Yr = year(s)

Introduction

Research shows that a high prevalence of foot and ankle injuries exist in sports activity, particularly ankle sprains, chronic ankle instability, and Achilles tendinitis/osis [1]. Occurrence rate estimates are around 10-30%, and in certain sports, particularly football, indoor volleyball, netball, and field events in track and field, this percentage is even higher [1,2]. Acute and chronic injury to the foot or ankle often limits an athlete's ability to run, jump, kick, or change direction, which can ultimately hamper participation in sports activity [3]. Given the physical limitations that can occur from the above injuries, adequate physical performance metrics should be utilized to assess for and make decisions on return to sport (RTS) readiness.

Physical performance tests (PPTs) are tools that qualify and quantify function and assist in the clinical RTS decision-making process [4,5]. Often these tests are used by healthcare professionals to determine when an athlete can safely return to sport following surgery or injury. Extensive literature on PPTs for knee and hip

rehabilitation and RTS has been published. A majority have focused on RTS criteria for anterior cruciate ligament (ACL) injuries, with the hop tests being the most utilized measures [6]. Specifically, these tests include the single-leg hop for distance, medial hop, triple hop, 6m timed hop, crossover hop, single-leg vertical hop[4,5]. Current evidence has begun to emphasize using a battery of tests along with functional testing algorithms for determining RTS readiness [7]. However, further work is still needed to validate whether these tests accurately determine RTS readiness [8]. While the tests performed for the hip and knee all involve the ankle, no literature has specifically addressed whether the tests are also valid for return to sport assessments in those with ankle injuries. Systematic reviews on RTS PPT have been done for the hip, specific foot and ankle conditions, and the entire lower extremity, but conclusions have remained the same that further research continues to be needed to establish appropriate reliability and validity for PPTs [4,5,9-12].

Despite the amount of literature on lower extremity PPTs, there is a dearth of information on each measure's standards for RTS with a foot and ankle musculoskeletal injury. Though the foot and ankle complex's physical demand requirements may vary from one sport to the next, foundational lower leg physical performance competencies and capacities are required across many sports [13]. Thus, similar PPT performance may be used as criteria for safe RTS activity regardless of the type of injury in similar weight-bearing sports. Though information on standardized predictive assessments and RTS is available for specific pathologies, such as lateral ankle sprains (LAS) and mid-portion Achilles tendinopathy, there remains a lack of consensus on RTS criteria [9,11,14]. Lack of agreement is especially concerning given the well-known high rates of reinjury, reduced percentage of individuals that return to their prior level of competition, and effects on long-term health and quality of life [7,15]. Due to the high rate of reinjury, which can be as high as 61% in some athletic populations for acute lateral ankle sprains, additional information is needed to guide clinicians on the appropriate use of PPTs to determine readiness for RTS post foot and ankle musculoskeletal injury [15-17].

The the current systematic review's primary purpose was to answer the following question: 'What are the current clinically applicable RTS PPTs to determine readiness in competitive athletes, ages 12 to 65, following musculoskeletal foot and ankle pathology?' The study question was framed using the PICOT format. The PICOT question variables, study elements, respective inclusion and exclusion criteria are shown in Table 1. The secondary purpose was to ensure ease of clinical application of the results by proposing a RTS functional test battery (FTB).

Table 1 & Figure 1: *PICOT Question and Study Design Inclusion and Exclusion Criteria*

Question Component	Inclusion Criteria	Exclusion Criteria
Population	• Adolescent (12-18 y/a) • Adults (18 - 65 y/a) • Musculoskeletal injuries of the foot and ankle • Humans • Returning to any level of sport, physical activity, or military duty	• Subsequent neurological or neuromuscular compromise or condition • Concomitant other musculoskeletal pathology beyond the foot and ankle • Healthy individuals (absence of foot or ankle injuries; this includes cohorts that were healthy at time 0 and became injured across the observational period)

Intervention	• Non-operative or conservation treatment • Surgery • A combination of the above	N/A
Comparison	• None	N/A
Outcome	• Clinically applicable physical performance test	• Laboratory required physical performance test • Expensive Laboratory Equipment (defined as a capital expense of >$1000) • PPT not specified or used in RTS eligibility
Study Design	• Randomized control trials (RCTs) • Prospective observational studies (Non-RCTs) • Hand-selected articles meeting the inclusion and exclusion criteria and identified through a review of article references or other resources	• Studies not written or translated into the English language • Case reports • Editorials • Letters • Conference or Poster Abstracts • Grey literature
Time	• After 2015	• Before 2015

Note: N/A, indicates information not applicable; PPT, physical performance tests; RCTs, randomized control trial; y/a, years of age.

Methods

A comprehensive systematic review of RTS PPTs for the hip and ankle was previously published [5]. The current investigation is not solely an update nor builds solely upon the previous work by Hegedus *et al.* (2015b) [5] as it focuses solely on the foot and ankle. Additional articles were also identified during the article search utilizing the strategy below that were not included in the prior work. The current investigation sought to identify RTS PPTs for the foot and ankle that have been published since 2015. An update to Hegedus's (2015b) [5] prior results was indicated due to: sufficient time elapsed, new evidence becoming available, and based on need or priority [18-20].

Eligibility Criteria

The inclusion and exclusion criteria for each of the PICOT question components are outlined in Table 1. The operational definition of PPT utilized by this review is defined as "a single test that attempts to measure constructs related to sports (strength, postural control, power, and agility)"[5]. The operational definition utilized by this review of the foot and ankle is defined as any anatomical structure at the level of and distal to the syndesmosis of the tibiofibular joint. The author's definition of RTS was defined as a Tegner Level 5 or above (i.e., recreational level of sport or higher) [21]. The justification for the selected age ranges from 12 to 65 years of age was (1) to capture studies that investigated or used PPT in high school populations and (2) the upper bound of age 65 was selected to be sure to capture senior athletes. A broad age range was selected to be as comprehensive as possible and capture PPT that would be pertinent across the life span. However, the limit at 65 was placed secondary to the age of 65 being defined as elderly (Tanaka, 2012).

Search Strategy

Relevant articles were identified by searching PubMed, CINHAL, and SportsDiscuss databases. The strategy's derivation was based on previous reviews [4,5]. Furthermore, it was audited by a senior author to ensure the appropriate use of Boolean modifiers, accurate translation of the search strategy across databases, and appropriateness of the search based upon the study's stated purpose. The intended search strategy for PubMed with the respective results is shown in Figure 1. The keywords used were variations and derivatives of: "return to sport," "musculoskeletal injuries," and "foot and ankle." Keywords of PPTs were not included in the search strategy due to artificially limiting the number of articles identified in preliminary database searches. The search strategies used for CINHAL and SPORTDiscus are shown in Appendix A.

Study Selection

Search results of the different databases were combined, duplicates deleted and filtered independently according to the specified inclusion and exclusion criteria by two research team members (AH, MJ) using a citation manager (Calarvate Analytics, EndNote, X9.2, Zotero). Discrepancies in filtering the search results were discussed by the two independent reviewers (AH, MJ). Discrepancies of the included article(s) that could not be resolved through discussion of the two reviewers were addressed by a *priori* identified third member of the research team (MK). Figure 2 outlines the study selection process in a PRISMA flow diagram.

Table 2 & Figure 2: *Summary of Included Studies*

First Author (Year)	Purpose	Population	Injured (I) Post-op (P)	Sport or Sports	Test(s) and Description	Results
Alves (2018) [22]	Compare Mulligan taping vs. placebo taping on static balance, lower extremity functional performance and latency time of the peroneus longus	16 subjects (10 M, 6 F), 21.5 ± 2.8 years. History (Hx) of at least one ankle sprain within 12 months, at least 1 interrupted day of activity, Cumberland Ankle Instability Tool (CAIT) < 25, Hx of giving way and/or instability	Chronic condition	Semi-professional basketball players	Motion Analysis of Single Leg Balance Eyes Closed, Figure-8 Hop Test, Lateral Hop Test, YoYo IRB Test, Peroneus Longus EMG Trap Door Reaction	No difference between Mulligan and placebo taping on postural control, Hop Tests or other lower extremity functional tests. Mulligan tape may decrease Peroneus Longus latency compared to placebo after running
Anguish (2018) [23]	Compare effects of hop to stabilization balance w/ SLB program on self-reported function, dynamic postural control, and proprioception (JPS)	18 subjects (2 F, 16 M), 18.38±1.81 years. Hx of CAI, ≥ 1 sprain/yr, with initial sprain >1 yr prior, reporting functional deficits at time of study, not currently injured	Chronic condition. Not currently injured	High school (14) and college (4) scholastic or recreational athletes (sports not specified)	SEBT– order randomized disability questionnaires -FAAM	Both groups improved SEBT

Bagherian (2018) [24]	Quantify lower extremity functional movement scores using Fusionetics scoring before/after fatigue. 2nd: quantify SEBT scores before/after fatigue in those with CAI	40 M collegiate athletes w/ CAI. Hx of moderate to severe unilateral ankle sprain >7 days sports time loss within last 5 yr, hx of 2 giving way episodes in last 12 mo, score < 90% on FAAM and <80% on FAAM Sport, no ankle sprains within 6 wk of study	Chronic condition. Not currently injured	Collegiate athletes (not specified)	SEBT on involved limb only in anterior, posterior and medial directions. Average of three trials.	Patients with CAI performed worse on SEBT post fatigue protocol
Bagherian (2019) [25]	Effect of corrective exercises on functional movement patterns sensorimotor function, reported function and fatigue sensitivity in college athletes with CAI	40 M college athletes, 18-35 y/a. Train x3/wk, hx of moderate/severe unilateral ankle sprain within last 5 yr, ≥2 giving way episodes in the last 12 mo	Chronic condition. Not currently injured	Intercollegiate sports (not specified)	SEBT w/ foot scanner, double limb squat, double limb squat with heel lift, single limb squat, ankle dorsiflexion, Biodex isokinetic dynamometer	Corrective exercises improve movement efficiency, sensorimotor function and self-reported functional. No change in fatigue
Best (2015) [26]	Evaluate early outcome of patients with acute ankle sprain w/ variable orthosis.	47 subjects, 16-50 y/a with acute LAS Grade II or higher with 77 healthy subjects as reference/controls	Injured	Not specified	FAOS, AOFAS, static balance test on movable balance platform, vertical drop jump, shuttle run and zig zag run	Marginal difference between various orthosis.
Cain (2017) [27]	Determine effectiveness of 4-wk BAPS protocol on balance of high school athletes with CAI	11 subjects with CAI (4 M, 7 F), 11 control (7 M, 4 F). high school students, no acute injuries. Chronic conditions only. 2 + moderate ankle sprains requiring medical attention, episodes of giving way	Chronic conditions. Not currently injured	High school athletes (not specified)	SEBT, Foot Lift Test, Single Leg Balance Test Eyes Closed, Side Hop Test	Improvement in all measures after 4 week of training
Cho (2019) [28]	Determine the effect of minimally invasive suture-tape augmentation on FAI	24 (9 M, 15 F) subjects with FAI, 2 episodes giving way, repeated ankle sprains, failed rehab, <2 points on CAIT	Post-operative	Not specified	Modified Rhomberg, Biodex II isokinetic dynamometer, CAIT, FAAM	Improvements in balance time

Coetzee (2019)	Brostrom repair and internal brace would accelerate rehab and return to activity in those with lateral ligament repair for CAI	81 subjects (30 M, 51 F)	Post-operative	Not listed, 68 returned to sport, 8 did not participate in sport prior	Single Leg Hop for Distance with Limb Symmetry Index, FAAM, AOFAS, Ankle Dorsiflexion, Calf Girth	Mean return to sport time was 84 days, athletes able to return to play earlier
Cruz-Diaz (2014) [29]	Determine the effect of 6 wk balance training program on patients with CAI	70 athletes with reported instability over 6 mo, no hx of lower extremity injuries or neuromuscular deficits	Chronic conditions. Not currently injured	Not specified	CAIT, SEBT	Large effect sizes in CAIT, SEBT PM and PL directions
Ferey-dounnia (2019) [30]	Assess impact of KT tape peroneus longus/glute med on dynamic balance, muscle strength, functional performance	30 subjects with 15 with FAI and 15 control from semi-pro male soccer team. Chronic symptoms with ≥1 LAS in last 6-12 mo, 2 episodes giving way, decreased functional due to hx of ankle sprains	Chronic conditions. Not currently injured	Soccer	Side Hop Test, SEBT, Figure 8 Hop Test	No significant differences between groups in Functional Performance Tests. However, KT tape improved performance in side hop, SEBT in ANT & PL directions immediately post
Golditz (2016) [31]	Explore potential associations between outcomes of different subject and objective assessments in a population of athletes with or without FAI	29 athletes with a history of FAI from previous study. 13 copers (10 M, 3 F) and 16 FAI (11 M, 5 F) with no current acute injuries	Chronic conditions. Not currently injured	Handball, Volleyball, Basketball, Soccer	SEBT, Isokinetic Dynamometer for JPS, Time to Stabilization Test	No associations between self-reported ankle function or sensorimotor tests/ MRI. Individuals with FAI, early degenerative changes related to reduced sensorimotor control
Hall (2018) [32]	Determine if balance/strength training protocols could improve strength/ balance and f performance deficits in CAI	39 subjects (21 M, 18 F), ≥1 substantial ankle sprain, 1 interrupted day of activity, multiple giving way episodes and feelings of instability 6 mo before study.	Chronic conditions. Not currently injured	Not Specified	SEBT, Isokinetic Strength Testing, BESS, Side Hop Test	Improvements in SEBT, BESS and Side Hop Tests

Harriss (2019) [33]	Determine if movement quality differs between collegiate athletes with/without CAI	99 division 1 athletes. (49 CAI (20 M, 29 F), 50 control (26 M, 24 F). ≥1 major episode at least 12 mo prior to study, 2 episodes of giving way 6 mo prior, >10 on IAC	Chronic conditions. Not currently injured	Women's crew, men's and women's soccer, men's and women's lacrosse, field hockey	LE Fusionetics Score, Landing Error Scoring System-17	Less-17 and Fusionetics did not differ between those with and without CAI. Abnormal trunk responses in CAI.
Kamali (2017) [34]	Evaluate effect of TCJ manipulation on athletes with CAI	40 athletes (18 M, 22 F). ≥1 LAS within last 6 wk or multiple episodes in last 12 mo. Must be able to perform 24m running test and have at least 80% strength of compared to healthy limb	Injured and chronic	Soccer, volleyball, basketball, martial arts	Single Leg Hop Test, YBT	Improvement in all Single Leg Hop Test, YBT following TCJ manipulation
Ko (2018) [35]	Determine if two common functional performance tests could identify functional performance deficits and how they relate to number of reported ankle sprains.	58 adolescent subjects (30 M, 28 F). 24 injured, 34 uninjured. No acute or injured. ≥1 major significant ankle sprain, no occurrence within prior 3 mo.	Chronic conditions. Not currently injured	Soccer	SEBT, Single Leg Hop Test	Adolescents with history of LAS demo decreased motion in all three direction on SEBT and decreased dynamic postural stability during Single Leg Hop Test
Madsen (2018) [36]	Identify functional performance tests that are sensitive to subjective and objective CAI deficits.	48 subjects. 24 with CAI (10 M, 14 F), 24 healthy (10 M, 14 F). 1 limb score at ≥ 11 score on IdFAI, Contralateral limb having no hx of instability or giving way, last ankle sprain occurred ≥ 3 mo prior to study	Chronic conditions. Not currently injured	Not Specified	Side Hop, 6-Meter Cross-Over Hop, Lateral Hop, & Figure 8 Hop Tests	Subjects with CAI perceive more instability with functional performance tests, however produce similar outcomes compared to healthy controls
McCann (2017)	Examine structural and functional impairments/activity limitations in athletes with acute lateral ankle sprain at return to play	50 patients (15 F, 35 M) high school and collegiate with potential injuries. Evaluated by AT, loss of least 1 day of activity.	Potential injury.	High school and collegiate athletes Not specified	SEBT-Anterior (normalized for leg length), Weight Bearing Lunge Test	Athletes with acute LAS continue to demonstrate deficits at return to play include ankle range of motion, joint laxity, and dynamic postural control

Powden (2019) [37]	Examine effect of 4-wk rehab program on common CAI impairments	20 participants (5 M, 15 F), 24.35 ± 6.95 y/a. Not injured. Excluded if ankle injury 6 wk prior. Inclusion: hx of ≥ 1 ankle sprain ≥ 6 mo prior to study, ≥ 2 episodes of giving way 3 mo prior.	Chronic conditions. Not currently injured	Not specified	YBT, Isometric Dynamometer (strength), Single Limb Stance Eyes Open/Closed on Force Plate, Weight Bearing Lunge Test	Improvements in range of motion, isometric strength, postural control and self-reported function
Ryu (2019) [38]	Assess correlation between YBT and ankle injury. Assess differences in YBT between different baseball positions	42 M baseball players from Korea baseball organization (one team) currently active on team. Those with current injuries or inability to perform YBT excluded.	Chronic conditions. Not currently injured	Baseball	YBT	YBT ANT position differences may reflect injury status in baseball. Different positions may have different levels of ankle stability.
Sierra-Guzmán (2018) [39]	Analyze peroneal reaction time, dynamic balance, and strength in those with CAI vs healthy.	105 recreational athletes (50 CAI, 55 healthy). Acute injuries excluded. History of at least 1 significant ankle sprain (>3 mo prior), ≥ 2 episodes of ankle giving wait, ≤ 24 on CAIT	Chronic conditions. Not currently injured	Recreational athletes (not listed)	SEBT, Balance on Biodex Stability System, Surface EMG, Isokinetic Strength Test	Greatest deficits in peroneal reaction time, dynamic balance, particularly of PM & PL directions of SEBT
Sierra-Guzmán (2018) [40]	Evaluate 6 wks of whole-body vibration on balance and body composition in recreational athletes with CAI	50 recreational athletes into three groups (vibration, non-vibration, control). Acute injuries excluded. Hx of ≥1 significant ankle sprain (>3 mo prior), ≥ 2 episodes of ankle giving wait, ≤ 24 on CAIT	Chronic conditions. Not currently injured	Not Listed	Biodex, SEBT	Improvements in balance with balance training with or without whole body vibration.
Someeh (2015) [41]	Determine if fibular reposition tape effects postural control in those with or without CAI	32 subjects 16 professional athletes with CAI, 16 pros healthy. Hx of ≥1 significant ankle sprain (within 6 mo prior), ≥ 2 episodes of ankle giving way. Acute injuries excluded	Chronic conditions. Not currently injured	Professional athletes in football, volleyball, and handball	SEBT	Improved acute postural control in both healthy and CAI athletes with taping.

Son (2017)	Examine walking neuromechanics between those with and without CAI	200 individuals (100 CAI, 100 control, 104 M, 96 F). Hx of LAS or ankle giving way in last 6 mo. Acute injuries excluded	Chronic conditions. Not currently injured	Not specified	Walking Gait Video Analysis	CAI group showed hip-dominant strategy compared to controls.
Toyooka (2017) [42]	Test whether 1-time heel raise and FWB 20-time heel raise are predictors of return to sport in those with acute Achilles tendon ruptures	96 patients post-Achilles tendon rupture and repair (23 F, 73 M). Post-Surgical Triple Bundle Technique by Uchiyama *et al.*	Post-operative	Badminton, Soccer, Volleyball	1-Time Full Body Weight Heel Raise Test, 20-Time Full Body Weight Heel Raise Test	Full body weight 1-time heel raise correlated to jogging capability and 20-time full body weight heel raise predict return to sport/daily life
Toyooka (2018) [43]	Evaluate relationship between single-limb stance with closed eyes and subjective function, instability, and ankle function	103 high school basketball players. No lower extremity injury in last 6 mo. Lifetime ankle sprains determined grouping (5 groups)	Chronic conditions. Not currently injured	Basketball	Win-pod platform single leg balance eyes closed	Little relationship between center of pressure analysis and subjective ankle function. May not accurately reflect function

Note: ANT, anterior; AOFAS, American Orthopedic Foot and Ankle Society Ankle-Hindfoot Score; BAPS, biomechanical ankle platform system; BESS, Balance Error Scoring System Test; CAI, chronic ankle instability; CAIT, Cumberland Ankle Instability Tool; F, females; FAAM, Foot and Ankle Ability Measure; FAI, functional ankle instability; FAOS, Foot and Ankle Outcome Score; Hx, history; IAC, International Ankle Consortium; IdFAI, Identification of Functional Ankle Instability questionnaire JPS, joint positional sense; KT, kinesiotape; LAS, lateral ankle sprain, M, males; mo., month; PL, posterior lateral; PM, posterior-medial; SEBT, star excursion balance test; SLB, single leg balance; wk; week(s); TCJ, talocrural joint; YBT, lower quarter y-balance test; yr, year(s); y/a, years of age.

Data Extraction

Data elements of identified full-text articles were prospectively determined based upon the PICOT question, the primary and secondary purposes of the current study, and examination of reviews previously published related to this topic [4,5]. These included: author, year, study design, sample size, subject demographic data, medical diagnosis(es), type and level of sport of subjects, clinically feasible PPT, information necessary for conducting quality and risk of bias assessments, and psychometric properties of PPT (reliability, agreement, hypothesis testing, responsiveness, criterion validity, etc.). The specific data elements were extracted by a member of the research team (MK), and all elements were double-checked by two other members (MJ and AH). A pre-piloted data collection sheet was used to collect the extracted study elements. Corresponding authors of primary studies were contacted in the case of missing data.

Summary Measures and Synthesis of Results

PPTs were categorized after data extraction as either a muscle performance, postural control, or functional test to clarify constructs measured and for ease of application by clinicians and healthcare providers. Additionally, the PPTs identified were summarized into a clinically recommended testing battery, consistent with proposed RTS decision-making models previously described for other body regions [7,44,45].

Risk of Bias Assessment

Consistent with the Cochrane Handbook [46], the risk of bias and quality appraisal of the included RCTs and non-RCTs were assessed. The risk of bias assessment (RoB) of included studies was performed using the respective RoB tools for RCTs and cohort studies developed by the CLARITY (Clinical Advances Through Research and Information Translation) from the McMaster University of Health Sciences [47,48]. The CLARITY RoB tool differs from the COSMIN (Consensus-based Standards for the selection of health Measurement Instruments) checklist, which is the quality appraisal tool used in previous reviews [4,5]. The justification for changing tools was due to the prior authors' acknowledging in their limitations that the COSMIN's measurement properties are not well understood [4,5,49]. The RoB assessments for RCTs and non-RCTs were performed by two independent research members (MJ and AH, respectively), and the assessment outcomes were audited by a third member of the research team (MK). Any discrepancies identified by the secondary review were clarified by a priori identified third member of the research team.

Results

Study Selection and Characteristics

Of the 119 articles read in total, 25 articles were deemed appropriate for final analysis. Nine were RCTs, five were case-control studies, ten were case series studies, and one study was a cross-sectional study. The following PPTs were identified: Star Excursion Balance Test, Modified Rhomberg Test, Side Hop Test, Foot Lift Test, Single Limb Hop for Distance Test, Balance Error Scoring System Test, Single Limb Heel Raise Test, 6 Meter Crossover Hop Test, Figure 8 Hop Test, Triple Crossover Hop Test, and the Lateral Hop Test. Additionally, through a review of identified articles references, an additional test, the Square Hop Test, was included in our selection [4,50]. For reasons unknown, the Square Hop Test was not included in Hegedus's previous review (2015a) [4]. Due to the Square Hop Test's ability to discriminate between healthy and injured limbs and meet the operational definition of a PPT, this test was also included within the current results. A summary of the test characteristics is provided in Appendix B. Study characteristics included authors, names and alternate names given to the test, the methodology by which the test was performed and scored, the measurement property, and the quality of the measurement property. The description of each of the included studies is provided in Table 2. provided in Appendix B. Study characteristics included authors, name and alternate names given to the test, the methodology by which the test was performed and scored, the measurement property, and the quality of the measurement property. The description of each of the included studies is provided in Table 2.

Risk of Bias Assessment

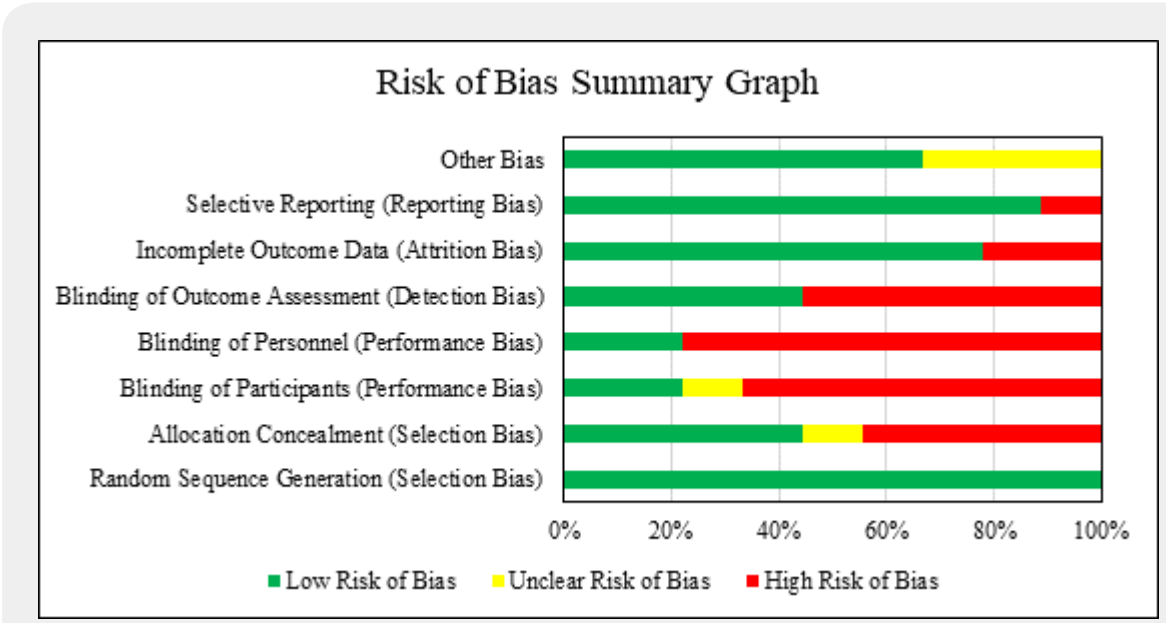
The RoB assessment results for RCTs are summarized in Table 3 and graphical representation of the results are shown in Figure 3. The RoB assessment results for each individual study, RCTs and non-RCTs, are provided in Appendix C. The highest risk of bias was in the blinding of participants, personnel, and outcome assessments. The lack of blinding in rehabilitation and physical therapy literature is well documented and the RoB assessment results in this review further corroborate this limitation [51]. However, a majority of RCTs were deemed to have a low level of selection, reporting and other biases (see Figure 3).

Table 3: *Summary of Risk of Bias Assessment for Randomized Control Trials*

First Author (Year)	Random Sequence Generation (Selection Bias)	Allocation Concealment (Selection Bias)	Blinding of Participants (Performance Bias)	Blinding of Personnel (Performance Bias)	Blinding of Outcome Assessment (Detection Bias)	Incomplete Outcome Data (Attrition Bias)	Selective Reporting (Reporting Bias)	Other Bias
Alves (2018) [22]	+	-	+	-	+	+	-	?
Anguish (2018) [23]	+	+	?	+	-	+	+	+
Bagherian (2019) [25]	+	+	-	-	-	+	+	+
Best (2015) [26]	+	-	-	-	+	-	+	?
Cain (2017) [27]	+	-	-	-	-	-	+	?
Cruz-Diaz (2015)	+	+	-	-	-	+	+	+
Hall (2018) [32]	+	-	-	-	-	+	+	+
Kamali (2017) [34]	+	?	+	+	+	+	+	+
Sierra-Guzmán (2018) [40]	+	+	-	-	+	+	+	+

Note. + = Low risk of bias, ? = Unclear risk of bias, - = High risk of bias

Figure 3: *Graphical Summary of Risk of Bias Assessment for Randomized Control Trials*



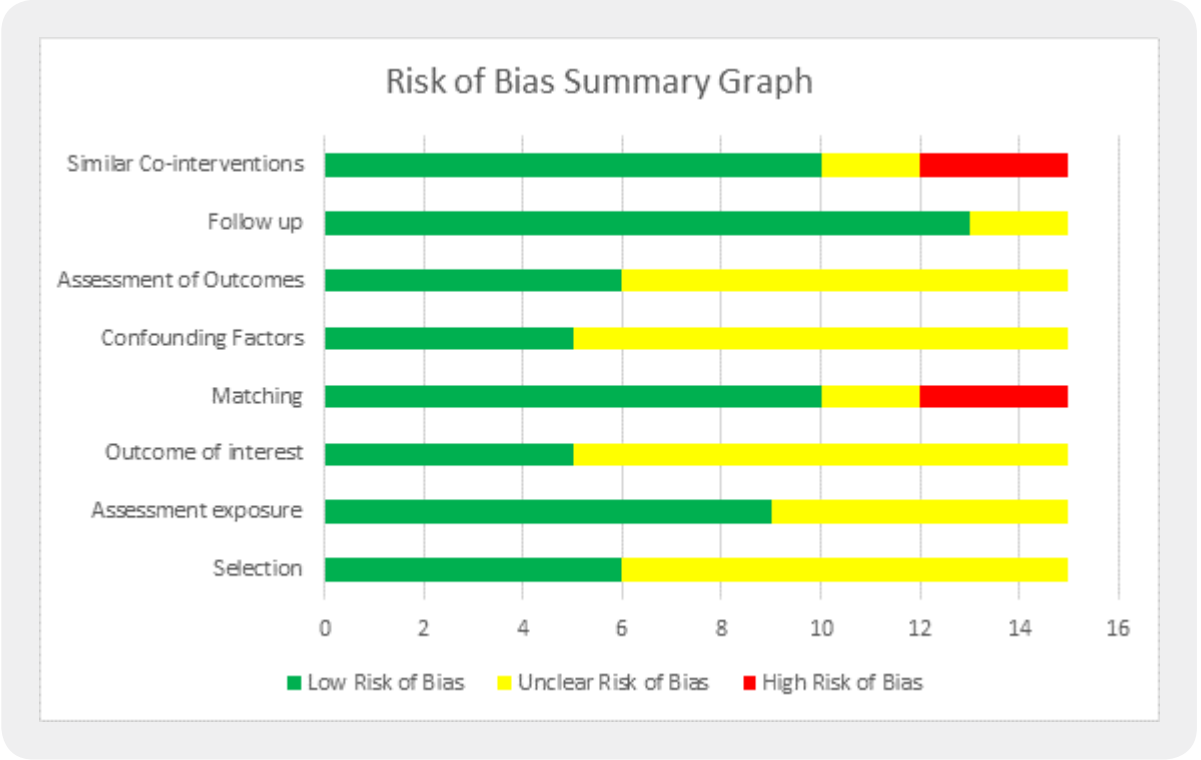
The RoB assessment results for non-RCTs are summarized in Table 4 and graphical representation of the results are shown in Figure 4. The highest RoB for non-RCTs is in the selection process as well as in the assessment of prognostic factors and outcomes. Overall, RoB appeared to be unclear in the majority of these studies. Selection of the cohort, being able to control for con-founding factors, and the inability to follow-up over time are documented limitations that contribute to RoB in cohort studies [52].

Table 4: *Summary of Risk of Bias Assessment for Non-randomized Control Trials*

First Author (Year)	Selection from same population	Assessment exposure	Outcome of interest not present	Match exposed and unexposed	Assessment of presence of prognostic factors	Assessment of outcome	Follow up of cohorts was adequate	Co-interventions similar between groups
Bagherian (2018) [24]	+	?	?	+	+	?	+	+
Cho (2019) [28]	+	+	+	-	?	?	+	-
Coetzee (2018) [53]	?	+	?	?	?	?	?	-
Fereydounnia (2019) [30]	?	?	?	+	?	+	+	+
Golditz (2016) [31]	?	+	+	+	+	?	+	+
Harriss (2019) [33]	+	+	+	+	?	+	+	+
Ko (2018) [42]	?	+	?	+	?	+	?	+
Madsen (2018) [36]	?	+	?	+	+	+	+	+
McCann (2018) [60]	?	?	?	-	?	+	+	?
Powden (2019) [37]	+	?	?	+	+	?	+	+
Ryu (2019) [38]	+	+	?	?	+	?	+	+
Sierra-Guzmán (2018) [39]	?	?	-	+	?	+	+	+
Someeh (2015) [41]	?	+	?	+	?	?	+	+
Toyooka (2017) [42]	?	+	+	-	?	?	+	-
Toyooka (2018) [43]	+	?	?	+	?	?	+	?

Note. + = Low risk of bias, ? = Unclear risk of bias, - = High risk of bias

Figure 4: *Graphical Summary of Risk of Bias Assessment for Non-randomized Control Trials*



Summary Measures and Synthesis of Results

The included studies that described each PPT and the respective reported measurement properties are synthesized in Appendix B. Summary statements that can be made regarding the PPT results are:

➢ Results suggest the 20 Times Single-Leg Heel Raise Test is a valid, reliable, and responsive test to assess readiness for return to sport post-Achilles tendon repair.

➢ There is moderate evidence the Side Hop Test and Square Hop Tests are reliable, valid, and responsive in assessing those with foot and ankle pathology.

➢ There is strong evidence that the Figure 8 Hop Test, Triple Crossover Hop Test, Lateral Hop for Distance are not capable of differentiating between healthy feet/ankles and those with chronic pathology.

➢ There is moderate evidence the SEBT is valid, responsive and reliable in assessing those with foot/ankle pathologies. This may further suggest that balance and proprioception are essential to assess in this population.

Quality of Statistical Properties of Identified Tests

The summary of statistical properties can be found in Table 5. Reporting and ratings of these properties varied somewhat, with gaps present among all the PPTs.

Table 5: *Summary of Measurement Properties by Test*

Physical Performance Test	Statistical Property	Values Reported by Included Study	Values Reported by Prior Study (Healthy Cohort)	Value Reported by Prior Study (Injured Cohort)
SEBT	Reliability	ICC 0.89-0.97 [35]	ICC 0.67-0.87 [54] 0.82-0.99* [55] 0.35-0.96* [56]	ICC 0.81-1.00 [57] 0.81-0.96 [58] 0.85-0.93 [59]
	Agreement	SEM 1.6cm [35]	-	-
	Construct Validity	Yes [30,35,38,40,41,60]	-	-
	Criterion Validity	-	-	-
	Responsiveness	Yes [24,25,27,29,32,37,39,41]	-	-
MODIFIED RHOMBERG	Reliability	ICC 0.998 [27]	-	ICC 034-0.69 [61]
	Agreement	-	-	-
	Construct Validity	-	-	-
	Criterion Validity	-	-	-
	Responsiveness	Yes [27,28]	-	-

SIDE HOP TEST	Reliability	ICC 0.84-0.99 [27,35,36]	-	ICC 0.96 [50]
	Agreement	SEM .06 sec [35]	-	SEM 0.37 Seconds [50] MDC 5.82 seconds [50]
	Construct Validity	Yes [35]	-	-
	Criterion Validity	-	-	-
	Responsiveness	Yes [27,30,32]	-	-
FOOT LIFT TEST	Reliability	0.989 [27]	-	-
	Agreement	-	-	-
	Construct Validity	-	-	-
	Criterion Validity	-	-	-
	Responsiveness	Yes [27]	-	-
SINGLE LEG HOP FOR DIS-TANCE	Reliability	-	-	-
	Agreement	-	-	-
	Construct Validity	-	-	-
	Criterion Validity	-	-	-
	Responsiveness	-	-	-
BESS TEST	Reliability	-	-	-
	Agreement	-	-	-
	Construct Validity	-	-	-
	Criterion Validity	-	-	-
	Responsiveness	Yes [32]	-	-
SINGLE LIMB HEEL RAISE TEST	Reliability	R2= .508 [42]	-	-
	Agreement	-	-	-
	Construct Validity	-	-	-
	Criterion Validity	Yes [42]	-	-
	Responsiveness	Yes [42]	-	-
6 METER CROSS-OVER HOP TEST	Reliability	ICC 0.84-0.96 [36]	-	ICC 0.96 [50]
	Agreement	-	-	SEM 0.37 seconds [50] MDC 1.03 seconds [50]
	Construct Validity	No [36]	-	-
	Criterion Validity	-	-	-
	Responsiveness	-	-	-

FIGURE OF EIGHT HOP TEST	Reliability	ICC 0.84-0.96 [36]	-	ICC 0.95 [50]
	Agreement	-	-	SEM 1.66 seconds [50] MDC 4.59 seconds [50]
	Construct Validity	No [30,36]	-	-
	Criterion Validity	-	-	-
	Responsiveness	No [30]	-	-
TRIPLE CROSSOVER HOP TEST	Reliability	0.93-0.96 [36]	-	-
	Agreement	-	-	-
	Construct Validity	No [36]	-	-
	Criterion Validity	-	-	-
	Responsiveness	-	-	-
LATERAL HOP TEST FOR DISTANCE	Reliability	0.93-0.96 [36]	-	-
	Agreement	-	-	-
	Construct Validity	No [36]	-	-
	Criterion Validity	-	-	-
	Responsiveness	-	-	-
SQUARE HOP TEST	Reliability	-	-	ICC 0.9 [50]
	Agreement	-	-	SEM 1.4 seconds [50] MDC 3.88 seconds [50]
	Construct Validity	-	-	-
	Criterion Validity	-	-	-
	Responsiveness	-	-	-

Note: *Cited by Hegedus, ICC, intraclass correlation coefficient; MDC, minimal detectable change; SEM, standard error of measure.

Reliability

Reliability was rated positive for 10 of the 12 PPTs. None of the included studies reported intraclass correlations (ICCs) for the BESS Test or the Single Leg Hop for Distance Test. All included studies that reported on reliability suggested good to excellent reliability for all 10 tests for which the ICCs were reported.

Agreement/Measurement

Of the 25 studies included in this review, only two reported on measurement error (SEM) or minimal detectable change (MDC). SEMs were reported for five PPTs, with the SEBT at 1.6cm, the Side Hop Test at 0.06 seconds to 0.37 seconds, the 6 Meter Cross-Over Hop Test at 0.37 seconds, the Figure-of-Eight Hop Test at 1.66 seconds and the Square Hop Test at 1.4 seconds [35,50]. MDCs were reported for four

hopping PPTs, with the Side Hop Test at 5.82 seconds, 6 Meter Hop Test at 1.03 seconds, the Figure-of-Eight Hop Test at 4.59 seconds and the Square Hop Test at 3.88 seconds [50]. However, as the MIC was not calculated for the above PPTs, a grade could not be determined.

Construct Validity

Construct validity is the ability of a PPT to be able to discriminate between healthy (i,e., athletes that are ready for RTS) and those that are unhealthy athletes (i.e., those not physically ready to RTS) based upon the PPT performance. Only six of the PPTs had any form of quality rating for construct validity. Of the six, only the SEBT and the Side Hop Test demonstrated positive quality ratings for construct validity. The SEBT had all six studies report positive quality ratings, while the Side Hop Test had one study address and report positive quality ratings for construct validity [30,35,38,40,41,60]. The 6 Meter Cross-Over Hop Test, Figure-of-Eight Hop Test, Triple Cross-Over Hop Test and Lateral Hop Test for Distance all received a negative rating on construct validity from the single study that assessed this [36]. None of the four tests with negative ratings were able to detect differences between limbs with histories of chronic lateral ankle sprains and healthy controls. No other PPTs had any study address construct validity ratings.

Criterion Validity

Only one of the 12 PPTs and one study of the 25 included in this review had any rating on criterion validity. The Single Limb Heel Raise Test was found to a positive rating on determining readiness for return to sport post Achilles tendon repair [42]. No other studies or PPTs had any reports or mentions of criterion validity.

Responsiveness

Six of the 12 PPTs had positive ratings for responsiveness. The SEBT, Modified Rhomberg, Side Hop Test, Foot Lift Test, BESS Test and Single Limb Heel Raise Test all demonstrated the ability to detect changes in function in relation to various interventions. The Figure-of-Eight Hop Test was found to not be responsive to the use of kinesio-tape on dynamic balance, muscle strength and functional performance in those with FAI [30].

Discussion

The current systematic review aimed to determine the clinically applicable PPT to assess RTS readiness in competitive athletes, ages 12 to 65, following musculoskeletal foot and ankle pathology. Across the 25 articles that were included, 12 PPTs were identified. One test assessed muscle performance, four tests measured postural control, and the remaining 7 PPT involved hopping in one or multiple directions for either distance or time. The results expand on existing knowledge by updating previous reviews, identifying additional PPTs, and recording each PPT's measurement properties. However, we found a lack of evidence on RTS standards among the currently available PPTs. Most included studies focused on CAI and LAS rather than specific RTS testing for musculoskeletal foot and ankle pathologies in general.

A prior systematic review by Hegedus *et al.* (2015b) [5] included 31 studies, identifying 14 lower extremity PPTs. The current systematic review included 25 studies and identifiedadditional PPTs (i.e., Foot Lift Test,

BESS Test, Modified Rhomberg Test, Side Hop Test, Single Limb Heel Raise Test, Square Hop Test, and the Figure-of-Eight Hop Test) while omitting six of the 14 PPTs identified by Hegedus *et al.*, (2015b) [5] (the 40-yard Sprint, Shuttle Run Test, Vertical Leap, T-Agility Test, and the Beep Test) (see Table 6). Reasons for the discrepancies between the current results and that of Hegedus *et al.* (2015b) [5] were due to one (or more) of the following reasons: (1) new available research that has been published since 2015; (2) Hegedus *et al.*, (2015b) [5] investigated PPTs for the hip and the knee, as well as the ankle, and (3) the current review only included studies that applied the PPTs to injured athletes. For example, the vertical jump and the multistage fitness test or "beep" test were excluded as the reported injuries were not specific to the foot and ankle [62].

Table 6: *Comparison of Identified Physical Performance Tests*

Body & Structure Measurement	**Specific Physical Performance Tests Identified**	
	Hegedus *et al.* (2015) [4,5]	**Current Review**
Static Postural Control		- Balance Error Scoring System Test (BESS) - Modified Romberg - Foot Lift Test
Dynamic Postural Control	- SEBT or Y-Balance Test	- SEBT or Y-Balance Test
Muscle Strength		- Heel Rise Test
Muscle Power	- Vertical Leap[a]	
Hop Test - Linear	- Single Hop for Distance - Triple Hop for Distance - 6-Meter Timed Hop - Lateral Hop for Distance - Medial Hop for Distance[c]	- Single Hop for Distance - Triple Hop for Distance[b] - 6-Meter Timed Hop[b] - Lateral Hop for Distance
Hop Test - Change of Direction	- 6-Meter Timed Crossover Hop - Triple Crossover Hop for Distance - Hexagon Hop Test	- 6-Meter Timed Crossover Hop - Triple Crossover Hop for Distance - Hexagon Hop Test - Side Hop - Figure 8 Hop Test - Square Hop Test[d]
Speed - Linear	- Sprint Test: 40 yards[e]	
Speed - Change of Direction	- T-Agility Test[f] - Shuttle Run[g]	

Cardiovascular Fitness	- Multistage Fitness or "beep" Test[h]	

Note: Red text indicates the physical performance test was not included in the other systematic review; green text indicates the physical performance test was identified in both systematic reviews. SEBT, Star Excursion Balance Test

aExcluded due to looking at knee injuries [62,63], excluded due to not specific to foot and ankle injuries [64], excluded due to healthy population [65-67]; bPPT described by Sekir *et al.* (2008; 2007) [68,69] and included from Hegedus *et al.*, (2015a)[4]; cExcluded due to looking at hip injuries [70]; dHand selected test from Caffrey *et al.*, (2009) [50] as it was able to discriminate between functional ankle instability (FAI) limb from uninvolved limb, for reasons unknown, the test was not included by Hegedus *et al.*, (2015a) [4]; eExcluded due to healthy population [71]; fExcluded due to healthy population [66,67,71]; gExcluded due to healthy population [65]; hExcluded due to looking at knee injuries [62,63]

More recent reviews related to RTS decision-making following foot and ankle injuries have been published [9,11,12]. Habets *et al.* (2018) [9] performed a systematic review to investigate RTS criteria for individuals with Achilles tendinopathy. Likewise, both Wikstrom *et al.* (2020) [12] and Tassignon *et al.* (2019) [11] reviewed prospective studies that used a criterion-based RTS decision-making process for patients with LAS. Each of these reviews was challenged to identify PPTs that helped to determine RTS readiness in their perspective injured populations [9,11,12]. Habets *et al.* (2018) [9] found that criteria for RTS as related to Achilles tendinopathy were determined by factoring the following criteria: level of pain, level of functional recovery, muscular strength, range of motion, endurance, medical advice, psychosocial factors, and anatomical/physiological properties of the Achilles tendon. Furthermore, there is evidence in sports literature suggesting that combining results of multiple functional performance tests has excellent clinical utility compared to a single stand-alone test [72]. These findings, combined with the current results, demonstrate the need to produce and study a criterion-based RTS test battery that can be used for individuals recovering from a foot or ankle injury.

Proposed Functional Performance Testing

Though several individual PPT and the respective measurement properties have been established in the athletes with ankle and foot musculoskeletal injuries, the clinical application of the results may remain ambiguous for several reasons:

1.The current results, nor previous reviews, have demonstrated high-quality evidence and consensus on RTS PPT criteria for the musculoskeletal injuries of the foot and ankle complex [5,11,12].

2.There are several qualitative and quantitative factors to consider in the RTS decision-making beyond the constructs that PPT can capture [11,13,73,74].

3.Further studies have sought to establish the predictive, or criterion, the validity of PPT by testing uninjured athletes, tracked the athletes over time, and correlate PPT performance to subsequent foot and ankle injury incidences [46,75-78]. Though studies that investigated the predictive ability of PPT performance of future foot and ankle injuries in healthy athletes were not included in the current review, the additional PPT that demonstrate injury prediction validity may still provide value in considering RTS decisions for injured athletes.

To summarize the current results, provide clinical application recommendations, and illustrate a framework for future research in the implementation, reliability, and validation of comprehensive RTS criteria, a functional testing battery is proposed (See Table 7). Previously published RTS functional testing algorithms for the upper and lower extremity, related systematic reviews, the synthesized results of the included studies, and relevant identified articles were used to compose the proposed functional testing battery using low cost (<$1000) equipment (See Table 7) [7,45,79-83].

Table 7: *Foot and Ankle Functional Testing Battery*

Body Structure & Function	Test	Criteria
Tissue Healing Timeline	1. Duration of time since injury	1. Dependent on pathoanatomical structure and severity of injury[a, b]
Pain & Inflammation	2. Visual Analog Scale (VAS) 3. Joint effusion	2. Pain/soreness/symptoms <2/10 at rest and with activity[a, b] 3. Minimal to none following activity[b]
Subjective Functional Outcome Measures[c, d, e, f, g]	4a. Foot and Ankle Disability Index (FADI) 4b. Foot and Ankle Ability Measure (FAAM) 4c. Victorian Institute of Sports Assessment - Achilles (VISA-A) 4d. Cumberland Ankle Instability Tool (CAIT) 4e. Lower Extremity Functional Scale (LEFS)	4. Criterion cutoff for RTS not established
Subjective Psychological Outcome Readiness[b]	5. Injury-psychological readiness to return sport scale	5. Score >50[o]
Range of Motion	6a. Dorsiflexion Heel Rocker Test[b] 6b. Goniometric[b] 6c. Weight Bearing Lunge Test[b, h]	6a. 10 successive heel rockers[b] 6b. Full range of motion[b] 6c. Patella 4" anterior to second ray[i]
Static Postural Control	One of the following: 7a. Balance Error Scoring System (BESS) 7b. Modified Rhomberg	7a. within one standard deviation of age & gender normative mean data[j] 7b. >90% LSI OR age & gender normative data[k]
Dynamic Postural Control	8. Star Excursion Balance Test / LE-YBT	8. (1) < 4 cm difference in any one reach direction between limbs; (2) >90% LSI composite score; (3) >90% composite score[l]

Strength	One of the following ankle dorsiflexion and plantarflexion strength tests: 9a. Heel Raise Test[b] 9b. Handheld Dynamometer Ankle[b] 9c. Manual Muscle Testing[b] AND One of the following hip abduction strength tests: 10a. Handheld Dynamometer Ankle[n,v] 10b. Manual Muscle Testing[n]	9a. >90% LSI number of repetitions 9b. >90% LSI 9c. 5/5 MMT 10a. >90% LSI 10b. 5/5 MMT
Hoping Tests - Linear	11a. SL Single Hop for Distance[p,r] 11b. SL Triple Hop for Distance[p,r] 11c. SL Lateral Hop for Distance 11d. SL Timed 6-Meter[p,r]	11a-d. >90% LSI[s,w] 11a. >80% Body Height[t,u] or 175±1.1% limb length[x] 11b. 2.2x Body Height[y] 11c. 2.0-2.2 x Body Height[aa]
Hoping Tests - Change of Direction	12a. SL Side Hop Test[q] 12b. SL Triple Cross Over Hop 12c. SL 6-Meter Cross Over Hop[p,r,q] 12d. SL Hexagon Hop Test[m] 12e. Figure 8 Hop Test[q] 12f. Square Hop Test[q]	12a-f. >90% LSI[s] 12a-c,e. Equal to predicted LSI predictive equation[w] 12d. 14 number of accurate hops per limb[m] 12f. Healthy Controls: 15.7 ± 0.4 sec/5 revolutions[q]
Full Lower Extremity Functional Test	13a. Lower Extremity Functional Tests (LEFT)[z] OR 13b. Functional Lower Extremity Evaluation (FLEE)	13a. LEFT: M: ≤100 sec | F: 117.2 sec[z] 13b. FLEE: *test provides pass/fail cutoff for all test batteries*
Sports Specific Tests	14. Test(s) will be specific to the athlete's sport	14. Sports specific reference values: *dependent on variables age, gender, level of sport*

Note. IKDC, international knee documentation committee subjective evaluation form; MMT, manualmuscle testing; HHD, hand-held dynamometer, CKC, closed kinetic chain; OKC, open kinetic chain; LSI, limb symmetry index = involved limb/non-involved limb; sec, seconds; SL, single limb; LE-YBT, Lower Extremity Y-Balance Test.

[a]See Axe & Snyder-Mackler, (2005) [84] for proposed tissue healing timelines for specific pathoanatomical structures; [b]Wikstrom *et al.*, (2020) [12]; [c]Delahunt *et al.*, (2018) [55]; [d]Shultz *et al.*, (2013) [86], [e]Martin & Irrgang, (2007) [87], [f]Tassignon *et al.*, (2019) [11], [g]Habets, (2018) [9], [h]Powden *et al.*, (2019) [44]; [i]Cook, (2010) [88]; [j]Ozinga *et al.*, (2018) [89]; [k]Springer *et al.*, (2007) [90]; [l]Functional Movement Systems, (2015) [24]; [m]Witchalls *et al.*, (2013) [75]; [n]Powers *et al.*, (2017) [78], [o]Monahan, (2018) [92]; [p]Sekir *et al.*, (2008) [68]. [q]Caffery *et al.*, (2009) [50]; [r]Sekir *et al.*, (2007) [69], [s]Greisberg *et al.*, (2019) [93], Gokeler *et al.*, (2017) [94], Logerstedt *et al.*, (2012) [95], Greenberg *et al.* (2020a) [96], Brumitt *et al.* (2013) [44]; [t]Davies & Zillmer (2000) [97]; [u]Witchalls *et al.*, (2013) [3]. [v]Yalfani *et al.*, (2017) [98]; [w]Madsen *et al.*, (2020) [99]; [x]Onate *et al.* (2018) [100]; [y]Greenberg *et al.* (2020b) [101]; [z]Brumitt *et al.*, (2013) [44], Brumitt *et al.*, (2018) [102], Haitz *et al.*, (2014) [103]; [aa]Hardesty *et al.*, (2017) [104]

Strengths and Limitations

A strength of the current review is that it contributes the following points to the existing body of knowledge of lower extremity PPTs to assess return to function following musculoskeletal foot and ankle injuries:

➢ A synthesis of PPT and clinical recommendations of how these might supplement a more comprehensive RTS test battery.

➢ There exists a lack of consensus on the appropriate standards and criteria for RTS following ankle and foot injuries in the athletic population.

➢ Several measurement properties of the identified PPTs have yet to be established. Nearly all the PPT lack predictive and criterion validity (i.e., the ability to predict a successful return to sport), agreement, and/or reliability measures on injured populations. The lack of these values is concerning. Future research is needed to establish the specified measurement properties to support these tests for rehab and return to sport decision-making.

The current results are derived from a majority (n = 15) of non-RCTs with varying levels of IV (prospective observational studies) and III evidence (cross-sectional studies). In combination with the consistent lack of blinding in the RCTs (see Figure 3) and lack of transparency of the non-RCTs in defining the assessment of outcomes, confounding factors in the outcomes, and the cohort selection (see Figure 4) the level and quality of existing evidence is a limitation of the current study. Clinicians need to consider these limitations when interpreting and implementing these results. Future RCTs that blind participants, personnel, and assessors, when feasible, are recommended to improve the quality and level of evidence that exists regarding PPT for the foot and ankle complex.

Limitations due to the methodology of the current systematic review include: (1) although a prospective protocol was written for the present review, it was not formally registered on the PROSPERO website, (2) the final search strategy and its translation across searched databases was not audited by a medical school or academic institution librarian, and (3) the inability to perform two independent data extractions, RoB, and quality assessments, as recommended in the Cochrane Handbook for Systematic Reviews of interventions secondary to the time allotted to conduct the review and the size of the research team [105]. Finally, to make the result of the current study most clinically applicable across a broad spectrum of clinical settings, studies including outcomes requiring one or more expensive (>$1000) "laboratory" equipment (i.e., isokinetic dynamometers, Biodex balance system, force plates, motion capture systems, and surface electromyography) were excluded from this review. When budgets allow, or such pieces of equipment are readily available, it is recommended that clinicians supplement the proposed RTS testing battery with isokinetic strength and power tests, ground reaction forces during hoping tasks, and/or center of pressure evaluations during dynamic and static lower extremity tasks [106].

Conclusion

Overall, there is a lack of consensus among RTS standards and criteria following foot and ankle injuries in the athletic population. Among the 12 PPT identified, several measurement properties have yet to be

established for these tests. However, the SEBT, Side Hop Test, and Square Hop Test were the most responsive and reliable in assessing foot and ankle pathologies. Furthermore, only the SEBT and Side Hop tests have established psychometric analysis supporting their validity. The 20 Times Single-Leg Heel Raise test was also a valid, reliable, and responsive test to assess RTS readiness, but it was specific to patients who are post-Achilles tendon repair.

Using the current review results and encompassing other tests identified before 2015, a comprehensive RTS test battery for individuals with foot and ankle injuries is proposed. Further research is needed to establish validity and reliability for the proposed test battery and each test within the test battery.

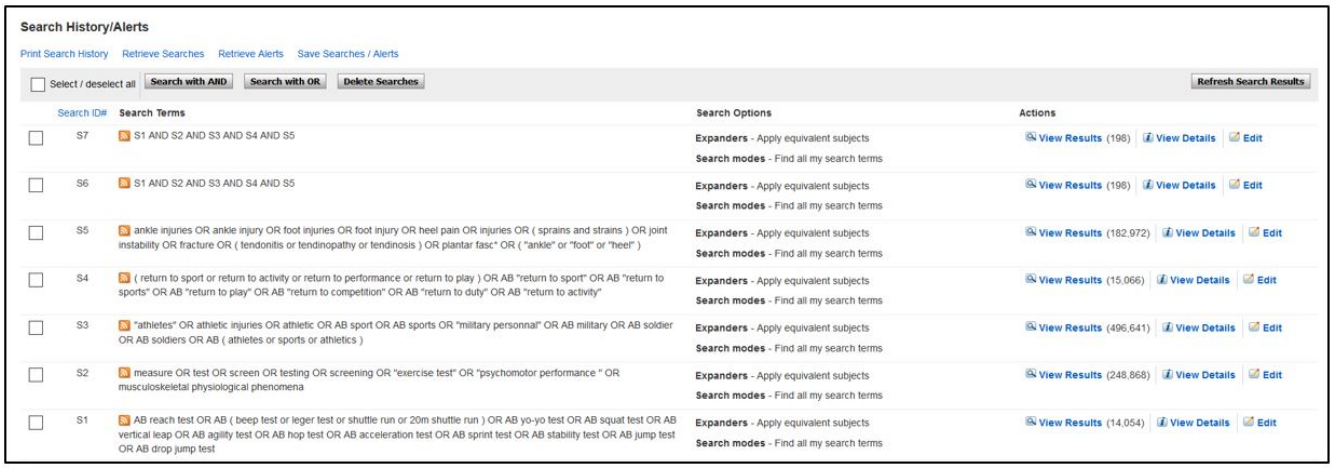
Supplemental Materials

Appendix A

CINAHL Search Strategy

S5	S1 AND S2 AND S3	**Limiters** - Age Groups: Adolescent: 13-18 years, Adult: 19-44 years, Middle Aged: 45-64 years, All Adult **Expanders** - Apply equivalent subjects **Search modes** - Find all my search terms	View Results (437) | View Details | Edit
S4	S1 AND S2 AND S3	**Expanders** - Apply equivalent subjects **Search modes** - Find all my search terms	View Results (906) | View Details | Edit
S3	((MM "Exercise Test") OR (MH "Physical Examination") OR "physical performance test") OR ((MH "Balance, Postural") OR (MH "Balance Training, Physical")) OR ((MH "Balance, Postural") OR (MH "Muscle Strength") OR "muscle strength" OR (MH "Muscle Strengthening")) OR "muscle power" OR "muscle endurance" OR ((MM "Physical Performance") OR (MM "Psychomotor Performance")) OR ((MM "Physical Performance") OR "physical performance" OR (MM "Psychomotor Performance")) OR ((MM "Functional Assessmen ...	**Expanders** - Apply equivalent subjects **Search modes** - Find all my search terms	View Results (103,846) | View Details | Edit
S2	((MM "Sports Re-Entry") OR (MH "Job Re-Entry") OR "return to sport") OR ((MM "Athletes") OR "athlete" OR (MH "Athletes, Professional") OR (MH "Athletes, College") OR (MH "Athletes, Master") OR (MH "Athletes, Disabled") OR (MH "Athletes, Amateur") OR (MH "Athletes, Male") OR (MH "Athletes, High School") OR (MH "Athletes, Female") OR (MH "Athletes, Elite"))	**Expanders** - Apply equivalent subjects **Search modes** - Find all my search terms	View Results (38,959) | View Details | Edit
S1	((MM "Ankle Injuries") OR (MM "Ankle") OR (MM "Ankle Sprain") OR (MM "Ankle Surgery") OR (MM "Lateral Ligament, Ankle") OR (MM "Ankle Fractures") OR (MM "Ankle Dislocation") OR (MM "Ankle Sprain, Syndesmosis") OR (MM "Ankle Joint") OR (MM "Talus") OR (MM "Arthroplasty, Replacement, Ankle")) OR "ankle injuries" OR ((MM "Foot Injuries") OR "foot injury") OR ((MH "Heel Pain") OR "heel pain") OR ((MM "Athletic Injuries") OR "athletic injuries") OR "sports injuries"	**Expanders** - Apply equivalent subjects **Search modes** - Find all my search terms	View Results (34,681) | View Details | Edit

SPORT Discus Search Strategy



Appendix B

Physical Performance Tests and Descriptions

TEST #1 SEBT - Star Excursion Balance Test or Y-Balance Test - Single Leg Balance with maximum reach of contralateral limb in anterior, posterior-lateral, and posterior-medial.					
Author (Year)	**Test Name**	**Alternate Test Description**	**Final Scoring Mechanism**	**Measurement Property**	**Quality of Measurement Property**
Anguish & Sandrey, (2018) [23]	SEBT	N/A	Distance	ICC used from previous studies [54,57,58]	0.35-0.96 0.67-0.87 0.81-0.96
Bagherian *et al.*, (2018) [24]	SEBT	N/A	Distance	ICC used from previous studies (Plisky *et al.*, 2009)	0.99-1.00
Cain *et al.*, (2017) [27]	SEBT	N/A	Distance	ICC used from previous studies [56]	- 0.81-0.93
Bagherian *et al.* (2019) [25]	SEBT	Normalized mean in relationship to leg length	Distance	ICC used from previous studies (Plisky *et al.*, 2009)	0.99-1.00
Cruz-Diaz *et al.*, (2014) [29]	SEBT	N/A	Distance	Not Listed	Not Listed
Fereydounnia *et al.*, (2019) [30]	SEBT	N/A	Distance	Not Listed	Not Listed

Golditz *et al.*, (2015) [31]	SEBT	N/A	Distance	Not Listed	Not Listed
Hall *et al.*, (2018) [32]	SEBT	N/A	Distance	Not Listed	Not Listed
Ko *et al.*, (2018) [35]	SEBT	N/A	Distance	ICC SEM	0.89-0.97 1.6cm
McCann *et al.*, (2017)	SEBT-ANT	Reach Distance of Anterior SEBT Test only.	Distance	Not listed	Not Listed
Powden *et al.*, (2019) [37]	Y Balance Test	N/A	Distance	ICC used from previous study (Shafer *et al.*, 2013)	Anterior 0.93 Posterior-Medial 0.91 Posterior Lateral 0.85
Ryu *et al.*, (2019) [38]	Y-Balance Test	N/A	Distance	Not Listed	Not Listed
Sierra-Guzman *et al.*, (2018) [39]	SEBT	N/A	Distance	Not Listed	Not Listed
Sierra-Guzman *et al.*, (2018) [40]	SEBT	N/A	Distance	Not Listed	Not Listed
Someeh *et al.*, (2015) [41]	SEBT	N/A	Distance	Not Listed	Not Listed

Test #2 Modified Rhomberg Test – Standing on one leg maintaining balance for an extended period.

Author (Year)	**Test Name**	**Alternate Test Description**	**Final Scoring Mechanism**	**Measurement Property**	**Quality of Measurement Property**
(Alves *et al.*, 2018) [22]	Single Leg Balance Eyes Closed	Center of Pressure, Speed, Anterior Posterior, Mediolateral Displacement Center of Pressure Area (VICON Motion Analysis & Force Platform)	Time	ICC used from previous study (Sharma *et al.*, 2011) [106]	0.80-0.89
Cain *et al.*, (2017) [27]	Single Leg Balance Eyes Closed	Time in Balance	Time	ICC	.998

Cho *et al.*, (2019) [28]	Modified Rhomberg	Time in Balance	Time	Not Listed	Not Listed
Powden *et al.*, (2019) [37]	Single Leg Balance Eyes open/ Closed	AP, MP position control in time to boundary with MATLAB	Motion outside of boundary	ICC used from previous study (Hoch & McKeon, 2015)	0.34-0.69

TEST #3 Side Hop Test - Hoping in a lateral direction back and forth, typically over a 30cm distance, for 10 repetitions.

Author (Year)	Test Name	Alternate Test Description	Final Scoring Mechanism	Measurement Property	Quality of Measurement Property
Alves *et al.*, (2018) [22]	Side Hop Test	Lateral Hop Test	Time	Not Listed	Not Listed
Cain *et al.*, (2017) [27]	Side Hop Test	N/A	Time	ICC	.999
Fereydounnia *et al.*, (2019) [30]	Side Hop Test	N/A	Time	Not Listed	Not Listed
Hall *et al.*, (2018) [32]	Side Hop Test	N/A	Time	Not Listed	Not Listed
Ko *et al.*, (2018) [35]	Single Leg Hop Test	N/A	Time	ICC SEM	.94 .06s
Madsen *et al.*, (2018) [36]	Side Hop Test	N/A	Time	ICC	.84-.96
Caffrey *et al.*, (2009) [50]	Side Hop Test	N/A	Time	ICC SEM MDC	0.84 2.1 seconds 5.82 seconds

TEST #4 Foot Lift Test – Assess how well individuals can maintain balance during 30 second trial

Author (Year)	Test Name	Alternate Test Description	Final Scoring Mechanism	Measurement Property	Quality of Measurement Property
Cain *et al.*, (2017) [27]	Foot Lift Test	Number of loss of contacts of parts of foot with ground	Errors or Balance Disturbances. Average of three trials	ICC	.989

TEST #5 Single Limb Hop for Distance (Limb Symmetry)- Hoping from a stationary standing single limb position as far forward as possible					
Author (Year)	**Test Name**	**Alternate Test Description**	**Final Scoring Mechanism**	**Measurement Property**	**Quality of Measurement Property**
Coetzee, *et al.*, (2018) [53]	Single Limb Hop for Distance	Starting on non-operative ankle first, maximum single leg hop distance. One practice trial and average of two successful trials	Limb Symmetry Index	Not Listed	Not Listed

TEST #6 Balance Error Scoring System (BESS) -measure of static balance. Double Limb Stance, Single limb Stance and Tandem Stance on firm and unstable surfaces					
Author (Year)	**Test Name**	**Alternate Test Description**	**Final Scoring Mechanism**	**Measurement Property**	**Quality of Measurement Property**
Hall *et al.*, (2018) [32]	Balance Error Scoring System	BESS Test with errors defined as loss of balance with or without correction. Max score per stance: 10 of each of the 6 trials. Given 1 practice trial and 1 test trial. Eyes closed with hands on hips	Total errors during 20 second trial in each of the 6 test trials	Not Listed	Not Listed

TEST #7 Single Limb Heel Raise Test - Standing on one limb, raise heel off ground while maintaining contact of forefoot					
Author (Year)	**Test Name**	**Alternate Test Description**	**Final Scoring Mechanism**	**Measurement Property**	**Quality of Measurement Property**
Toyooka *et al.*, (2017) [42]	Single Limb Heel Raise Test (SLHRT)	Half body weight single leg heel raise, 1 full body weight single leg heel raise, 20 full body weight single leg heel raises	Time in weeks to be able to perform 1 full body weight heel raise, 20 full body weight heel raises	Correlation	FBW 1: R^2 =.317 FBW 20: R^2 =.508

TEST #8 6 Meter Cross-Over Hop Test - hop diagonally over a 15cm wide line for 6 meters as quickly as possible					
Author (Year)	**Test Name**	**Alternate Test Description**	**Final Scoring Mechanism**	**Measurement Property**	**Quality of Measurement Property**
Madsen *et al.*, (2018) [36]	6 Meter Cross Over Hop Test	N/A	Time	ICC	0.84-0.96

TEST #9 Figure-of-Eight Hop Test – hop in a Figure-of-Eight fashion around 2 cones 5 m apart, 2 times.					
Author (Year)	**Test Name**	**Alternate Test Description**	**Final Scoring Mechanism**	**Measurement Property**	**Quality of Measurement Property**
Alves *et al.*, (2018) [22]	Figure-of-8 Hop Test	N/A	Time	Not Listed	Not Listed
Madsen *et al.*, (2018) [36]	Figure-of-Eight Hop Test	N/A	Time	ICC	0.84-0.96
Fereydounnia *et al.*, (2019) [30]	Figure-of-Eight Hop Test	Total time to completion with average of three trials used for final evaluation	Time	Not Listed	Not Listed
Caffrey *et al.*, (2009) [50]	Figure-of-Eight Hop Test	N/A	Time	ICC SEM MDC	0.95 1.66 seconds 4.59 seconds

TEST #10 Triple Crossover Hop Test - Subjects jump diagonally over 15cm wide line as far forward as possible.					
Author (Year)	**Test Name**	**Alternate Test Description**	**Final Scoring Mechanism**	**Measurement Property**	**Quality of Measurement Property**
Madsen *et al.*, (2018) [36]	Triple Cross Over Hop	N/A	Total Distance of Three Hops	ICC	0.93-0.96

TEST #11 Lateral Hop for Distance - Subjects jump laterally three times as far as possible.					
Author (Year)	**Test Name**	**Alternate Test Description**	**Final Scoring Mechanism**	**Measurement Property**	**Quality of Measurement Property**
Madsen *et al.*, (2018) [36]	Lateral Distance Hop	N/A	Total Distance of Three Hops	ICC	0.93-0.96

TEST #12 Square Hop Test- Subjects jump around 5 m course outlined by two cones in figure 8 position.					
Author (Year)	**Test Name**	**Alternate Test Description**	**Final Scoring Mechanism**	**Measurement Property**	**Quality of Measurement Property**
Caffrey *et al.*, 2009) [50]	Square Hop Test	N/A	Time	ICC SEM MDC	0.9 1.4 seconds 1.88 seconds

Appendix C

Tool to Assess Risk of Bias (RoB) Assessment of Randomized Control Trials

Alves *et al.*, (2018) [22]

Question	Example	Definitely Yes (Low RoB)	Probably Yes	Probably No	Definitely No (High RoB)
1. Was the allocation sequence adequately generated?*	**Examples of low risk of bias:** • Referring to a random number table • Using a computer random number generator • Coin tossing • Shuffling cards or envelopes • Throwing dice • Drawing of lots • Minimization with or without a random element **Examples of high risk of bias:** • Sequence generated by odd or even date of birth • Sequence generated by some rule based on date (or day) of admission • Sequence generated by some rule based on hospital or clinic record number • Allocation by judgement of the clinician • Allocation by preference of the participant • Allocation based on the results of a series laboratory test or series of tests • Allocation by availability of the intervention	X			
2. Was the allocation adequately concealed?	**Examples of low risk of bias allocation concealment techniques:** • Central allocation (including telephone, web-based, and pharmacy-controlled, randomization) **Examples of possible low risk of bias:** • Sequentially numbered drug containers of identical appearance • Sequentially numbered, opaque, sealed envelopes **Examples of high risk of bias allocation concealment techniques:** • Using an open random allocation schedule (e.g. a list of random numbers) • Assignment envelopes were used without appropriate safeguards (e.g. if envelopes were unsealed or non-opaque or not sequentially numbered) • Alternation or rotation • Date of birth • Case record number • Any other explicitly unconcealed procedure				X

3. Blinding: Was knowledge of the allocated interventions adequately prevented?*					
3.a. Were patients blinded?	**Examples of low risk of bias:** • No blinding but the review authors judge that the outcome and the outcome measurement are not likely influenced by lack of blinding • Blinding of participants and key study personnel ensured, and unlikely that blinding could have been broken • Either participants or some key study personnel were not blinded, but outcome assessment was blinded and the nonblinding of others unlikely to introduce bias **Examples of high risk of bias:** • No blinding or incomplete blinding, and the outcome or outcome measurement is likely to be influenced by lack of blinding • Blinding of key study participants and personnel attempted, but likely that the blinding could have been broken • Either participants or some key study personnel were not blinded, and the non-blinding of others likely to introduce bias	X			
3.b. Were healthcare providers blinded?					X
3.c. Were data collectors blinded?		X			
3.d. Were outcome assessors blinded?		X			
3.e. Were data analysts blinded?			X		
4. Was loss to follow-up (missing outcome data) infrequent?	**Examples of low risk of bias:** • No missing outcome data • Reasons for missing outcome data unlikely to be related to outcome (for survival data, censoring unlikely to be introducing bias) • Missing outcome data balanced in numbers across intervention groups, with similar reasons for missing data across groups • For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk not enough to have an important impact on the intervention effect estimate • For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes not enough to have an important impact on observed effect size • Missing data have been imputed using appropriate methods **Examples of high risk of bias** • Reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across intervention groups • For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk enough to induce important bias in intervention effect estimate • For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes enough to induce clinically relevant bias in observed effect size • 'As-treated' analysis done with substantial departure of the intervention received from that assigned at randomization • Potentially inappropriate application of simple imputation	X			

5. Are reports of the study free of selective outcome reporting?*	**Examples of low risk of bias:** • The study protocol is available and all of the study's pre-specified (primary and secondary) outcomes that are of interest in the review have been reported in the pre-specified way • The study protocol is not available but it is clear that the published reports include all expected outcomes, including those that were pre-specified (convincing text of this nature may be uncommon) **Examples of high risk of bias** • Not all of the study's pre-specified primary outcomes have been reported • One or more primary outcomes is reported using measurements, analysis methods or subsets of the data (e.g. subscales) that were not pre-specified • One or more reported primary outcomes were not pre-specified (unless clear justification for their reporting is provided, such as an unexpected adverse effect) • One or more outcomes of interest in the review are reported incompletely so that they cannot be entered in a meta-analysis • The study report fails to include results for a key outcome that would be expected to have been reported for such a study	X			
6. Was the study apparently free of other problems that could put it at a risk of bias?*	**Examples of low risk of bias:** • The study appears to be free of other sources of bias **Examples of high risk of bias** • Had a potential source of bias related to the specific study design used • Stopped early due to some data-dependent process (including a formal-stopping rule) • Had extreme baseline imbalance • Has been claimed to have been fraudulent • Had some other problem	X			

Note. *May omit this item

Anguish & Sandrey, (2018) [23]

Question	Example	Definitely Yes (Low RoB)	Probably Yes	Probably No	Definitely No (High RoB)
1. Was the allocation sequence adequately generated?*	**Examples of low risk of bias:** • Referring to a random number table • Using a computer random number generator • Coin tossing • Shuffling cards or envelopes • Throwing dice • Drawing of lots • Minimization with or without a random element **Examples of high risk of bias:** • Sequence generated by odd or even date of birth • Sequence generated by some rule based on date (or day) of admission • Sequence generated by some rule based on hospital or clinic record number • Allocation by judgement of the clinician • Allocation by preference of the participant • Allocation based on the results of a series laboratory test or series of tests • Allocation by availability of the intervention	X			
2. Was the allocation adequately concealed?	**Examples of low risk of bias allocation concealment techniques:** • Central allocation (including telephone, web-based, and pharmacy-controlled, randomization) **Examples of possible low risk of bias:** • Sequentially numbered drug containers of identical appearance • Sequentially numbered, opaque, sealed envelopes **Examples of high risk of bias allocation concealment techniques:** • Using an open random allocation schedule (e.g. a list of random numbers) • Assignment envelopes were used without appropriate safeguards (e.g. if envelopes were unsealed or non-opaque or not sequentially numbered) • Alternation or rotation • Date of birth • Case record number • Any other explicitly unconcealed procedure	X			

3. Blinding: Was knowledge of the allocated interventions adequately prevented?*					
3.a. Were patients blinded?	**Examples of low risk of bias:** • No blinding but the review authors judge that the outcome and the outcome measurement are not likely influenced by lack of blinding • Blinding of participants and key study personnel ensured, and unlikely that blinding could have been broken • Either participants or some key study personnel were not blinded, but outcome assessment was blinded and the nonblinding of others unlikely to introduce bias **Examples of high risk of bias:** • No blinding or incomplete blinding, and the outcome or outcome measurement is likely to be influenced by lack of blinding • Blinding of key study participants and personnel attempted, but likely that the blinding could have been broken • Either participants or some key study personnel were not blinded, and the non-blinding of others likely to introduce bias		X		
3.b. Were healthcare providers blinded?		X			
3.c. Were data collectors blinded?					X
3.d. Were outcome assessors blinded?					X
3.e. Were data analysts blinded?				X	
4. Was loss to follow-up (missing outcome data) infrequent?	**Examples of low risk of bias:** • No missing outcome data • Reasons for missing outcome data unlikely to be related to outcome (for survival data, censoring unlikely to be introducing bias) • Missing outcome data balanced in numbers across intervention groups, with similar reasons for missing data across groups • For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk not enough to have an important impact on the intervention effect estimate • For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes not enough to have an important impact on observed effect size • Missing data have been imputed using appropriate methods **Examples of high risk of bias** • Reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across intervention groups • For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk enough to induce important bias in intervention effect estimate • For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes enough to induce clinically relevant bias in observed effect size • 'As-treated' analysis done with substantial departure of the intervention received from that assigned at randomization • Potentially inappropriate application of simple imputation	X			

5. Are reports of the study free of selective outcome reporting?*	**Examples of low risk of bias:** • The study protocol is available and all of the study's pre-specified (primary and secondary) outcomes that are of interest in the review have been reported in the pre-specified way • The study protocol is not available but it is clear that the published reports include all expected outcomes, including those that were pre-specified (convincing text of this nature may be uncommon) **Examples of high risk of bias** • Not all of the study's pre-specified primary outcomes have been reported • One or more primary outcomes is reported using measurements, analysis methods or subsets of the data (e.g. subscales) that were not pre-specified • One or more reported primary outcomes were not pre-specified (unless clear justification for their reporting is provided, such as an unexpected adverse effect) • One or more outcomes of interest in the review are reported incompletely so that they cannot be entered in a meta-analysis • The study report fails to include results for a key outcome that would be expected to have been reported for such a study	X			
6. Was the study apparently free of other problems that could put it at a risk of bias?*	**Examples of low risk of bias:** • The study appears to be free of other sources of bias **Examples of high risk of bias** • Had a potential source of bias related to the specific study design used • Stopped early due to some data-dependent process (including a formal-stopping rule) • Had extreme baseline imbalance • Has been claimed to have been fraudulent • Had some other problem	X			

Bagherian *et al.*, (2019) [25]

Question	Example	Definitely Yes (Low RoB)	Probably Yes	Probably No	Definitely No (High RoB)
1. Was the allocation sequence adequately generated?*	**Examples of low risk of bias:** • Referring to a random number table • Using a computer random number generator • Coin tossing • Shuffling cards or envelopes • Throwing dice • Drawing of lots • Minimization with or without a random element **Examples of high risk of bias:** • Sequence generated by odd or even date of birth • Sequence generated by some rule based on date (or day) of admission • Sequence generated by some rule based on hospital or clinic record number • Allocation by judgement of the clinician • Allocation by preference of the participant • Allocation based on the results of a series laboratory test or series of tests • Allocation by availability of the intervention	X			
2. Was the allocation adequately concealed?	**Examples of low risk of bias allocation concealment techniques:** • Central allocation (including telephone, web-based, and pharmacy-controlled, randomization) **Examples of possible low risk of bias:** • Sequentially numbered drug containers of identical appearance • Sequentially numbered, opaque, sealed envelopes **Examples of high risk of bias allocation concealment techniques:** • Using an open random allocation schedule (e.g. a list of random numbers) • Assignment envelopes were used without appropriate safeguards (e.g. if envelopes were unsealed or non-opaque or not sequentially numbered) • Alternation or rotation • Date of birth • Case record number • Any other explicitly unconcealed procedure	X			

3. Blinding: Was knowledge of the allocated interventions adequately prevented?*					
3.a. Were patients blinded?	**Examples of low risk of bias:** • No blinding but the review authors judge that the outcome and the outcome measurement are not likely influenced by lack of blinding • Blinding of participants and key study personnel ensured, and unlikely that blinding could have been broken • Either participants or some key study personnel were not blinded, but outcome assessment was blinded and the nonblinding of others unlikely to introduce bias **Examples of high risk of bias:** • No blinding or incomplete blinding, and the outcome or outcome measurement is likely to be influenced by lack of blinding • Blinding of key study participants and personnel attempted, but likely that the blinding could have been broken • Either participants or some key study personnel were not blinded, and the non-blinding of others likely to introduce bias				X
3.b. Were healthcare providers blinded?					X
3.c. Were data collectors blinded?				X	
3.d. Were outcome assessors blinded?					X
3.e. Were data analysts blinded?				X	
4. Was loss to follow-up (missing outcome data) infrequent?	**Examples of low risk of bias:** • No missing outcome data • Reasons for missing outcome data unlikely to be related to outcome (for survival data, censoring unlikely to be introducing bias) • Missing outcome data balanced in numbers across intervention groups, with similar reasons for missing data across groups • For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk not enough to have an important impact on the intervention effect estimate • For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes not enough to have an important impact on observed effect size • Missing data have been imputed using appropriate methods **Examples of high risk of bias** • Reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across intervention groups • For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk enough to induce important bias in intervention effect estimate • For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes enough to induce clinically relevant bias in observed effect size • 'As-treated' analysis done with substantial departure of the intervention received from that assigned at randomization • Potentially inappropriate application of simple imputation	X			

5. Are reports of the study free of selective outcome reporting?*	**Examples of low risk of bias:** • The study protocol is available and all of the study's pre-specified (primary and secondary) outcomes that are of interest in the review have been reported in the pre-specified way • The study protocol is not available but it is clear that the published reports include all expected outcomes, including those that were pre-specified (convincing text of this nature may be uncommon) **Examples of high risk of bias** • Not all of the study's pre-specified primary outcomes have been reported • One or more primary outcomes is reported using measurements, analysis methods or subsets of the data (e.g. subscales) that were not pre-specified • One or more reported primary outcomes were not pre-specified (unless clear justification for their reporting is provided, such as an unexpected adverse effect) • One or more outcomes of interest in the review are reported incompletely so that they cannot be entered in a meta-analysis • The study report fails to include results for a key outcome that would be expected to have been reported for such a study	X			
6. Was the study apparently free of other problems that could put it at a risk of bias?*	**Examples of low risk of bias:** • The study appears to be free of other sources of bias **Examples of high risk of bias** • Had a potential source of bias related to the specific study design used • Stopped early due to some data-dependent process (including a formal-stopping rule) • Had extreme baseline imbalance • Has been claimed to have been fraudulent • Had some other problem	X			

Best *et al.*, (2015) [26]

Question	Example	Definitely Yes (Low RoB)	Probably Yes	Probably No	Definitely No (High RoB)
1. Was the allocation sequence adequately generated?*	**Examples of low risk of bias:** • Referring to a random number table • Using a computer random number generator • Coin tossing • Shuffling cards or envelopes • Throwing dice • Drawing of lots • Minimization with or without a random element **Examples of high risk of bias:** • Sequence generated by odd or even date of birth • Sequence generated by some rule based on date (or day) of admission • Sequence generated by some rule based on hospital or clinic record number • Allocation by judgement of the clinician • Allocation by preference of the participant • Allocation based on the results of a series laboratory test or series of tests • Allocation by availability of the intervention		X		
2. Was the allocation adequately concealed?	**Examples of low risk of bias allocation concealment techniques:** • Central allocation (including telephone, web-based, and pharmacy-controlled, randomization) **Examples of possible low risk of bias:** • Sequentially numbered drug containers of identical appearance • Sequentially numbered, opaque, sealed envelopes **Examples of high risk of bias allocation concealment techniques:** • Using an open random allocation schedule (e.g. a list of random numbers) • Assignment envelopes were used without appropriate safeguards (e.g. if envelopes were unsealed or non-opaque or not sequentially numbered) • Alternation or rotation • Date of birth • Case record number • Any other explicitly unconcealed procedure				X

3. Blinding: Was knowledge of the allocated interventions adequately prevented?*					
3.a. Were patients blinded?	**Examples of low risk of bias:** • No blinding but the review authors judge that the outcome and the outcome measurement are not likely influenced by lack of blinding • Blinding of participants and key study personnel ensured, and unlikely that blinding could have been broken • Either participants or some key study personnel were not blinded, but outcome assessment was blinded and the nonblinding of others unlikely to introduce bias **Examples of high risk of bias:** • No blinding or incomplete blinding, and the outcome or outcome measurement is likely to be influenced by lack of blinding • Blinding of key study participants and personnel attempted, but likely that the blinding could have been broken • Either participants or some key study personnel were not blinded, and the non-blinding of others likely to introduce bias				X
3.b. Were healthcare providers blinded?					X
3.c. Were data collectors blinded?		X			
3.d. Were outcome assessors blinded?		X			
3.e. Were data analysts blinded?			X		
4. Was loss to follow-up (missing outcome data) infrequent?	**Examples of low risk of bias:** • No missing outcome data • Reasons for missing outcome data unlikely to be related to outcome (for survival data, censoring unlikely to be introducing bias) • Missing outcome data balanced in numbers across intervention groups, with similar reasons for missing data across groups • For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk not enough to have an important impact on the intervention effect estimate • For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes not enough to have an important impact on observed effect size • Missing data have been imputed using appropriate methods **Examples of high risk of bias** • Reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across intervention groups • For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk enough to induce important bias in intervention effect estimate • For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes enough to induce clinically relevant bias in observed effect size • 'As-treated' analysis done with substantial departure of the intervention received from that assigned at randomization • Potentially inappropriate application of simple imputation				X

5. Are reports of the study free of selective outcome reporting?*	**Examples of low risk of bias:** • The study protocol is available and all of the study's pre-specified (primary and secondary) outcomes that are of interest in the review have been reported in the pre-specified way • The study protocol is not available but it is clear that the published reports include all expected outcomes, including those that were pre-specified (convincing text of this nature may be uncommon) **Examples of high risk of bias** • Not all of the study's pre-specified primary outcomes have been reported • One or more primary outcomes is reported using measurements, analysis methods or subsets of the data (e.g. subscales) that were not pre-specified • One or more reported primary outcomes were not pre-specified (unless clear justification for their reporting is provided, such as an unexpected adverse effect) • One or more outcomes of interest in the review are reported incompletely so that they cannot be entered in a meta-analysis • The study report fails to include results for a key outcome that would be expected to have been reported for such a study	X			
6. Was the study apparently free of other problems that could put it at a risk of bias?*	**Examples of low risk of bias:** • The study appears to be free of other sources of bias **Examples of high risk of bias** • Had a potential source of bias related to the specific study design used • Stopped early due to some data-dependent process (including a formal-stopping rule) • Had extreme baseline imbalance • Has been claimed to have been fraudulent • Had some other problem		X		

Cain *et al.*, (2017) [27]

Question	Example	Definitely Yes (Low RoB)	Probably Yes	Probably No	Definitely No (High RoB)
1. Was the allocation sequence adequately generated?*	**Examples of low risk of bias:** • Referring to a random number table • Using a computer random number generator • Coin tossing • Shuffling cards or envelopes • Throwing dice • Drawing of lots • Minimization with or without a random element **Examples of high risk of bias:** • Sequence generated by odd or even date of birth • Sequence generated by some rule based on date (or day) of admission • Sequence generated by some rule based on hospital or clinic record number • Allocation by judgement of the clinician • Allocation by preference of the participant • Allocation based on the results of a series laboratory test or series of tests • Allocation by availability of the intervention		X		
2. Was the allocation adequately concealed?	**Examples of low risk of bias allocation concealment techniques:** • Central allocation (including telephone, web-based, and pharmacy-controlled, randomization) **Examples of possible low risk of bias:** • Sequentially numbered drug containers of identical appearance • Sequentially numbered, opaque, sealed envelopes **Examples of high risk of bias allocation concealment techniques:** • Using an open random allocation schedule (e.g. a list of random numbers) • Assignment envelopes were used without appropriate safeguards (e.g. if envelopes were unsealed or non-opaque or not sequentially numbered) • Alternation or rotation • Date of birth • Case record number • Any other explicitly unconcealed procedure				X

3. Blinding: Was knowledge of the allocated interventions adequately prevented?*					
3.a. Were patients blinded?	**Examples of low risk of bias:** • No blinding but the review authors judge that the outcome and the outcome measurement are not likely influenced by lack of blinding • Blinding of participants and key study personnel ensured, and unlikely that blinding could have been broken • Either participants or some key study personnel were not blinded, but outcome assessment was blinded and the nonblinding of others unlikely to introduce bias **Examples of high risk of bias:** • No blinding or incomplete blinding, and the outcome or outcome measurement is likely to be influenced by lack of blinding • Blinding of key study participants and personnel attempted, but likely that the blinding could have been broken • Either participants or some key study personnel were not blinded, and the non-blinding of others likely to introduce bias				X
3.b. Were healthcare providers blinded?					X
3.c. Were data collectors blinded?					X
3.d. Were outcome assessors blinded?					X
3.e. Were data analysts blinded?				X	
4. Was loss to follow-up (missing outcome data) infrequent?	**Examples of low risk of bias:** • No missing outcome data • Reasons for missing outcome data unlikely to be related to outcome (for survival data, censoring unlikely to be introducing bias) • Missing outcome data balanced in numbers across intervention groups, with similar reasons for missing data across groups • For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk not enough to have an important impact on the intervention effect estimate • For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes not enough to have an important impact on observed effect size • Missing data have been imputed using appropriate methods **Examples of high risk of bias** • Reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across intervention groups • For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk enough to induce important bias in intervention effect estimate • For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes enough to induce clinically relevant bias in observed effect size • 'As-treated' analysis done with substantial departure of the intervention received from that assigned at randomization • Potentially inappropriate application of simple imputation				X

5. Are reports of the study free of selective outcome reporting?*	**Examples of low risk of bias:** • The study protocol is available and all of the study's pre-specified (primary and secondary) outcomes that are of interest in the review have been reported in the pre-specified way • The study protocol is not available but it is clear that the published reports include all expected outcomes, including those that were pre-specified (convincing text of this nature may be uncommon) **Examples of high risk of bias** • Not all of the study's pre-specified primary outcomes have been reported • One or more primary outcomes is reported using measurements, analysis methods or subsets of the data (e.g. subscales) that were not pre-specified • One or more reported primary outcomes were not pre-specified (unless clear justification for their reporting is provided, such as an unexpected adverse effect) • One or more outcomes of interest in the review are reported incompletely so that they cannot be entered in a meta-analysis • The study report fails to include results for a key outcome that would be expected to have been reported for such a study	X			
6. Was the study apparently free of other problems that could put it at a risk of bias?*	**Examples of low risk of bias:** • The study appears to be free of other sources of bias **Examples of high risk of bias** • Had a potential source of bias related to the specific study design used • Stopped early due to some data-dependent process (including a formal-stopping rule) • Had extreme baseline imbalance • Has been claimed to have been fraudulent • Had some other problem		X		

Cruz-Diaz *et al.*, (2014) [29]

Question	Example	Definitely Yes (Low RoB)	Probably Yes	Probably No	Definitely No (High RoB)
1. Was the allocation sequence adequately generated?*	**Examples of low risk of bias:** • Referring to a random number table • Using a computer random number generator • Coin tossing • Shuffling cards or envelopes • Throwing dice • Drawing of lots • Minimization with or without a random element **Examples of high risk of bias:** • Sequence generated by odd or even date of birth • Sequence generated by some rule based on date (or day) of admission • Sequence generated by some rule based on hospital or clinic record number • Allocation by judgement of the clinician • Allocation by preference of the participant • Allocation based on the results of a series laboratory test or series of tests • Allocation by availability of the intervention	X			
2. Was the allocation adequately concealed?	**Examples of low risk of bias allocation concealment techniques:** • Central allocation (including telephone, web-based, and pharmacy-controlled, randomization) **Examples of possible low risk of bias:** • Sequentially numbered drug containers of identical appearance • Sequentially numbered, opaque, sealed envelopes **Examples of high risk of bias allocation concealment techniques:** • Using an open random allocation schedule (e.g. a list of random numbers) • Assignment envelopes were used without appropriate safeguards (e.g. if envelopes were unsealed or non-opaque or not sequentially numbered) • Alternation or rotation • Date of birth • Case record number • Any other explicitly unconcealed procedure	X			

3. Blinding: Was knowledge of the allocated interventions adequately prevented?*					
3.a. Were patients blinded?	**Examples of low risk of bias:** • No blinding but the review authors judge that the outcome and the outcome measurement are not likely influenced by lack of blinding • Blinding of participants and key study personnel ensured, and unlikely that blinding could have been broken • Either participants or some key study personnel were not blinded, but outcome assessment was blinded and the nonblinding of others unlikely to introduce bias **Examples of high risk of bias:** • No blinding or incomplete blinding, and the outcome or outcome measurement is likely to be influenced by lack of blinding • Blinding of key study participants and personnel attempted, but likely that the blinding could have been broken • Either participants or some key study personnel were not blinded, and the nonblinding of others likely to introduce bias				X
3.b. Were healthcare providers blinded?					X
3.c. Were data collectors blinded?		X			
3.d. Were outcome assessors blinded?		X			
3.e. Were data analysts blinded?				X	
4. Was loss to follow-up (missing outcome data) infrequent?	**Examples of low risk of bias:** • No missing outcome data • Reasons for missing outcome data unlikely to be related to outcome (for survival data, censoring unlikely to be introducing bias) • Missing outcome data balanced in numbers across intervention groups, with similar reasons for missing data across groups • For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk not enough to have an important impact on the intervention effect estimate • For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes not enough to have an important impact on observed effect size • Missing data have been imputed using appropriate methods **Examples of high risk of bias** • Reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across intervention groups • For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk enough to induce important bias in intervention effect estimate • For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes enough to induce clinically relevant bias in observed effect size • 'As-treated' analysis done with substantial departure of the intervention received from that assigned at randomization • Potentially inappropriate application of simple imputation	X			

5. Are reports of the study free of selective outcome reporting?*	**Examples of low risk of bias:** • The study protocol is available and all of the study's pre-specified (primary and secondary) outcomes that are of interest in the review have been reported in the pre-specified way • The study protocol is not available but it is clear that the published reports include all expected outcomes, including those that were pre-specified (convincing text of this nature may be uncommon) **Examples of high risk of bias** • Not all of the study's pre-specified primary outcomes have been reported • One or more primary outcomes is reported using measurements, analysis methods or subsets of the data (e.g. subscales) that were not pre-specified • One or more reported primary outcomes were not pre-specified (unless clear justification for their reporting is provided, such as an unexpected adverse effect) • One or more outcomes of interest in the review are reported incompletely so that they cannot be entered in a meta-analysis • The study report fails to include results for a key outcome that would be expected to have been reported for such a study	X			
6. Was the study apparently free of other problems that could put it at a risk of bias?*	**Examples of low risk of bias:** • The study appears to be free of other sources of bias **Examples of high risk of bias** • Had a potential source of bias related to the specific study design used • Stopped early due to some data-dependent process (including a formal-stopping rule) • Had extreme baseline imbalance • Has been claimed to have been fraudulent • Had some other problem		X		

Hall *et al.*, (2018a) [32]

Question	Example	Definitely Yes (Low RoB)	Probably Yes	Probably No	Definitely No (High RoB)
1. Was the allocation sequence adequately generated?*	**Examples of low risk of bias:** • Referring to a random number table • Using a computer random number generator • Coin tossing • Shuffling cards or envelopes • Throwing dice • Drawing of lots • Minimization with or without a random element **Examples of high risk of bias:** • Sequence generated by odd or even date of birth • Sequence generated by some rule based on date (or day) of admission • Sequence generated by some rule based on hospital or clinic record number • Allocation by judgement of the clinician • Allocation by preference of the participant • Allocation based on the results of a series laboratory test or series of tests • Allocation by availability of the intervention		X		
2. Was the allocation adequately concealed?	**Examples of low risk of bias allocation concealment techniques:** • Central allocation (including telephone, web-based, and pharmacy-controlled, randomization) **Examples of possible low risk of bias:** • Sequentially numbered drug containers of identical appearance • Sequentially numbered, opaque, sealed envelopes **Examples of high risk of bias allocation concealment techniques:** • Using an open random allocation schedule (e.g. a list of random numbers) • Assignment envelopes were used without appropriate safeguards (e.g. if envelopes were unsealed or non-opaque or not sequentially numbered) • Alternation or rotation • Date of birth • Case record number • Any other explicitly unconcealed procedure				X

3. Blinding: Was knowledge of the allocated interventions adequately prevented?*					
3.a. Were patients blinded?	**Examples of low risk of bias:** • No blinding but the review authors judge that the outcome and the outcome measurement are not likely influenced by lack of blinding • Blinding of participants and key study personnel ensured, and unlikely that blinding could have been broken • Either participants or some key study personnel were not blinded, but outcome assessment was blinded and the nonblinding of others unlikely to introduce bias **Examples of high risk of bias:** • No blinding or incomplete blinding, and the outcome or outcome measurement is likely to be influenced by lack of blinding • Blinding of key study participants and personnel attempted, but likely that the blinding could have been broken • Either participants or some key study personnel were not blinded, and the non-blinding of others likely to introduce bias				X
3.b. Were healthcare providers blinded?					X
3.c. Were data collectors blinded?					X
3.d. Were outcome assessors blinded?					X
3.e. Were data analysts blinded?				X	
4. Was loss to follow-up (missing outcome data) infrequent?	**Examples of low risk of bias:** • No missing outcome data • Reasons for missing outcome data unlikely to be related to outcome (for survival data, censoring unlikely to be introducing bias) • Missing outcome data balanced in numbers across intervention groups, with similar reasons for missing data across groups • For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk not enough to have an important impact on the intervention effect estimate • For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes not enough to have an important impact on observed effect size • Missing data have been imputed using appropriate methods **Examples of high risk of bias** • Reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across intervention groups • For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk enough to induce important bias in intervention effect estimate • For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes enough to induce clinically relevant bias in observed effect size • 'As-treated' analysis done with substantial departure of the intervention received from that assigned at randomization • Potentially inappropriate application of simple imputation	X			

5. Are reports of the study free of selective outcome reporting?*	**Examples of low risk of bias:** • The study protocol is available and all of the study's pre-specified (primary and secondary) outcomes that are of interest in the review have been reported in the pre-specified way • The study protocol is not available but it is clear that the published reports include all expected outcomes, including those that were pre-specified (convincing text of this nature may be uncommon) **Examples of high risk of bias** • Not all of the study's pre-specified primary outcomes have been reported • One or more primary outcomes is reported using measurements, analysis methods or subsets of the data (e.g. subscales) that were not pre-specified • One or more reported primary outcomes were not pre-specified (unless clear justification for their reporting is provided, such as an unexpected adverse effect) • One or more outcomes of interest in the review are reported incompletely so that they cannot be entered in a meta-analysis • The study report fails to include results for a key outcome that would be expected to have been reported for such a study	X			
6. Was the study apparently free of other problems that could put it at a risk of bias?*	**Examples of low risk of bias:** • The study appears to be free of other sources of bias **Examples of high risk of bias** • Had a potential source of bias related to the specific study design used • Stopped early due to some data-dependent process (including a formal-stopping rule) • Had extreme baseline imbalance • Has been claimed to have been fraudulent • Had some other problem		X		

Kamali *et al.*, (2017) [34]

Question	Example	Definitely Yes (Low RoB)	Probably Yes	Probably No	Definitely No (High RoB)
1. Was the allocation sequence adequately generated?*	**Examples of low risk of bias:** • Referring to a random number table • Using a computer random number generator • Coin tossing • Shuffling cards or envelopes • Throwing dice • Drawing of lots • Minimization with or without a random element **Examples of high risk of bias:** • Sequence generated by odd or even date of birth • Sequence generated by some rule based on date (or day) of admission • Sequence generated by some rule based on hospital or clinic record number • Allocation by judgement of the clinician • Allocation by preference of the participant • Allocation based on the results of a series laboratory test or series of tests • Allocation by availability of the intervention		X		
2. Was the allocation adequately concealed?	**Examples of low risk of bias allocation concealment techniques:** • Central allocation (including telephone, web-based, and pharmacy-controlled, randomization) **Examples of possible low risk of bias:** • Sequentially numbered drug containers of identical appearance • Sequentially numbered, opaque, sealed envelopes **Examples of high risk of bias allocation concealment techniques:** • Using an open random allocation schedule (e.g. a list of random numbers) • Assignment envelopes were used without appropriate safeguards (e.g. if envelopes were unsealed or non-opaque or not sequentially numbered) • Alternation or rotation • Date of birth • Case record number • Any other explicitly unconcealed procedure		X		

3. Blinding: Was knowledge of the allocated interventions adequately prevented?*					
3.a. Were patients blinded?	**Examples of low risk of bias:** • No blinding but the review authors judge that the outcome and the outcome measurement are not likely influenced by lack of blinding • Blinding of participants and key study personnel ensured, and unlikely that blinding could have been broken • Either participants or some key study personnel were not blinded, but outcome assessment was blinded and the nonblinding of others unlikely to introduce bias **Examples of high risk of bias:** • No blinding or incomplete blinding, and the outcome or outcome measurement is likely to be influenced by lack of blinding • Blinding of key study participants and personnel attempted, but likely that the blinding could have been broken • Either participants or some key study personnel were not blinded, and the non-blinding of others likely to introduce bias	X			
3.b. Were healthcare providers blinded?		X			
3.c. Were data collectors blinded?		X			
3.d. Were outcome assessors blinded?		X			
3.e. Were data analysts blinded?				X	
4. Was loss to follow-up (missing outcome data) infrequent?	**Examples of low risk of bias:** • No missing outcome data • Reasons for missing outcome data unlikely to be related to outcome (for survival data, censoring unlikely to be introducing bias) • Missing outcome data balanced in numbers across intervention groups, with similar reasons for missing data across groups • For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk not enough to have an important impact on the intervention effect estimate • For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes not enough to have an important impact on observed effect size • Missing data have been imputed using appropriate methods **Examples of high risk of bias** • Reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across intervention groups • For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk enough to induce important bias in intervention effect estimate • For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes enough to induce clinically relevant bias in observed effect size • 'As-treated' analysis done with substantial departure of the intervention received from that assigned at randomization • Potentially inappropriate application of simple imputation	X			

5. Are reports of the study free of selective outcome reporting?*	**Examples of low risk of bias:** • The study protocol is available and all of the study's pre-specified (primary and secondary) outcomes that are of interest in the review have been reported in the pre-specified way • The study protocol is not available but it is clear that the published reports include all expected outcomes, including those that were pre-specified (convincing text of this nature may be uncommon) **Examples of high risk of bias** • Not all of the study's pre-specified primary outcomes have been reported • One or more primary outcomes is reported using measurements, analysis methods or subsets of the data (e.g. subscales) that were not pre-specified • One or more reported primary outcomes were not pre-specified (unless clear justification for their reporting is provided, such as an unexpected adverse effect) • One or more outcomes of interest in the review are reported incompletely so that they cannot be entered in a meta-analysis • The study report fails to include results for a key outcome that would be expected to have been reported for such a study	X			
6. Was the study apparently free of other problems that could put it at a risk of bias?*	**Examples of low risk of bias:** • The study appears to be free of other sources of bias **Examples of high risk of bias** • Had a potential source of bias related to the specific study design used • Stopped early due to some data-dependent process (including a formal-stopping rule) • Had extreme baseline imbalance • Has been claimed to have been fraudulent • Had some other problem		X		

Sierra-Guzmán *et al.*, (2018) [40]

Question	Example	Definitely Yes (Low RoB)	Probably Yes	Probably No	Definitely No (High RoB)
1. Was the allocation sequence adequately generated?*	**Examples of low risk of bias:** • Referring to a random number table • Using a computer random number generator • Coin tossing • Shuffling cards or envelopes • Throwing dice • Drawing of lots • Minimization with or without a random element **Examples of high risk of bias:** • Sequence generated by odd or even date of birth • Sequence generated by some rule based on date (or day) of admission • Sequence generated by some rule based on hospital or clinic record number • Allocation by judgement of the clinician • Allocation by preference of the participant • Allocation based on the results of a series laboratory test or series of tests • Allocation by availability of the intervention	X			
2. Was the allocation adequately concealed?	**Examples of low risk of bias allocation concealment techniques:** • Central allocation (including telephone, web-based, and pharmacy-controlled, randomization) **Examples of possible low risk of bias:** • Sequentially numbered drug containers of identical appearance • Sequentially numbered, opaque, sealed envelopes **Examples of high risk of bias allocation concealment techniques:** • Using an open random allocation schedule (e.g. a list of random numbers) • Assignment envelopes were used without appropriate safeguards (e.g. if envelopes were unsealed or non-opaque or not sequentially numbered) • Alternation or rotation • Date of birth • Case record number • Any other explicitly unconcealed procedure	X			

3. Blinding: Was knowledge of the allocated interventions adequately prevented?*					
3.a. Were patients blinded?	**Examples of low risk of bias:** • No blinding but the review authors judge that the outcome and the outcome measurement are not likely influenced by lack of blinding • Blinding of participants and key study personnel ensured, and unlikely that blinding could have been broken • Either participants or some key study personnel were not blinded, but outcome assessment was blinded and the nonblinding of others unlikely to introduce bias **Examples of high risk of bias:** • No blinding or incomplete blinding, and the outcome or outcome measurement is likely to be influenced by lack of blinding • Blinding of key study participants and personnel attempted, but likely that the blinding could have been broken • Either participants or some key study personnel were not blinded, and the non-blinding of others likely to introduce bias				X
3.b. Were healthcare providers blinded?					X
3.c. Were data collectors blinded?		X			
3.d. Were outcome assessors blinded?		X			
3.e. Were data analysts blinded?				X	
4. Was loss to follow-up (missing outcome data) infrequent?	**Examples of low risk of bias:** • No missing outcome data • Reasons for missing outcome data unlikely to be related to outcome (for survival data, censoring unlikely to be introducing bias) • Missing outcome data balanced in numbers across intervention groups, with similar reasons for missing data across groups • For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk not enough to have an important impact on the intervention effect estimate • For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes not enough to have an important impact on observed effect size • Missing data have been imputed using appropriate methods **Examples of high risk of bias** • Reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across intervention groups • For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk enough to induce important bias in intervention effect estimate • For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes enough to induce clinically relevant bias in observed effect size • 'As-treated' analysis done with substantial departure of the intervention received from that assigned at randomization • Potentially inappropriate application of simple imputation	X			

5. Are reports of the study free of selective outcome reporting?*	**Examples of low risk of bias:** • The study protocol is available and all of the study's pre-specified (primary and secondary) outcomes that are of interest in the review have been reported in the pre-specified way • The study protocol is not available but it is clear that the published reports include all expected outcomes, including those that were pre-specified (convincing text of this nature may be uncommon) **Examples of high risk of bias** • Not all of the study's pre-specified primary outcomes have been reported • One or more primary outcomes is reported using measurements, analysis methods or subsets of the data (e.g. subscales) that were not pre-specified • One or more reported primary outcomes were not pre-specified (unless clear justification for their reporting is provided, such as an unexpected adverse effect) • One or more outcomes of interest in the review are reported incompletely so that they cannot be entered in a meta-analysis • The study report fails to include results for a key outcome that would be expected to have been reported for such a study	X			
6. Was the study apparently free of other problems that could put it at a risk of bias?*	**Examples of low risk of bias:** • The study appears to be free of other sources of bias **Examples of high risk of bias** • Had a potential source of bias related to the specific study design used • Stopped early due to some data-dependent process (including a formal-stopping rule) • Had extreme baseline imbalance • Has been claimed to have been fraudulent • Had some other problem		X		

Tool to Assess Risk of Bias (RoB) in Cohort Studies

Bagherian *et al.* (2018) [24]

Question	Example	Definitely Yes (Low RoB)	Probably Yes	Probably No	Definitely No (High RoB)
1. Was selection of exposed and non-exposed cohorts drawn from the same population?	**Low RoB:** - Exposed and unexposed drawn for same administrative data base of patients presenting at same points of care over the same time frame **High RoB:** - Exposed and unexposed presenting to different points of care over a different time frame	x			

2. Can we be confident in the assessment of exposure?	**Low RoB:** - Secure record (e.g. surgical records, pharmacy records) - Repeated interview or other ascertainment asking about current use/exposure **Higher RoB:** - Structured interview at a single point in time - Written self-report - Individuals who are asked retrospectively confirm their exposure status may be subject to recall bias - less likely to recall an exposure if they have not developed an adverse outcome, and more likely to recall an exposure (whether an exposure occurred or not) if they have developed an adverse outcome **High RoB:** - Uncertain how exposure information obtained		x		
3. Can we be confident that the outcome of interest was not present at start of study?	**NA**		x		
4. Did the study match exposed and unexposed for all variables that are associated with the outcome of interest or did the statistical analysis adjust for these prognostic variables?	**Low RoB:** - Comprehensive matching or adjustment for all plausible prognostic variables **Higher RoB:** - Matching or adjustment for most plausible prognostic variables **HIGH RoB:** - Matching or adjustment for a minority of plausible prognostic variables - No matching or adjustment of plausible prognostic variables - Statements of no differences between groups - Statements that differences were not statistically significant are not sufficient for establishing comparability	x			

5. Can we be confident in the assessment of the presence or absence of prognostic factors?	**Low RoB:** - Interview of all participants - Self-completed survey from all participants - Review of charts with reproducibility demonstrated - From data base with documentation of accuracy of abstraction of prognostic data **Higher RoB:** - Chart review without demonstration of reproducibility - Data base with uncertain quality of abstraction of prognostic information **High RoB:** Prognostic information from data base with no available documentation of quality of abstraction of prognostic variables	x			
6. Can we be confident in the assessment of outcome?	**Low RoB:** - Independent blind assessment - Record linkage - For some outcomes (e.g. fractured hip), reference to the medical record is sufficient to satisfy the requirement for confirmation of the fracture **Higher RoB:** - Independent assessment unblinded - Self-report - For some outcomes (e.g. vertebral fracture where reference to x-rays would be required) reference to the medical record would not be adequate outcomes **High RoB:** - Uncertain (no description)		x		

7. Was the follow up of cohorts adequate?	**Low RoB:** - No missing outcome data - Reasons for missing outcome data unlikely to be related to true outcome (for survival data, censoring is unlikely to introduce bias) - Missing outcome data balanced in numbers across intervention groups, with similar reasons for missing data across groups - For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk is not enough to have an important impact on the intervention effect estimate - For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes is not large enough to have an important impact on the observed effect size - Missing data have been imputed using appropriated methods **High RoB:** - Reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across intervention groups - For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk is enough to induce important bias in intervention effect estimate - For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes is large enough to induce clinically relevant bias in the observed effect size	x			

8. Were co-interventions similar between groups?	**Low RoB:** - Most or all relevant co-interventions that might influence the outcome of interest are documented to be similar in the exposed and unexposed **High RoB:** - Few or no relevant co-interventions that might influence the outcome of interest are documented to be similar in the exposed and unexposed	x			

Cho *et al.*, (2019) [28]

Question	Example	Definitely Yes (Low RoB)	Probably Yes	Probably No	Definitely No (High RoB)
1. Was selection of exposed and non-exposed cohorts drawn from the same population?	**Low RoB:** - Exposed and unexposed drawn for same administrative data base of patients presenting at same points of care over the same time frame **High RoB:** - Exposed and unexposed presenting to different points of care over a different time frame	x			
2. Can we be confident in the assessment of exposure?	**Low RoB:** - Secure record (e.g. surgical records, pharmacy records) - Repeated interview or other ascertainment asking about current use/exposure **Higher RoB:** - Structured interview at a single point in time - Written self-report - Individuals who are asked retrospectively confirm their exposure status may be subject to recall bias - less likely to recall an exposure if they have not developed an adverse outcome, and more likely to recall an exposure (whether an exposure occurred or not) if they have developed an adverse outcome **High RoB:** - Uncertain how exposure information obtained	x			

3. Can we be confident that the outcome of interest was not present at start of study?	**NA**	x			
4. Did the study match exposed and unexposed for all variables that are associated with the outcome of interest or did the statistical analysis adjust for these prognostic variables?	**Low RoB:** - Comprehensive matching or adjustment for all plausible prognostic variables **Higher RoB:** - Matching or adjustment for most plausible prognostic variables **HIGH RoB:** - Matching or adjustment for a minority of plausible prognostic variables - No matching or adjustment of plausible prognostic variables - Statements of no differences between groups - Statements that differences were not statistically significant are not sufficient for establishing comparability				x
5. Can we be confident in the assessment of the presence or absence of prognostic factors?	**Low RoB:** - Interview of all participants - Self-completed survey from all participants - Review of charts with reproducibility demonstrated - From data base with documentation of accuracy of abstraction of prognostic data **Higher RoB:** - Chart review without demonstration of reproducibility - Data base with uncertain quality of abstraction of prognostic information **High RoB:** Prognostic information from data base with no available documentation of quality of abstraction of prognostic variables		x		

6. Can we be confident in the assessment of outcome?	**Low RoB:** - Independent blind assessment - Record linkage - For some outcomes (e.g. fractured hip), reference to the medical record is sufficient to satisfy the requirement for confirmation of the fracture **Higher RoB:** - Independent assessment unblinded - Self-report - For some outcomes (e.g. vertebral fracture where reference to x-rays would be required) reference to the medical record would not be adequate outcomes **High RoB:** - Uncertain (no description)		x		
7. Was the follow up of cohorts adequate?	**Low RoB:** - No missing outcome data - Reasons for missing outcome data unlikely to be related to true outcome (for survival data, censoring is unlikely to introduce bias) - Missing outcome data balanced in numbers across intervention groups, with similar reasons for missing data across groups - For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk is not enough to have an important impact on the intervention effect estimate - For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes is not large enough to have an important impact on the observed effect size - Missing data have been imputed using appropriated methods **High RoB:** - Reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across intervention groups - For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk is enough to induce important bias in intervention effect estimate - For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes is large enough to induce clinically relevant bias in the observed effect size	x			

8. Were co-interventions similar between groups?	**Low RoB:** - Most or all relevant co-interventions that might influence the outcome of interest are documented to be similar in the exposed and unexposed **High RoB:** - Few or no relevant co-interventions that might influence the outcome of interest are documented to be similar in the exposed and unexposed				x

Coetzee *et al.*, (2018) [53]

Question	Example	Definitely Yes (Low RoB)	Probably Yes	Probably No	Definitely No (High RoB)
1. Was selection of exposed and non-exposed cohorts drawn from the same population?	**Low RoB:** - Exposed and unexposed drawn for same administrative data base of patients presenting at same points of care over the same time frame **High RoB:** - Exposed and unexposed presenting to different points of care over a different time frame		x		
2. Can we be confident in the assessment of exposure?	**Low RoB:** - Secure record (e.g. surgical records, pharmacy records) - Repeated interview or other ascertainment asking about current use/exposure **Higher RoB:** - Structured interview at a single point in time - Written self-report - Individuals who are asked retrospectively confirm their exposure status may be subject to recall bias - less likely to recall an exposure if they have not developed an adverse outcome, and more likely to recall an exposure (whether an exposure occurred or not) if they have developed an adverse outcome **High RoB:** - Uncertain how exposure information obtained	x			

3. Can we be confident that the outcome of interest was not present at start of study?	**NA**		x		
4. Did the study match exposed and unexposed for all variables that are associated with the outcome of interest or did the statistical analysis adjust for these prognostic variables?	**Low RoB:** - Comprehensive matching or adjustment for all plausible prognostic variables **Higher RoB:** - Matching or adjustment for most plausible prognostic variables **HIGH RoB:** - Matching or adjustment for a minority of plausible prognostic variables - No matching or adjustment of plausible prognostic variables - Statements of no differences between groups - Statements that differences were not statistically significant are not sufficient for establishing comparability		x		
5. Can we be confident in the assessment of the presence or absence of prognostic factors?	**Low RoB:** - Interview of all participants - Self-completed survey from all participants - Review of charts with reproducibility demonstrated - From data base with documentation of accuracy of abstraction of prognostic data **Higher RoB:** - Chart review without demonstration of reproducibility - Data base with uncertain quality of abstraction of prognostic information **High RoB:** Prognostic information from data base with no available documentation of quality of abstraction of prognostic variables		x		

6. Can we be confident in the assessment of outcome?	**Low RoB:** - Independent blind assessment - Record linkage - For some outcomes (e.g. fractured hip), reference to the medical record is sufficient to satisfy the requirement for confirmation of the fracture **Higher RoB:** - Independent assessment unblinded - Self-report - For some outcomes (e.g. vertebral fracture where reference to x-rays would be required) reference to the medical record would not be adequate outcomes **High RoB:** - Uncertain (no description)		x		
7. Was the follow up of cohorts adequate?	**Low RoB:** - No missing outcome data - Reasons for missing outcome data unlikely to be related to true outcome (for survival data, censoring is unlikely to introduce bias) - Missing outcome data balanced in numbers across intervention groups, with similar reasons for missing data across groups - For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk is not enough to have an important impact on the intervention effect estimate - For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes is not large enough to have an important impact on the observed effect size - Missing data have been imputed using appropriated methods **High RoB:** - Reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across intervention groups - For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk is enough to induce important bias in intervention effect estimate - For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes is large enough to induce clinically relevant bias in the observed effect size		x		

8. Were co-interventions similar between groups?	**Low RoB:** - Most or all relevant co-interventions that might influence the outcome of interest are documented to be similar in the exposed and unexposed **High RoB:** - Few or no relevant co-interventions that might influence the outcome of interest are documented to be similar in the exposed and unexposed				x

Fereydounnia *et al.*, (2019) [30]

Question	Example	Definitely Yes (Low RoB)	Probably Yes	Probably No	Definitely No (High RoB)
1. Was selection of exposed and non-exposed cohorts drawn from the same population?	**Low RoB:** - Exposed and unexposed drawn for same administrative data base of patients presenting at same points of care over the same time frame **High RoB:** - Exposed and unexposed presenting to different points of care over a different time frame		X		
2. Can we be confident in the assessment of exposure?	**Low RoB:** - Secure record (e.g. surgical records, pharmacy records) - Repeated interview or other ascertainment asking about current use/exposure **Higher RoB:** - Structured interview at a single point in time - Written self-report - Individuals who are asked retrospectively confirm their exposure status may be subject to recall bias - less likely to recall an exposure if they have not developed an adverse outcome, and more likely to recall an exposure (whether an exposure occurred or not) if they have developed an adverse outcome **High RoB:** - Uncertain how exposure information obtained		X		

3. Can we be confident that the outcome of interest was not present at start of study?	**NA**		X		
4. Did the study match exposed and unexposed for all variables that are associated with the outcome of interest or did the statistical analysis adjust for these prognostic variables?	**Low RoB:** - Comprehensive matching or adjustment for all plausible prognostic variables **Higher RoB:** - Matching or adjustment for most plausible prognostic variables **HIGH RoB:** - Matching or adjustment for a minority of plausible prognostic variables - No matching or adjustment of plausible prognostic variables - Statements of no differences between groups - Statements that differences were not statistically significant are not sufficient for establishing comparability	X			
5. Can we be confident in the assessment of the presence or absence of prognostic factors?	**Low RoB:** - Interview of all participants - Self-completed survey from all participants - Review of charts with reproducibility demonstrated - From data base with documentation of accuracy of abstraction of prognostic data **Higher RoB:** - Chart review without demonstration of reproducibility - Data base with uncertain quality of abstraction of prognostic information **High RoB:** Prognostic information from data base with no available documentation of quality of abstraction of prognostic variables		X		

6. Can we be confident in the assessment of outcome?	**Low RoB:** - Independent blind assessment - Record linkage - For some outcomes (e.g. fractured hip), reference to the medical record is sufficient to satisfy the requirement for confirmation of the fracture **Higher RoB:** - Independent assessment unblinded - Self-report - For some outcomes (e.g. vertebral fracture where reference to x-rays would be required) reference to the medical record would not be adequate outcomes **High RoB:** - Uncertain (no description)	X			
7. Was the follow up of cohorts adequate?	**Low RoB:** - No missing outcome data - Reasons for missing outcome data unlikely to be related to true outcome (for survival data, censoring is unlikely to introduce bias) - Missing outcome data balanced in numbers across intervention groups, with similar reasons for missing data across groups - For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk is not enough to have an important impact on the intervention effect estimate - For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes is not large enough to have an important impact on the observed effect size - Missing data have been imputed using appropriated methods **High RoB:** - Reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across intervention groups - For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk is enough to induce important bias in intervention effect estimate - For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes is large enough to induce clinically relevant bias in the observed effect size	X			

8. Were co-interventions similar between groups?	**Low RoB:** - Most or all relevant co-interventions that might influence the outcome of interest are documented to be similar in the exposed and unexposed **High RoB:** - Few or no relevant co-interventions that might influence the outcome of interest are documented to be similar in the exposed and unexposed	X			

Golditz *et al.*, (2016) [31]

Question	Example	Definitely Yes (Low RoB)	Probably Yes	Probably No	Definitely No (High RoB)
1. Was selection of exposed and non-exposed cohorts drawn from the same population?	**Low RoB:** - Exposed and unexposed drawn for same administrative data base of patients presenting at same points of care over the same time frame **High RoB:** - Exposed and unexposed presenting to different points of care over a different time frame		x		
2. Can we be confident in the assessment of exposure?	**Low RoB:** - Secure record (e.g. surgical records, pharmacy records) - Repeated interview or other ascertainment asking about current use/exposure **Higher RoB:** - Structured interview at a single point in time - Written self-report - Individuals who are asked retrospectively confirm their exposure status may be subject to recall bias - less likely to recall an exposure if they have not developed an adverse outcome, and more likely to recall an exposure (whether an exposure occurred or not) if they have developed an adverse outcome **High RoB:** - Uncertain how exposure information obtained	x			

3. Can we be confident that the outcome of interest was not present at start of study?	**NA**	x			
4. Did the study match exposed and unexposed for all variables that are associated with the outcome of interest or did the statistical analysis adjust for these prognostic variables?	**Low RoB:** - Comprehensive matching or adjustment for all plausible prognostic variables **Higher RoB:** - Matching or adjustment for most plausible prognostic variables **HIGH RoB:** - Matching or adjustment for a minority of plausible prognostic variables - No matching or adjustment of plausible prognostic variables - Statements of no differences between groups - Statements that differences were not statistically significant are not sufficient for establishing comparability	x			
5. Can we be confident in the assessment of the presence or absence of prognostic factors?	**Low RoB:** - Interview of all participants - Self-completed survey from all participants - Review of charts with reproducibility demonstrated - From data base with documentation of accuracy of abstraction of prognostic data **Higher RoB:** - Chart review without demonstration of reproducibility - Data base with uncertain quality of abstraction of prognostic information **High RoB:** Prognostic information from data base with no available documentation of quality of abstraction of prognostic variables	x			

6. Can we be confident in the assessment of outcome?	**Low RoB:** - Independent blind assessment - Record linkage - For some outcomes (e.g. fractured hip), reference to the medical record is sufficient to satisfy the requirement for confirmation of the fracture **Higher RoB:** - Independent assessment unblinded - Self-report - For some outcomes (e.g. vertebral fracture where reference to x-rays would be required) reference to the medical record would not be adequate outcomes **High RoB:** - Uncertain (no description)		x		
7. Was the follow up of cohorts adequate?	**Low RoB:** - No missing outcome data - Reasons for missing outcome data unlikely to be related to true outcome (for survival data, censoring is unlikely to introduce bias) - Missing outcome data balanced in numbers across intervention groups, with similar reasons for missing data across groups - For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk is not enough to have an important impact on the intervention effect estimate - For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes is not large enough to have an important impact on the observed effect size - Missing data have been imputed using appropriated methods **High RoB:** - Reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across intervention groups - For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk is enough to induce important bias in intervention effect estimate	x			

	- For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes is large enough to induce clinically relevant bias in the observed effect size				
8. Were co-interventions similar between groups?	**Low RoB:** - Most or all relevant co-interventions that might influence the outcome of interest are documented to be similar in the exposed and unexposed **High RoB:** - Few or no relevant co-interventions that might influence the outcome of interest are documented to be similar in the exposed and unexposed	x			

Harriss *et al*, (2019) [33]

Question	Example	Definitely Yes (Low RoB)	Probably Yes	Probably No	Definitely No (High RoB)
1. Was selection of exposed and non-exposed cohorts drawn from the same population?	**Low RoB:** - Exposed and unexposed drawn for same administrative data base of patients presenting at same points of care over the same time frame **High RoB:** - Exposed and unexposed presenting to different points of care over a different time frame	x			
2. Can we be confident in the assessment of exposure?	**Low RoB:** - Secure record (e.g. surgical records, pharmacy records) - Repeated interview or other ascertainment asking about current use/exposure **Higher RoB:** - Structured interview at a single point in time - Written self-report - Individuals who are asked retrospectively confirm their exposure status may be subject to recall bias - less likely to recall an exposure if they have not developed an adverse outcome, and more likely to recall an exposure (whether an exposure occurred or not) if they have developed an adverse outcome	x			

	High RoB: - Uncertain how exposure information obtained				
3. Can we be confident that the outcome of interest was not present at start of study?	**NA**	x			
4. Did the study match exposed and unexposed for all variables that are associated with the outcome of interest or did the statistical analysis adjust for these prognostic variables?	**Low RoB:** - Comprehensive matching or adjustment for all plausible prognostic variables **Higher RoB:** - Matching or adjustment for most plausible prognostic variables **HIGH RoB:** - Matching or adjustment for a minority of plausible prognostic variables - No matching or adjustment of plausible prognostic variables - Statements of no differences between groups - Statements that differences were not statistically significant are not sufficient for establishing comparability	x			
5. Can we be confident in the assessment of the presence or absence of prognostic factors?	**Low RoB:** - Interview of all participants - Self-completed survey from all participants - Review of charts with reproducibility demonstrated - From data base with documentation of accuracy of abstraction of prognostic data **Higher RoB:** - Chart review without demonstration of reproducibility - Data base with uncertain quality of abstraction of prognostic information **High RoB:** Prognostic information from data base with no available documentation of quality of abstraction of prognostic variables		x		

6. Can we be confident in the assessment of outcome?	**Low RoB:** - Independent blind assessment - Record linkage - For some outcomes (e.g. fractured hip), reference to the medical record is sufficient to satisfy the requirement for confirmation of the fracture **Higher RoB:** - Independent assessment unblinded - Self-report - For some outcomes (e.g. vertebral fracture where reference to x-rays would be required) reference to the medical record would not be adequate outcomes **High RoB:** - Uncertain (no description)	x			
7. Was the follow up of cohorts adequate?	**Low RoB:** - No missing outcome data - Reasons for missing outcome data unlikely to be related to true outcome (for survival data, censoring is unlikely to introduce bias) - Missing outcome data balanced in numbers across intervention groups, with similar reasons for missing data across groups - For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk is not enough to have an important impact on the intervention effect estimate - For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes is not large enough to have an important impact on the observed effect size - Missing data have been imputed using appropriated methods **High RoB:** - Reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across intervention groups - For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk is enough to induce important bias in intervention effect estimate - For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes is large enough to induce clinically relevant bias in the observed effect size	x			

8. Were co-interventions similar between groups?	**Low RoB:** - Most or all relevant co-interventions that might influence the outcome of interest are documented to be similar in the exposed and unexposed **High RoB:** - Few or no relevant co-interventions that might influence the outcome of interest are documented to be similar in the exposed and unexposed	x			

Ko *et al.*, (2018) [35]

Question	Example	Definitely Yes (Low RoB)	Probably Yes	Probably No	Definitely No (High RoB)
1. Was selection of exposed and non-exposed cohorts drawn from the same population?	**Low RoB:** - Exposed and unexposed drawn for same administrative data base of patients presenting at same points of care over the same time frame **High RoB:** - Exposed and unexposed presenting to different points of care over a different time frame		x		
2. Can we be confident in the assessment of exposure?	**Low RoB:** - Secure record (e.g. surgical records, pharmacy records) - Repeated interview or other ascertainment asking about current use/exposure **Higher RoB:** - Structured interview at a single point in time - Written self-report - Individuals who are asked retrospectively confirm their exposure status may be subject to recall bias - less likely to recall an exposure if they have not developed an adverse outcome, and more likely to recall an exposure (whether an exposure occurred or not) if they have developed an adverse outcome **High RoB:** - Uncertain how exposure information obtained	x			

3. Can we be confident that the outcome of interest was not present at start of study?	**NA**		x		
4. Did the study match exposed and unexposed for all variables that are associated with the outcome of interest or did the statistical analysis adjust for these prognostic variables?	**Low RoB:** - Comprehensive matching or adjustment for all plausible prognostic variables **Higher RoB:** - Matching or adjustment for most plausible prognostic variables **HIGH RoB:** - Matching or adjustment for a minority of plausible prognostic variables - No matching or adjustment of plausible prognostic variables - Statements of no differences between groups - Statements that differences were not statistically significant are not sufficient for establishing comparability	x			
5. Can we be confident in the assessment of the presence or absence of prognostic factors?	**Low RoB:** - Interview of all participants - Self-completed survey from all participants - Review of charts with reproducibility demonstrated - From data base with documentation of accuracy of abstraction of prognostic data **Higher RoB:** - Chart review without demonstration of reproducibility - Data base with uncertain quality of abstraction of prognostic information **High RoB:** Prognostic information from data base with no available documentation of quality of abstraction of prognostic variables		x		

6. Can we be confident in the assessment of outcome?	**Low RoB:** - Independent blind assessment - Record linkage - For some outcomes (e.g. fractured hip), reference to the medical record is sufficient to satisfy the requirement for confirmation of the fracture **Higher RoB:** - Independent assessment unblinded - Self-report - For some outcomes (e.g. vertebral fracture where reference to x-rays would be required) reference to the medical record would not be adequate outcomes **High RoB:** - Uncertain (no description)	x			
7. Was the follow up of cohorts adequate?	**Low RoB:** - No missing outcome data - Reasons for missing outcome data unlikely to be related to true outcome (for survival data, censoring is unlikely to introduce bias) - Missing outcome data balanced in numbers across intervention groups, with similar reasons for missing data across groups - For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk is not enough to have an important impact on the intervention effect estimate - For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes is not large enough to have an important impact on the observed effect size - Missing data have been imputed using appropriated methods **High RoB:** - Reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across intervention groups - For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk is enough to induce important bias in intervention effect estimate - For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes is large enough to induce clinically relevant bias in the observed effect size		x		

8. Were co-interventions similar between groups?	**Low RoB:** - Most or all relevant co-interventions that might influence the outcome of interest are documented to be similar in the exposed and unexposed **High RoB:** - Few or no relevant co-interventions that might influence the outcome of interest are documented to be similar in the exposed and unexposed	x			

Madsen *et al.*, (2018) [36]

Question	Example	Definitely Yes (Low RoB)	Probably Yes	Probably No	Definitely No (High RoB)
1. Was selection of exposed and non-exposed cohorts drawn from the same population?	**Low RoB:** - Exposed and unexposed drawn for same administrative data base of patients presenting at same points of care over the same time frame **High RoB:** - Exposed and unexposed presenting to different points of care over a different time frame		x		
2. Can we be confident in the assessment of exposure?	**Low RoB:** - Secure record (e.g. surgical records, pharmacy records) - Repeated interview or other ascertainment asking about current use/exposure **Higher RoB:** - Structured interview at a single point in time - Written self-report - Individuals who are asked retrospectively confirm their exposure status may be subject to recall bias - less likely to recall an exposure if they have not developed an adverse outcome, and more likely to recall an exposure (whether an exposure occurred or not) if they have developed an adverse outcome **High RoB:** - Uncertain how exposure information obtained	x			

3. Can we be confident that the outcome of interest was not present at start of study?	**NA**		x		
4. Did the study match exposed and unexposed for all variables that are associated with the outcome of interest or did the statistical analysis adjust for these prognostic variables?	**Low RoB:** - Comprehensive matching or adjustment for all plausible prognostic variables **Higher RoB:** - Matching or adjustment for most plausible prognostic variables **HIGH RoB:** - Matching or adjustment for a minority of plausible prognostic variables - No matching or adjustment of plausible prognostic variables - Statements of no differences between groups - Statements that differences were not statistically significant are not sufficient for establishing comparability	x			
5. Can we be confident in the assessment of the presence or absence of prognostic factors?	**Low RoB:** - Interview of all participants - Self-completed survey from all participants - Review of charts with reproducibility demonstrated - From data base with documentation of accuracy of abstraction of prognostic data **Higher RoB:** - Chart review without demonstration of reproducibility - Data base with uncertain quality of abstraction of prognostic information **High RoB:** Prognostic information from data base with no available documentation of quality of abstraction of prognostic variables	x			

6. Can we be confident in the assessment of outcome?	**Low RoB:** - Independent blind assessment - Record linkage - For some outcomes (e.g. fractured hip), reference to the medical record is sufficient to satisfy the requirement for confirmation of the fracture **Higher RoB:** - Independent assessment unblinded - Self-report - For some outcomes (e.g. vertebral fracture where reference to x-rays would be required) reference to the medical record would not be adequate outcomes **High RoB:** - Uncertain (no description)	x			
7. Was the follow up of cohorts adequate?	**Low RoB:** - No missing outcome data - Reasons for missing outcome data unlikely to be related to true outcome (for survival data, censoring is unlikely to introduce bias) - Missing outcome data balanced in numbers across intervention groups, with similar reasons for missing data across groups - For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk is not enough to have an important impact on the intervention effect estimate - For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes is not large enough to have an important impact on the observed effect size - Missing data have been imputed using appropriated methods **High RoB:** - Reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across intervention groups	x			

	- For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes is large enough to induce clinically relevant bias in the observed effect size				
8. Were co-interventions similar between groups?	**Low RoB:** - Most or all relevant co-interventions that might influence the outcome of interest are documented to be similar in the exposed and unexposed **High RoB:** - Few or no relevant co-interventions that might influence the outcome of interest are documented to be similar in the exposed and unexposed	x			

McCann *et al.*, (2018) [60]

Question	Example	Definitely Yes (Low RoB)	Probably Yes	Probably No	Definitely No (High RoB)
1. Was selection of exposed and non-exposed cohorts drawn from the same population?	**Low RoB:** - Exposed and unexposed drawn for same administrative data base of patients presenting at same points of care over the same time frame **High RoB:** - Exposed and unexposed presenting to different points of care over a different time frame		X		
2. Can we be confident in the assessment of exposure?	**Low RoB:** - Secure record (e.g. surgical records, pharmacy records) - Repeated interview or other ascertainment asking about current use/exposure **Higher RoB:** - Structured interview at a single point in time - Written self-report - Individuals who are asked retrospectively confirm their exposure status may be subject to recall bias - less likely to recall an exposure if they have not developed an adverse outcome, and more likely to recall an exposure (whether an exposure occurred or not) if they have developed an adverse outcome		X		

	High RoB: - Uncertain how exposure information obtained				
3. Can we be confident that the outcome of interest was not present at start of study?	**NA**		X		
4. Did the study match exposed and unexposed for all variables that are associated with the outcome of interest or did the statistical analysis adjust for these prognostic variables?	**Low RoB:** - Comprehensive matching or adjustment for all plausible prognostic variables **Higher RoB:** - Matching or adjustment for most plausible prognostic variables **HIGH RoB:** - Matching or adjustment for a minority of plausible prognostic variables - No matching or adjustment of plausible prognostic variables - Statements of no differences between groups - Statements that differences were not statistically significant are not sufficient for establishing comparability				X
5. Can we be confident in the assessment of the presence or absence of prognostic factors?	**Low RoB:** - Interview of all participants - Self-completed survey from all participants - Review of charts with reproducibility demonstrated - From data base with documentation of accuracy of abstraction of prognostic data **Higher RoB:** - Chart review without demonstration of reproducibility Data base with uncertain quality of abstraction of prognostic information **High RoB:** Prognostic information from data base with no available documentation of quality of abstraction of prognostic variables		X		

6. Can we be confident in the assessment of outcome?	**Low RoB:** - Independent blind assessment - Record linkage - For some outcomes (e.g. fractured hip), reference to the medical record is sufficient to satisfy the requirement for confirmation of the fracture **Higher RoB:** - Independent assessment unblinded - Self-report - For some outcomes (e.g. vertebral fracture where reference to x-rays would be required) reference to the medical record would not be adequate outcomes **High RoB:** - Uncertain (no description)	X			
7. Was the follow up of cohorts adequate?	**Low RoB:** - No missing outcome data - Reasons for missing outcome data unlikely to be related to true outcome (for survival data, censoring is unlikely to introduce bias) - Missing outcome data balanced in numbers across intervention groups, with similar reasons for missing data across groups - For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk is not enough to have an important impact on the intervention effect estimate - For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes is not large enough to have an important impact on the observed effect size - Missing data have been imputed using appropriated methods **High RoB:** - Reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across intervention groups - For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk is enough to induce important bias in intervention effect estimate - For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes is large enough to induce clinically relevant bias in the observed effect size	X			

8. Were co-interventions similar between groups?	**Low RoB:** - Most or all relevant co-interventions that might influence the outcome of interest are documented to be similar in the exposed and unexposed **High RoB:** - Few or no relevant co-interventions that might influence the outcome of interest are documented to be similar in the exposed and unexposed		X		

Powden *et al*, (2019) [37]

Question	Example	Definitely Yes (Low RoB)	Probably Yes	Probably No	Definitely No (High RoB)
1. Was selection of exposed and non-exposed cohorts drawn from the same population?	**Low RoB:** - Exposed and unexposed drawn for same administrative data base of patients presenting at same points of care over the same time frame **High RoB:** - Exposed and unexposed presenting to different points of care over a different time frame	X			
2. Can we be confident in the assessment of exposure?	**Low RoB:** - Secure record (e.g. surgical records, pharmacy records) - Repeated interview or other ascertainment asking about current use/exposure **Higher RoB:** - Structured interview at a single point in time - Written self-report - Individuals who are asked retrospectively confirm their exposure status may be subject to recall bias - less likely to recall an exposure if they have not developed an adverse outcome, and more likely to recall an exposure (whether an exposure occurred or not) if they have developed an adverse outcome **High RoB:** - Uncertain how exposure information obtained		X		

3. Can we be confident that the outcome of interest was not present at start of study?	**NA**		X		
4. Did the study match exposed and unexposed for all variables that are associated with the outcome of interest or did the statistical analysis adjust for these prognostic variables?	**Low RoB:** - Comprehensive matching or adjustment for all plausible prognostic variables **Higher RoB:** - Matching or adjustment for most plausible prognostic variables **HIGH RoB:** - Matching or adjustment for a minority of plausible prognostic variables - No matching or adjustment of plausible prognostic variables - Statements of no differences between groups - Statements that differences were not statistically significant are not sufficient for establishing comparability	X			
5. Can we be confident in the assessment of the presence or absence of prognostic factors?	**Low RoB:** - Interview of all participants - Self-completed survey from all participants - Review of charts with reproducibility demonstrated - From data base with documentation of accuracy of abstraction of prognostic data **Higher RoB:** - Chart review without demonstration of reproducibility - Data base with uncertain quality of abstraction of prognostic information **High RoB:** Prognostic information from data base with no available documentation of quality of abstraction of prognostic variables	X			

6. Can we be confident in the assessment of outcome?	**Low RoB:** - Independent blind assessment - Record linkage - For some outcomes (e.g. fractured hip), reference to the medical record is sufficient to satisfy the requirement for confirmation of the fracture **Higher RoB:** - Independent assessment unblinded - Self-report - For some outcomes (e.g. vertebral fracture where reference to x-rays would be required) reference to the medical record would not be adequate outcomes **High RoB:** - Uncertain (no description)		X		
7. Was the follow up of cohorts adequate?	**Low RoB:** - No missing outcome data - Reasons for missing outcome data unlikely to be related to true outcome (for survival data, censoring is unlikely to introduce bias) - Missing outcome data balanced in numbers across intervention groups, with similar reasons for missing data across groups - For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk is not enough to have an important impact on the intervention effect estimate - For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes is not large enough to have an important impact on the observed effect size - Missing data have been imputed using appropriated methods **High RoB:** - Reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across intervention groups - For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk is enough to induce important bias in intervention effect estimate - For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes is large enough to induce clinically relevant bias in the observed effect size	X			

8. Were co-interventions similar between groups?	**Low RoB:** - Most or all relevant co-interventions that might influence the outcome of interest are documented to be similar in the exposed and unexposed **High RoB:** - Few or no relevant co-interventions that might influence the outcome of interest are documented to be similar in the exposed and unexposed	X			

Ryu *et al.* (2019) [38]

Question	Example	Definitely Yes (Low RoB)	Probably Yes	Probably No	Definitely No (High RoB)
1. Was selection of exposed and non-exposed cohorts drawn from the same population?	**Low RoB:** - Exposed and unexposed drawn for same administrative data base of patients presenting at same points of care over the same time frame **High RoB:** - Exposed and unexposed presenting to different points of care over a different time frame	X			
2. Can we be confident in the assessment of exposure?	**Low RoB:** - Secure record (e.g. surgical records, pharmacy records) - Repeated interview or other ascertainment asking about current use/exposure **Higher RoB:** - Structured interview at a single point in time - Written self-report - Individuals who are asked retrospectively confirm their exposure status may be subject to recall bias - less likely to recall an exposure if they have not developed an adverse outcome, and more likely to recall an exposure (whether an exposure occurred or not) if they have developed an adverse outcome **High RoB:** - Uncertain how exposure information obtained	x			

3. Can we be confident that the outcome of interest was not present at start of study?	**NA**		X		
4. Did the study match exposed and unexposed for all variables that are associated with the outcome of interest or did the statistical analysis adjust for these prognostic variables?	**Low RoB:** - Comprehensive matching or adjustment for all plausible prognostic variables **Higher RoB:** - Matching or adjustment for most plausible prognostic variables **HIGH RoB:** - Matching or adjustment for a minority of plausible prognostic variables - No matching or adjustment of plausible prognostic variables - Statements of no differences between groups - Statements that differences were not statistically significant are not sufficient for establishing comparability		x		
5. Can we be confident in the assessment of the presence or absence of prognostic factors?	**Low RoB:** - Interview of all participants - Self-completed survey from all participants - Review of charts with reproducibility demonstrated - From data base with documentation of accuracy of abstraction of prognostic data **Higher RoB:** - Chart review without demonstration of reproducibility - Data base with uncertain quality of abstraction of prognostic information **High RoB:** Prognostic information from data base with no available documentation of quality of abstraction of prognostic variables	X			

6. Can we be confident in the assessment of outcome?	**Low RoB:** - Independent blind assessment - Record linkage - For some outcomes (e.g. fractured hip), reference to the medical record is sufficient to satisfy the requirement for confirmation of the fracture **Higher RoB:** - Independent assessment unblinded - Self-report - For some outcomes (e.g. vertebral fracture where reference to x-rays would be required) reference to the medical record would not be adequate outcomes **High RoB:** - Uncertain (no description)		X		
7. Was the follow up of cohorts adequate?	**Low RoB:** - No missing outcome data - Reasons for missing outcome data unlikely to be related to true outcome (for survival data, censoring is unlikely to introduce bias) - Missing outcome data balanced in numbers across intervention groups, with similar reasons for missing data across groups - For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk is not enough to have an important impact on the intervention effect estimate - For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes is not large enough to have an important impact on the observed effect size - Missing data have been imputed using appropriated methods **High RoB:** - Reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across intervention groups - For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk is enough to induce important bias in intervention effect estimate - For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes is large enough to induce clinically relevant bias in the observed effect size	X			

8. Were co-interventions similar between groups?	**Low RoB:** - Most or all relevant co-interventions that might influence the outcome of interest are documented to be similar in the exposed and unexposed **High RoB:** - Few or no relevant co-interventions that might influence the outcome of interest are documented to be similar in the exposed and unexposed	X			

Sierra-Guzman *et al.*, (2018) [40]

Question	Example	Definitely Yes (Low RoB)	Probably Yes	Probably No	Definitely No (High RoB)
1. Was selection of exposed and non-exposed cohorts drawn from the same population?	**Low RoB:** - Exposed and unexposed drawn for same administrative data base of patients presenting at same points of care over the same time frame **High RoB:** - Exposed and unexposed presenting to different points of care over a different time frame		x		
2. Can we be confident in the assessment of exposure?	**Low RoB:** - Secure record (e.g. surgical records, pharmacy records) - Repeated interview or other ascertainment asking about current use/exposure **Higher RoB:** - Structured interview at a single point in time - Written self-report - Individuals who are asked retrospectively confirm their exposure status may be subject to recall bias - less likely to recall an exposure if they have not developed an adverse outcome, and more likely to recall an exposure (whether an exposure occurred or not) if they have developed an adverse outcome **High RoB:** - Uncertain how exposure information obtained		x		

3. Can we be confident that the outcome of interest was not present at start of study?	**NA**				x
4. Did the study match exposed and unexposed for all variables that are associated with the outcome of interest or did the statistical analysis adjust for these prognostic variables?	**Low RoB:** - Comprehensive matching or adjustment for all plausible prognostic variables **Higher RoB:** - Matching or adjustment for most plausible prognostic variables **HIGH RoB:** - Matching or adjustment for a minority of plausible prognostic variables - No matching or adjustment of plausible prognostic variables - Statements of no differences between groups - Statements that differences were not statistically significant are not sufficient for establishing comparability	x			
5. Can we be confident in the assessment of the presence or absence of prognostic factors?	**Low RoB:** - Interview of all participants - Self-completed survey from all participants - Review of charts with reproducibility demonstrated - From data base with documentation of accuracy of abstraction of prognostic data **Higher RoB:** - Chart review without demonstration of reproducibility - Data base with uncertain quality of abstraction of prognostic information **High RoB:** Prognostic information from data base with no available documentation of quality of abstraction of prognostic variables		x		

6. Can we be confident in the assessment of outcome?	**Low RoB:** - Independent blind assessment - Record linkage - For some outcomes (e.g. fractured hip), reference to the medical record is sufficient to satisfy the requirement for confirmation of the fracture **Higher RoB:** - Independent assessment unblinded - Self-report - For some outcomes (e.g. vertebral fracture where reference to x-rays would be required) reference to the medical record would not be adequate outcomes **High RoB:** - Uncertain (no description)	x			
7. Was the follow up of cohorts adequate?	**Low RoB:** - No missing outcome data - Reasons for missing outcome data unlikely to be related to true outcome (for survival data, censoring is unlikely to introduce bias) - Missing outcome data balanced in numbers across intervention groups, with similar reasons for missing data across groups - For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk is not enough to have an important impact on the intervention effect estimate - For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes is not large enough to have an important impact on the observed effect size - Missing data have been imputed using appropriated methods **High RoB:** - Reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across intervention groups - For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk is enough to induce important bias in intervention effect estimate - For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes is large enough to induce clinically relevant bias in the observed effect size	x			

8. Were co-interventions similar between groups?	**Low RoB:** - Most or all relevant co-interventions that might influence the outcome of interest are documented to be similar in the exposed and unexposed **High RoB:** - Few or no relevant co-interventions that might influence the outcome of interest are documented to be similar in the exposed and unexposed	x			

Someeh *et al.*, (2015) [41]

Question	Example	Definitely Yes (Low RoB)	Probably Yes	Probably No	Definitely No (High RoB)
1. Was selection of exposed and non-exposed cohorts drawn from the same population?	**Low RoB:** - Exposed and unexposed drawn for same administrative data base of patients presenting at same points of care over the same time frame **High RoB:** - Exposed and unexposed presenting to different points of care over a different time frame		x		
2. Can we be confident in the assessment of exposure?	**Low RoB:** - Secure record (e.g. surgical records, pharmacy records) - Repeated interview or other ascertainment asking about current use/exposure **Higher RoB:** - Structured interview at a single point in time - Written self-report - Individuals who are asked retrospectively confirm their exposure status may be subject to recall bias - less likely to recall an exposure if they have not developed an adverse outcome, and more likely to recall an exposure (whether an exposure occurred or not) if they have developed an adverse outcome **High RoB:** - Uncertain how exposure information obtained	x			

3. Can we be confident that the outcome of interest was not present at start of study?	**NA**		x		
4. Did the study match exposed and unexposed for all variables that are associated with the outcome of interest or did the statistical analysis adjust for these prognostic variables?	**Low RoB:** - Comprehensive matching or adjustment for all plausible prognostic variables **Higher RoB:** - Matching or adjustment for most plausible prognostic variables **HIGH RoB:** - Matching or adjustment for a minority of plausible prognostic variables - No matching or adjustment of plausible prognostic variables - Statements of no differences between groups - Statements that differences were not statistically significant are not sufficient for establishing comparability	x			
5. Can we be confident in the assessment of the presence or absence of prognostic factors?	**Low RoB:** - Interview of all participants - Self-completed survey from all participants - Review of charts with reproducibility demonstrated - From data base with documentation of accuracy of abstraction of prognostic data **Higher RoB:** - Chart review without demonstration of reproducibility - Data base with uncertain quality of abstraction of prognostic information **High RoB:** Prognostic information from data base with no available documentation of quality of abstraction of prognostic variables		x		

6. Can we be confident in the assessment of outcome?	**Low RoB:** - Independent blind assessment - Record linkage - For some outcomes (e.g. fractured hip), reference to the medical record is sufficient to satisfy the requirement for confirmation of the fracture **Higher RoB:** - Independent assessment unblinded - Self-report - For some outcomes (e.g. vertebral fracture where reference to x-rays would be required) reference to the medical record would not be adequate outcomes **High RoB:** - Uncertain (no description)		x		
7. Was the follow up of cohorts adequate?	**Low RoB:** - No missing outcome data - Reasons for missing outcome data unlikely to be related to true outcome (for survival data, censoring is unlikely to introduce bias) - Missing outcome data balanced in numbers across intervention groups, with similar reasons for missing data across groups - For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk is not enough to have an important impact on the intervention effect estimate - For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes is not large enough to have an important impact on the observed effect size - Missing data have been imputed using appropriated methods **High RoB:** - Reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across intervention groups - For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk is enough to induce important bias in intervention effect estimate - For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes is large enough to induce clinically relevant bias in the observed effect size	x			

8. Were co-interventions similar between groups?	**Low RoB:** - Most or all relevant co-interventions that might influence the outcome of interest are documented to be similar in the exposed and unexposed **High RoB:** - Few or no relevant co-interventions that might influence the outcome of interest are documented to be similar in the exposed and unexposed	x			

Toyooka *et al.*, (2017) [42]

Question	Example	Definitely Yes (Low RoB)	Probably Yes	Probably No	Definitely No (High RoB)
1. Was selection of exposed and non-exposed cohorts drawn from the same population?	**Low RoB:** - Exposed and unexposed drawn for same administrative data base of patients presenting at same points of care over the same time frame **High RoB:** - Exposed and unexposed presenting to different points of care over a different time frame		x		
2. Can we be confident in the assessment of exposure?	**Low RoB:** - Secure record (e.g. surgical records, pharmacy records) - Repeated interview or other ascertainment asking about current use/exposure **Higher RoB:** - Structured interview at a single point in time - Written self-report - Individuals who are asked retrospectively confirm their exposure status may be subject to recall bias - less likely to recall an exposure if they have not developed an adverse outcome, and more likely to recall an exposure (whether an exposure occurred or not) if they have developed an adverse outcome **High RoB:** - Uncertain how exposure information obtained	x			

3. Can we be confident that the outcome of interest was not present at start of study?	**NA**	x			
4. Did the study match exposed and unexposed for all variables that are associated with the outcome of interest or did the statistical analysis adjust for these prognostic variables?	**Low RoB:** - Comprehensive matching or adjustment for all plausible prognostic variables **Higher RoB:** - Matching or adjustment for most plausible prognostic variables **HIGH RoB:** - Matching or adjustment for a minority of plausible prognostic variables - No matching or adjustment of plausible prognostic variables - Statements of no differences between groups - Statements that differences were not statistically significant are not sufficient for establishing comparability				x
5. Can we be confident in the assessment of the presence or absence of prognostic factors?	**Low RoB:** - Interview of all participants - Self-completed survey from all participants - Review of charts with reproducibility demonstrated - From data base with documentation of accuracy of abstraction of prognostic data **Higher RoB:** - Chart review without demonstration of reproducibility - Data base with uncertain quality of abstraction of prognostic information **High RoB:** Prognostic information from data base with no available documentation of quality of abstraction of prognostic variables		x		

6. Can we be confident in the assessment of outcome?	**Low RoB:** - Independent blind assessment - Record linkage - For some outcomes (e.g. fractured hip), reference to the medical record is sufficient to satisfy the requirement for confirmation of the fracture **Higher RoB:** - Independent assessment unblinded - Self-report - For some outcomes (e.g. vertebral fracture where reference to x-rays would be required) reference to the medical record would not be adequate outcomes **High RoB:** - Uncertain (no description)		x		
7. Was the follow up of cohorts adequate?	**Low RoB:** - No missing outcome data - Reasons for missing outcome data unlikely to be related to true outcome (for survival data, censoring is unlikely to introduce bias) - Missing outcome data balanced in numbers across intervention groups, with similar reasons for missing data across groups - For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk is not enough to have an important impact on the intervention effect estimate - For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes is not large enough to have an important impact on the observed effect size - Missing data have been imputed using appropriated methods **High RoB:** - Reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across intervention groups - For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk is enough to induce important bias in intervention effect estimate - For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes is large enough to induce clinically relevant bias in the observed effect size	x			

8. Were co-interventions similar between groups?	**Low RoB:** - Most or all relevant co-interventions that might influence the outcome of interest are documented to be similar in the exposed and unexposed **High RoB:** - Few or no relevant co-interventions that might influence the outcome of interest are documented to be similar in the exposed and unexposed	x			

Toyooka *et al.*, (2018) [43]

Question	Example	Definitely Yes (Low RoB)	Probably Yes	Probably No	Definitely No (High RoB)
1. Was selection of exposed and non-exposed cohorts drawn from the same population?	**Low RoB:** - Exposed and unexposed drawn for same administrative data base of patients presenting at same points of care over the same time frame **High RoB:** - Exposed and unexposed presenting to different points of care over a different time frame	x			
2. Can we be confident in the assessment of exposure?	**Low RoB:** - Secure record (e.g. surgical records, pharmacy records) - Repeated interview or other ascertainment asking about current use/exposure **Higher RoB:** - Structured interview at a single point in time - Written self-report - Individuals who are asked retrospectively confirm their exposure status may be subject to recall bias - less likely to recall an exposure if they have not developed an adverse outcome, and more likely to recall an exposure (whether an exposure occurred or not) if they have developed an adverse outcome **High RoB:** - Uncertain how exposure information obtained		x		

3. Can we be confident that the outcome of interest was not present at start of study?	**NA**		x		
4. Did the study match exposed and unexposed for all variables that are associated with the outcome of interest or did the statistical analysis adjust for these prognostic variables?	**Low RoB:** - Comprehensive matching or adjustment for all plausible prognostic variables **Higher RoB:** - Matching or adjustment for most plausible prognostic variables **HIGH RoB:** - Matching or adjustment for a minority of plausible prognostic variables - No matching or adjustment of plausible prognostic variables - Statements of no differences between groups - Statements that differences were not statistically significant are not sufficient for establishing comparability	x			
5. Can we be confident in the assessment of the presence or absence of prognostic factors?	**Low RoB:** - Interview of all participants - Self-completed survey from all participants - Review of charts with reproducibility demonstrated - From data base with documentation of accuracy of abstraction of prognostic data **Higher RoB:** - Chart review without demonstration of reproducibility - Data base with uncertain quality of abstraction of prognostic information **High RoB:** Prognostic information from data base with no available documentation of quality of abstraction of prognostic variables		x		

6. Can we be confident in the assessment of outcome?	**Low RoB:** - Independent blind assessment - Record linkage - For some outcomes (e.g. fractured hip), reference to the medical record is sufficient to satisfy the requirement for confirmation of the fracture **Higher RoB:** - Independent assessment unblinded - Self-report - For some outcomes (e.g. vertebral fracture where reference to x-rays would be required) reference to the medical record would not be adequate outcomes **High RoB:** - Uncertain (no description)		x		
7. Was the follow up of cohorts adequate?	**Low RoB:** - No missing outcome data - Reasons for missing outcome data unlikely to be related to true outcome (for survival data, censoring is unlikely to introduce bias) - Missing outcome data balanced in numbers across intervention groups, with similar reasons for missing data across groups - For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk is not enough to have an important impact on the intervention effect estimate - For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes is not large enough to have an important impact on the observed effect size - Missing data have been imputed using appropriated methods **High RoB:** - Reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across intervention groups - For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk is enough to induce important bias in intervention effect estimate - For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes is large enough to induce clinically relevant bias in the observed effect size	x			

8. Were co-interventions similar between groups?	**Low RoB:** - Most or all relevant co-interventions that might influence the outcome of interest are documented to be similar in the exposed and unexposed **High RoB:** - Few or no relevant co-interventions that might influence the outcome of interest are documented to be similar in the exposed and unexposed		x		

Acknowledgements

There were no sources of funding or grants that supported the current systematic review or the investigators. The authors would like to acknowledge the guidance, consultation and support of Dr. Abebaw Yohannes, Ph.D. throughout the duration of this initiative.

Conflict of Interest

The authors declare no conflict of interest in this work.

Bibliography

1. Fong, D. T. P., Hong, Y., Chan, L. K., Yung, P. S. H. & Chan, K. M. (2007). A Systematic Review on Ankle Injury and Ankle Sprain in Sports. *Sports Medicine, 37*(1), 73-94.

2. Hootman, J. M., Dick, R. & Agel, J. (2007). Epidemiology of collegiate injuries for 15 sports: Summary and recommendations for injury prevention initiatives. *Journal of Athletic Training, 42*(2), 311-319.

3. Witchalls, J. B., Newman, P., Waddington, G., Adams, R. &Blanch, P. (2013). Functional performance deficits associated with ligamentous instability at the ankle. *Journal of Science and Medicine in Sport, 16*(2), 89-93.

4. Hegedus, E. J., McDonough, S. M., Bleakley, C. M., Baxter, G. D. & Cook, C. E. (2015a). Clinician-friendly lower extremity physical performance measures in athletes: a systematic review of measurement properties and correlation with injury, part 1. *British Journal of Sports Medicine, 49*(10), 642-648.

5. Hegedus, E. J., McDonough, S. M., Bleakley, C. M., Baxter, G. D. & Cook, C. E. (2015b). Clinician-friendly lower extremity physical performance tests in athletes: a systematic review of measurement properties and correlation with injury. Part 2-the tests for the hip, thigh, foot and ankle including the star excursion balance test. *British Journal of Sports Medicine, 49*(10), 649-656.

6. Harris, J. D., Abrams, G. D., Bach, B. R., Williams, D., Heidloff, D., Bush-Joseph, C. A., Verma, N. N., Forsythe, B. & Cole, B. J. (2014). Return to sport after ACL reconstruction. *Orthopedics, 37*(2), 103-108.

7. Davies, McCarty, E., Provencher, M. & Manske, R. C. (2017). ACL Return to Sport Guidelines and Criteria. *Current Reviews in Musculoskeletal Medicine, 10*(3), 307-314.

8. Webster, K. E. &Hewett, T. E. (2019). What is the Evidence for and Validity of Return-to-Sport Testing after Anterior Cruciate Ligament Reconstruction Surgery? *A Systematic Review and Meta-Analysis. Sports Medicine, 49*(6), 917-929.

9. Habets, B., van den Broek, A. G., Huisstede, B. M. A., Backx, F. J. G. & van Cingel, R. E. H. (2018). Return to Sport in Athletes with Midportion Achilles Tendinopathy: A Qualitative Systematic Review Regarding Definitions and Criteria. *Sports Medicine, 48*(3), 705-723.

10. Kivlan, B. R. & Martin, R. L. (2012). Functional Performance Testing of the Hip in Athletes. *International Journal of Sports Physical Therapy, 7*(4), 402-412.

11. Tassignon, B., Verschueren, J., Delahunt, E., Smith, M., Vicenzino, B., Verhagen, E. & Meeusen, R. (2019). Criteria-Based Return to Sport Decision-Making Following Lateral Ankle Sprain Injury: a Systematic Review and Narrative Synthesis. *Sports Medicine, 49*(4), 601-619.

12. Wikstrom, E. A., Mueller, C. &Cain, M. S. (2020). Lack of Consensus on Return-to-Sport Criteria Following Lateral Ankle Sprain: A Systematic Review of Expert Opinions. *Journal of Sport Rehabilitation, 29*(2), 231-237.

13. Clanton, T. O., Matheny, L. M., Jarvis, H. C. & Jeronimus, A. B. (2012). Return to Play in Athletes Following Ankle Injuries. *Sports Health: A Multidisciplinary Approach, 4*(6), 471-474.

14. Vogler, J. H., Csiernik, A. J., Yorgey, M. K., Harrison, J. J. & Games, K. E. (2017). Clinician-friendly physical performance tests for the hip, ankle, and foot. *Journal of Athletic Training, 52*(9), 861-862.

15. Fulton, J., Wright, K., Kelly, M., Zebrosky, B., Zanis, M., Drvol, C. & Butler, R. (2014). Injury risk is altered by previous injury: a systematic review of the literature and presentation of causative neuromuscular factors. *International Journal of Sports Physical Therapy, 9*(5), 583-595.

16. Attenborough, A. S., Hiller, C. E., Smith, R. M., Stuelcken, M., Greene, A. & Sinclair, P. J. (2014). Chronic Ankle Instability in Sporting Populations. *Sports Medicine, 44*(11), 1545-1556.

17. Malliaropoulos, N., Ntessalen, M., Papacostas, E., Giuseppe Longo, U. &Maffulli, N. (2009). Reinjury after Acute Lateral Ankle Sprains in Elite Track and Field Athletes. *The American Journal of Sports Medicine, 37*(9), 1755-1761.

18. Chalmers, I., Enkin, M. & Keirse, M. J. N. C. (1993). Preparing and Updating Systematic Reviews of Randomized Controlled Trials of Health Care. *The Milbank Quarterly, 71*(3), 411.

19. Garner, P., Hopewell, S., Chandler, J., MacLehose, H., Schünemann, H. J., Akl, E. A., Beyene, J., *et al.* (2016). When and how to update systematic reviews: Consensus and checklist. *BMJ (Online), 354,* 1-10.

20. Higgins, J. &Green, S. (2008). Chapter 22: Overview of reviews. Cochrane handbook for systematic reviews of interventions. *Cochrane Database of Systematic Reviews,* 187-235.

21. Tegner, Y. & Lysholm, J. (1985). Rating systems in the evaluation of knee ligament injuries. *Clinical Orthopaedics and Related Research, 198,* 43-49.

22. Alves, Y., Ribeiro, F. & Silva, A. G. (2018). Effect of fibular repositioning taping in adult basketball players with chronic ankle instability: A randomized, placebo-controlled, crossover trial. *Journal of Sports Medicine and Physical Fitness, 58*(10), 1465-1473.

23. Anguish, B.& Sandrey, M. A. (2018). Two 4-week balance-training programs for chronic ankle instability. *Journal of Athletic Training, 53*(7), 662-671.

24. Bagherian, S., Rahnama, N., Wikstrom, E. A., Clark, M. A., Rostami, F. & Donovan, L. (2018). Characterizing lower extremity movement scores before and after fatigue in collegiate athletes with chronic ankle instability. *International Journal of Athletic Therapy and Training, 23*(1), 27-32.

25. Bagherian, S., Rahnama, N. &Wikstrom, E. A. (2019). Corrective Exercises Improve Movement Efficiency and Sensorimotor Function but Not Fatigue Sensitivity in Chronic Ankle Instability Patients. *Clinical Journal of Sport Medicine, 29*(3), 193-202.

26. Best, R., Böhle, C., Schiffer, T., Petersen, W., Ellermann, A., Brueggemann, G. P. & Liebau, C. (2015). Early functional outcome of two different orthotic concepts in ankle sprains: a randomized controlled trial. *Archives of Orthopaedic and Trauma Surgery, 135*(7), 993-1001.

27. Cain, M. S., Garceau, S. W. & Linens, S. W. (2017). Effects of a 4-week biomechanical ankle platform system protocol on balance in high school athletes with chronic ankle instability. *Journal of Sport Rehabilitation, 26*(1), 1-7.

28. Cho, B. K., Hong, S. H. & Jeon, J. H. (2019). Effect of Lateral Ligament Augmentation Using Suture-Tape on Functional Ankle Instability. *Foot and Ankle International, 40*(4), 447-456.

29. Cruz-Diaz, D., Lomas-Vega, R., Osuna-Pérez, M. C., Contreras, F. H. & Martínez-Amat, A. (2014). Effects of 6 Weeks of Balance Training on Chronic Ankle Instability in Athletes: A Randomized Controlled Trial. *International Journal of Sports Medicine, 36*(9), 754-760.

30. Fereydounnia, S., Shadmehr, A., Attarbashi Moghadam, B., Talebian Moghadam, S., Mir, S. M., Salemi, S. & Pourkazemi, F. (2019). Improvements in strength and functional performance after Kinesio taping in semi-professional male soccer players with and without functional ankle instability. *Foot, 41,* 12-18.

31. Golditz, T., Welsch, G. H., Pachowsky, M., Hennig, F. F., Pfeifer, K. & Steib, S. (2016). A multimodal approach to ankle instability: Interrelations between subjective and objective assessments of ankle status in athletes. *Journal of Orthopaedic Research, 34*(3), 525-532.

32. Hall, E. A., Chomistek, A. K., Kingma, J. J. & Docherty, C. L. (2018a). Balance- and strength-training protocols to improve chronic ankle instability deficits, part I: Assessing clinical outcome measures. *Journal of Athletic Training, 53*(6), 568-577.

33. Harriss, J., Khan, A., Song, K., Register-Mihalik, J. K. & Wikstrom, E. A. (2019). Clinical movement assessments do not differ between collegiate athletes with and without chronic ankle instability. *Physical Therapy in Sport, 36*, 22-27.

34. Kamali, F., Sinaei, E. &Bahadorian, S. (2017). The immediate effect of talocrural joint manipulation on functional performance of 15-40 years old athletes with chronic ankle instability: A double-blind randomized clinical trial. *Journal of Bodywork and Movement Therapies, 21*(4), 830-834.

35. Ko, J., Rosen, A. B. & Brown, C. N. (2018). Functional performance deficits in adolescent athletes with a history of lateral ankle sprain(s). *Physical Therapy in Sport, 33,* 125-132.

36. Madsen, L. P., Hall, E. A. & Docherty, C. L. (2018). Assessing Outcomes in People With Chronic Ankle Instability: The Ability of Functional Performance Tests to Measure Deficits in Physical Function and Perceived Instability. *Journal of Orthopaedic & Sports Physical Therapy, 48*(5), 372-380.

37. Powden, C. J., Hoch, J. M., Jamali, B. E. & Hoch, M. C. (2019). A 4-week multimodal intervention for individuals with chronic ankle instability: Examination of disease-oriented and patient-oriented outcomes. *Journal of Athletic Training, 54*(4), 384-396.

38. Ryu, C. H., Park, J., Kang, M., Oh, J. H., Kim, Y. K., Kim, Y. Il, Lee, H. S. & Seo, S. G. (2019). Differences in lower quarter Y-balance test with player position and ankle injuries in professional baseball players. *Journal of Orthopaedic Surgery, 27*(1), 1-7.

39. Sierra-Guzman, R., Jimenez, F. & Abian-Vicen, J. (2018). Predictors of chronic ankle instability: Analysis of peroneal reaction time, dynamic balance and isokinetic strength. *Clinical Biomechanics, 54,* 28-33.

40. Sierra-Guzman, R., Jimenez-Diaz, F., Ramirez, C., Esteban, P. & Abian-Vicen, J. (2018). Whole-body-vibration training and balance in recreational athletes with chronic ankle instability. *Journal of Athletic Training, 53*(4), 355-363.

41. Someeh, M., Norasteh, A. A., Daneshmandi, H. & Asadi, A. (2015). Immediate effects of Mulligan's fibular repositioning taping on postural control in athletes with and without chronic ankle instability. *Physical Therapy in Sport, 16*(2), 135-139.

42. Toyooka, S., Takeda, H., Nakajima, K., Masujima, A., Miyamoto, W., Pagliazzi, G., Nakagawa, T. & Kawano, H. (2017). Correlation Between Recovery of Triceps Surae Muscle Strength and Level of Activity After Open Repair of Acute Achilles Tendon Rupture. *Foot & Ankle International, 38*(12), 1324-1330.

43. Toyooka, T., Urabe, Y., Sugiura, S., Takata, A., Shinozaki, M., Takata, Y., Ishizaki, T., *et al.* (2018). Does the single-limb stance reflect chronic ankle instability in an athlete? *Gait and Posture, 66,* 242-246.

44. Brumitt, J., Heiderscheit, B. C., Manske, R. C., Niemuth, P. E. & Rauh, M. J. (2013). Lower extremity functional tests and risk of injury in division iii collegiate athletes. *International Journal of Sports Physical Therapy, 8*(3), 216-227.

45. Davies, George J. &Riemann, B. (2013). *The Use of a Functional Testing Algorithm (FTA) to Make Qualitative and Quantitative Decisions to Return Athletes Back to Sports Following Shoulder Injuries.* (Pp. 1-13).

46. Eechaute, C., Leemans, L., De Mesmaeker, M., De Ridder, R., Beckwée, D., Struyf, F., Roosen, F., *et al.* (2020). The predictive value of the multiple hop test for first-time noncontact lateral ankle sprains. *Journal of Sports Sciences, 38*(1), 86-93.

47. Guyatt, G. & Busse, J. (2015a). Methods commentary: Risk of bias in randomized trials 1.

48. Guyatt, G. & Busse, J. (2015b). No TitleMethods commentary: Risk of cohort studies.

49. Bartels, B., de Groot, J. F. & Terwee, C. B. (2013). The Six-Minute Walk Test in Chronic Pediatric Conditions: A Systematic Review of Measurement Properties. *Physical Therapy, 93*(4), 529-541.

50. Caffrey, E., Docherty, C. L., Schrader, J. & Klossner, J. (2009). The ability of 4 single-limb hopping tests to detect functional performance deficits in individuals with functional ankle instability. *Journal of Orthopaedic and Sports Physical Therapy, 39*(11), 799-806.

51. Armijo-Olivo, S., Fuentes, J., Da Costa, B. R., Saltaji, H., Ha, C. & Cummings, G. G. (2017). Blinding in Physical Therapy Trials and Its Association with Treatment Effects: A Meta-epidemiological Study. *American Journal of Physical Medicine and Rehabilitation, 96*(1), 34-44.

52. Grimes, D. A. & Schulz, K. F. (2002). Cohort studies: Marching towards outcomes. *Lancet, 359*(9303), 341-345.

53. Coetzee, J. C., Ellington, J. K., Ronan, J. A. & Stone, R. M. (2018). Functional Results of Open Broström Ankle Ligament Repair Augmented With a Suture Tape. *Foot and Ankle International, 39*(3), 304-310.

54. Kinzey, S. J. &Armstrong, C. W. (1998). The reliability of the star-excursion test in assessing dynamic balance. *Journal of Orthopaedic and Sports Physical Therapy, 27*(5), 356-360.

55. Plisky, P. J., Rauh, M. J., Kaminski, T. W. & Underwood, F. B. (2006). Star excursion balance test as a predictor of lower extremity injury in high school basketball players. *Journal of Orthopaedic and Sports Physical Therapy, 36*(12), 911-919.

56. Hertel, J., Miller, S. J. & Denegar, C. R. (2000). Intratester and Intertester Reliability during the Star Excursion Balance Tests. *Journal of Sport Rehabilitation, 9*(2), 104-116.

57. Hertel, J., Braham, R. A., Hale, S. A. & Olmsted-Kramer, L. C. (2006). Simplifying the star excursion balance test: Analyses of subjects with and without chronic ankle instability. *Journal of Orthopaedic and Sports Physical Therapy, 36*(3), 131-137.

58. Olmsted-Kramer, L. C., Carcia, C. R., Hertel, J. & Shultz, S. J. (2002). Efficacy of the Star Excursion Balance Tests in Detecting Reach Deficits in Subjects With Chronic Ankle Instability. *Journal of Athletic Training, 37*(4), 501-506.

59. Shaffer, S. W., Teyhen, D. S., Lorenson, C. L., Warren, R. L., Koreerat, C. M., Straseske, C. A. &Childs, J. D. (2013). Y-Balance Test: A Reliability Study Involving Multiple Raters. *Military Medicine, 178*(11), 1264-1270.

60. McCann, R., Kosik, K., Terada, M. &Gribble, P. (2018). Residual impairments and activity limitations at return to play from a lateral ankle sprain. *International Journal of Athletic Therapy and Training, 23*(2), 83-88.

61. Hoch, M. C., Staton, G. S. & McKeon, P. O. (2011). Dorsiflexion range of motion significantly influences dynamic balance. *Journal of Science and Medicine in Sport, 14*(1), 90-92.

62. Ostenberg, A. & Roos, H. (2000). Injury risk factors in female European football. A prospective study of 123 players during one season. *Scandinavian Journal of Medicine and Science in Sports, 10*(5), 279-285.

63. Noyes, F. R., Barber-Westin, S. D., Smith, S. T. & Campbell, T. (2011). Atraining program to improve neuromuscular indices in female high school volleyball players. *Journal of Strength and Conditioning Research, 25*(8), 2151-2160.

64. Hjelm, N., Werner, S. & Renstrom, P. (2012). Injury risk factors in junior tennis players: a prospective 2-year study. *Scandinavian Journal of Medicien & Science in Sports, 22*(1), 40-48.

65. Cross, K. M., Wilson, R. W. & Perrin, D. H. (1996). Functional performance following an ice immersion to the lower extremity. *Journal of Athletic Training, 31*(2), 113-116.

66. Pienaar, C. &Coetzee, B. (2013). Changes in selected physical, motor performance and anthropometric components of university-level rugby players after one microcycle of a combined rugby conditioning and plyometric training program. *Journal of Strength and Conditioning Research, 27*(2), 398-415.

67. Váczi, M., Tollár, J., Meszler, B., Juhász, I. & Karsai, I. (2013). Short-term high intensity plyometric training program improves strength, power and agility in male soccer players. *Journal of Human Kinetics, 36*(1), 17-26.

68. Sekir, U, Yildiz, Y., Hazneci, B., Ors, F., Saka, T. & Aydin, T. (2008). Reliability of a functional test battery evaluating functionality, proprioception, and strength in recreational athletes with functional ankle instability. *European Journal of Physical and Rehabilitation Medici., 44*(4), 407-415.

69. Sekir, Ufuk, Yildiz, Y., Hazneci, B., Ors, F. & Aydin, T. (2007). Effect of isokinetic training on strength, functionality and proprioception in athletes with functional ankle instability. *Knee Surgery, Sports Traumatology, Arthroscopy, 15*(5), 654-664.

70. Kivlan, B. R., Clemente, F. R. & Phelps, A. L. (2013). Functional hip tests for dancers, *8*(4), 360-369.

71. Putnam, A. R., Bandolin, S. N. & Krabak, B. J. (2012). Impact of Ankle Bracing on Skill Performance in Recreational Soccer Players. *PM&R, 4*(8), 574-579.

72. Ko, J., Rosen, A. B. & Brown, C. N. (2017). Comparison Between Single and Combined Clinical Postural Stability Tests in Individuals With and Without Chronic Ankle Instability. *Clinical Journal of Sport Medicine, 27*(4), 394-399.

73. Arden, C. L., Glasgow, P., Schneiders, A., Witvrouw, E., Clarsen, B., Cools, A., Gojanovic, B., *et al.* (2016). 2016 Consensus statement on return to sport from the First World Congress in Sports Physical Therapy, Bern. *Br J Sports Med., 50*(14), 853-864.

74. Creighton, D. W., Shrier, I., Shultz, R., Meeuwisse, W. H. & Matheson, G. O. (2010). Return-to-Play in Sport: A Decision-based Model. *Clinical Journal of Sport Medicine, 20*(5), 379-385.

75. Chang, W. D., Chou, L. W., Chang, N. J. & Chen, S. (2020). Comparison of Functional Movement Screen, Star Excursion Balance Test, and Physical Fitness in Junior Athletes with Different Sports Injury Risk. *BioMed Research International, 2020,* 1-8.

76. Eagle, S. R., Kessels, M., Johnson, C. D., Nijst, B., Lovalekar, M., Krajewski, K., Flanagan, S. D., *et al.* (2019). Bilateral strength asymmetries and unilateral strength imbalance: Predicting ankle injury when considered with higher body mass in US special forces. *Journal of Athletic Training, 54*(5), 497-504.

77. Fransz, D. P., Huurnink, A., Kingma, I., de Boode, V. A., Heyligers, I. C. &van Dieën, J. H. (2018). Performance on a Single-Legged Drop-Jump Landing Test Is Related to Increased Risk of Lateral Ankle Sprains Among Male Elite Soccer Players: A 3-Year Prospective Cohort Study. *The American Journal of Sports Medicine, 46*(14), 3454-3462.

78. Powers, C. M., Ghoddosi, N., Straub, R. K. & Khayambashi, K. (2017). Hip Strength as a Predictor of Ankle Sprains in Male Soccer Players: A Prospective Study. *Journal of Athletic Training, 52*(11), 1048-1055.

79. Ellenbecker, T. S. & Davies, G. J. (2001). *Closed kinetic chain exercise: a comprehensive guide to multiple joint exercise*. Human Kinetics.

80. Etnoyer, J., Greenstein, J. &Bishop, B. (2013). FUNHAB®: A Science-based, Multimodal Approach for Musculoskeletal Conditions. *Topics of Integrated Health, 4*(2).

81. Giangarra, C. E. &Manske, R. C. (2017). *Clinical Orthopaedic Rehabilitation: A Team Approach* E-book. Elsevier Health Sciences.

82. Page, P. & Frank, C. (2007). The janda approach to chronic musculoskeletal pain.

83. Page, P., Frank, C. C. & Lardner, R. (2010). *Assessment and treatment of muscle imbalance: the Janda approach*. Human kinetics.

84. Axe, M. J. & Snyder-Mackler, L. (2005). *Operative and post-operative management of the knee.* Orthopaedic Section Independent Study Course, 15.

85. Delahunt, E., Bleakley, C. M., Bossard, D. S., Caulfield, B. M., Docherty, C. L., Doherty, C., Fourchet, F., *et al*. (2018). Clinical assessment of acute lateral ankle sprain injuries (ROAST): 2019 consensus statement and recommendations of the International Ankle Consortium. *British Journal of Sports Medicine, 52*(20), 1304-1310.

86. Shultz, S., Olszewski, A., Ramsey, O., Schmitz, M., Wyatt, V. & Cook, C. (2013). A systematic review of outcome tools used to measure lower leg conditions. *International Journal of Sports Physical Therapy, 8*(6), 838-848.

87. Martin, R. L. & Irrgang, J. J. (2007). A survey of self-reported outcome instruments for the foot and ankle. *Journal of Orthopaedic and Sports Physical Therapy, 37*(2), 72-84.

88. Cook, G. (2010). *Movement: Functional movement systems: Screening, assessment, corrective strategies. On Target Publications*. Clinical Journal of Sport Medicine.

89. Ozinga, S. J., Linder, S. M., Koop, M. M., Dey, T., Figler, R., Russman, A. N., So, R., Rosenthal, A. H., Cruickshank, J. & Alberts, J. L. (2018). Normative performance on the balance error scoring system by youth, high school, and collegiate athletes. *Journal of Athletic Training, 53*(7), 636-645.

90. Springer, B. A., Marin, R., Cyhan, T., Roberts, H. &Gill, N. W. (2007). Normative values for the unipedal stance test with eyes open and closed. *Journal of Geriatric Physical Therapy, 30*(1), 8-15.

91. Functional Movement Systems. (2015). Y Balance Test (YBT) Online Manual: Version 1.

92. Monahan, A. C. (2018). *Psychological Readiness of Athletes to Return to Play Following Injury.*Electronic Theses and Dissertations, (Pp. 1-65).

93. Greisberg, J., Gould, P., Vosseller, J. T., Greisberg, M., Bandasak, N., Dolar, R. & Ahmad, C. (2019). Performance Function Tests in Assessing Ankle Fitness. *JAAOS: Global Research and Reviews, 3*(1), e096.

94. Gokeler, A., Welling, W., Benjaminse, A., Lemmink, K., Seil, R. & Zaffagnini, S. (2017). A critical analysis of limb symmetry indices of hop tests in athletes after anterior cruciate ligament reconstruction: A case control study. *Orthopaedics and Traumatology: Surgery and Research, 103*(6), 947-951.

95. Logerstedt, D., Grindem, H., Lynch, A., Eitzen, I., Engebretsen, L., Risberg, M. A., Axe, M. J. & Snyder-Mackler, L. (2012). Single-Legged Hop Tests as Predictors of Self-Reported Knee Function After Anterior Cruciate Ligament Reconstruction. *The American Journal of Sports Medicine, 40*(10), 2348-2356.

96. Greenberg, E. M., Dyke, J., Leung, A., Karl, M., Lawrence, J. T. & Ganley, T. (2020a). Uninjured Youth Athlete Performance on Single-Leg Hop Testing: How Many Can Achieve Recommended Return-to-Sport Criterion? *Sports Health, 12*(6), 552-558.

97. Davies, G. J. & Zillmer, D. A. (2000). Functional progression of a patient through a rehabilitation program. *Orthopaedic Physical Therapy Clinics of North America, 9*(2), 103-118.

98. Yalfani, A., Gandomi, F. & Kohboomi, M. (2017). The effect of G-max and G-med muscles fatigue on functional performance and balance in athletes with and without chronic ankle instability. *Asian Journal of Sports Medicine, 8*(3).

99. Madsen, L. P., Booth, R. L., Volz, J. D. & Docherty, C. L. (2020). Using Normative Data and Unilateral Hopping Tests to Reduce Ambiguity in Return-to-Play Decisions. *Journal of Athletic Training, 55*(7).

100. Onate, J. A., Starkel, C., Clifton, D. R., Best, T. M., Borchers, J., Chaudhari, A., Dawn Comstock, R., *et al.* (2018). Normative functional performance values in high school athletes: The functional pre-participation evaluation project. *Journal of Athletic Training, 53*(1), 35-42.

101. Greenberg, E. M., Karl, M., Leung, A., Lawrence, J. T. & Ganley, T. (2020b). Limb Symmetry Is Not Enough: Establishment of Height Normalized Hop Distances Within Healthy Youth Athletes. *Orthopaedic Journal of Sports Medicine, 8*(4_suppl3).

102. Brumitt, J., Heiderscheit, B. C., Manske, R. C., Niemuth, P. E., Mattocks, A. & Rauh, M. J. (2018). Preseason functional test scores are associated with future sports injury in female collegiate athletes. *Journal of Strength and Conditioning Research, 32*(6), 1692-1701.

103. Haitz, K., Shultz, R., Hodgins, M. & Matheson, G. O. (2014). Test-retest and interrater reliability of the functional lower extremity evaluation. *Journal of Orthopaedic and Sports Physical Therapy, 44*(12), 947-954.

104. Hardesty, K., Hegedus, E. J., Ford, K. R., Nguyen, A.-D. & Taylor, J. B. (2017). Determination of Clinically Relevant Differences in Frontal Plane Hop Tests in Women'S Collegiate Basketball and Soccer Players. *International Journal of Sports Physical Therapy, 12*(2), 182-189.

105. Higgins, Thomas, J., Cumpston, M., Li, T., Page, M. J. & Welch, V. A. (2019). *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons.

106. Sharma, N., Sharma, A. & Sandhu, J. S. (2011). Functional performance testing in athletes with functional ankle instability. *Asian Journal of Sports Medicine, 2*(4), 249-258.