Title	Reliability, Validity and Utility of Inertial Sensor Systems for Postural Control Assessment in Sport Science and Medicine Applications: A Systematic Review
Authors(s)	Johnston, William, Argent, Rob, O'Reilly, Martin, Caulfield, Brian
Publication date	2019-03-22
Publication information	Johnston, William, Rob Argent, Martin O'Reilly, and Brian Caulfield. "Reliability, Validity and Utility of Inertial Sensor Systems for Postural Control Assessment in Sport Science and Medicine Applications: A Systematic Review" 49, no. 5 (March 22, 2019).
Publisher	Springer
Item record/more information	http://hdl.handle.net/10197/10761
Publisher's statement	This is a post-peer-review, pre-copyedit version of an article published in Sports Medicine. The final authenticated version is available online at: http://dx.doi.org/10.1007/s40279-019-01095-9
Publisher's version (DOI)	10.1007/s40279-019-01095-9

Reliability, validity and utility of inertial sensor systems for postural control assessment in sport science and medicine applications: A systematic review

Short Title: Inertial sensor systems for postural control assessment in sports science and medicine.

William Johnston[1,2], Martin O'Reilly[1,2], Rob Argent[1,2] and Brian Caulfield[1,2]

1. Insight Centre for Data Analytics, University College Dublin, Ireland.
2. School of Public Health, Physiotherapy and Sports Science, University College Dublin, Ireland.

ORCID ID

William Johnston: https://orcid.org/0000-0003-0525-6577

Martin O'Reilly: https://orcid.org/0000-0003-2425-5393

Rob Argent: https://orcid.org/0000-0002-3522-9409

Brian Caulfield: https://orcid.org/0000-0003-0290-9587

Corresponding Author: William Johnston – William.Johnston@ucdconnect.ie

Compliance with Ethical Standards

Funding

William Johnston and Brian Caulfield were funded by Science Foundation Ireland under their grant for the Insight Centre for Data Analytics (SFI/12/RC/2289). Martin O'Reilly receives funding from Enterprise Ireland (CF/2017/0739). Rob Argent receives funding under the Marie Sklodowska-Curie grant agreement (no. 676201). These funding bodies had no influence on the data collection, data analysis, data interpretation or approval/disapproval of publication.

Conflicts of interest

William Johnston, Martin O'Reilly, Rob Argent and Brian Caulfield declare that they have no conflicts of interested relevant to the content of this review.

Abstract

Background: Recent advances in mobile sensing and computing technology has provided a means to objectively quantify postural control in unobtrusive environment. This has resulted in the rapid development and evaluation of a series of wearable inertial sensor-based assessments. However, the validity, reliability and clinical utility of such systems is not fully understood.

Objectives: This systematic review aims to synthesise and evaluate studies which have investigated the ability of wearable inertial sensor systems to validly and reliability quantify postural control performance in sports science and medicine applications.

Methods: A systematic search strategy, utilising the PRISMA guidelines, was employed to identify eligible articles through ScienceDirect, Embase and PubMed databases. Forty-seven articles met the inclusion criteria and were evaluated and qualitatively synthesised under two main headings: measurement validity and measurement reliability. Furthermore, studies which investigated the utility of these systems in clinical populations were summarised and discussed.

Results: After duplicate removal 4,374 articles were identified with the search strategy, with 47 papers included in the final review. Twenty-eight studies investigated validity in healthy populations, while 15 studies investigated validity in clinical populations. Thirteen studies investigated the measurement reliability of these sensor-based systems.

Conclusions: The application of wearable inertial sensors for sports science and medicine postural control applications is an evolving field. To date, research has primarily focused on evaluating the validity and reliability of a heterogeneous set of assessment protocols, in a laboratory environment. While researchers have begun to investigate their utility in clinical use-cases such as concussion and musculoskeletal injury, most studies have leveraged small sample sizes, have low study quality, and use a variety of descriptive variables, assessment protocols and sensor mounting locations. Future research should evaluate the clinical utility of these systems in large high-quality prospective cohort studies, to establish the role they may play in injury risk-identification, diagnosis and management. This systematic review was registered with the International Prospective Register of Systematic Reviews on the 10/08/2018 (PROSPERO registration: CRD42018106363): https://www.crd.york.ac.uk/PROSPERO/display_record.php?RecordID=106363.

Key Points

- Wearable inertial sensor systems can provide a valid and reliable measure of postural control performance which may be more sensitive to change than traditional clinical sports medicine postural control assessments.
- While sensor-based systems demonstrate promise, there is currently insufficient evidence to indicate the use of many of these assessments in sports medicine clinical practice.
- Further high-quality prospective research is required to establish the role these systems may play in sports injury-risk identification, diagnosis and management.
- When developing such systems, researchers should establish the validity (convergence; construct; discriminant) and the test-retest reliability (intra-session; inter-session) of the system, using consistent and well-defined methodologies.

1 INTRODUCTION

Postural control (PC) assessments are frequently used in sports science and medicine as outcome measures in performance testing, injury-risk screening, injury rehabilitation, and assessment of readiness to return to play [1-3]. PC can be broadly divided into two main components: dynamic PC and static PC. Static PC is defined as the ability to maintain the centre of gravity (COG) within the base of support, while dynamic PC involves the maintenance of an upright posture, while the COG moves outside of the base of support [4, 5]. In sports medicine, static PC assessments are frequently used in the acute-subacute evaluation of injuries, while dynamic assessments are typically introduced as the athlete progresses towards recovery, as a means to challenge the sensorimotor subsystems.

Despite the widespread use of sports science and medicine PC assessments, many rely on subjective ratings of performance, have significant floor and ceiling effects, and lack sensitivity and specificity [2], resulting in the need for more objective alternatives [6, 7]. The technological revolution of the 21st century has seen the development of ubiquitous wearable inertial sensor and mobile computing technology. Inertial sensors, coupled with mobile phones and tablets, provide a cheap ($50 - $2000) and accessible means to efficiently collect and process large amounts of human movement data, in an unconstrained environment [8]. The tri-axial accelerometer, gyroscope and/or magnetometer signals can then be used to describe aspects of 'stability' and 'strategy' during the PC assessment, enhancing the capabilities of researchers and clinicians' to objectively quantifying human movement, outside of the confines of the laboratory environment [8, 9]. For example, alterations in 'stability' post knee-injury may capture general sensorimotor deficits, while changes in knee frontal plane 'strategy' may indicate localised strength or control issues. While in their infancy in sports science and medicine, inertial sensor systems for PC assessment have demonstrated their utility in older adult populations, identifying those at an increased risk of falls [10] and capturing functional alterations in neurological populations, such as Parkinson's disease [11]. As their value in other clinical populations is being realised, the utility of such technology in sports science and medicine domains such as concussion, is starting to be explored [12-17]. However, despite promising early results, it is imperative that researchers and clinicians acknowledge the evolving nature of this area of research, and thoroughly evaluate the validity and reliability of such systems to ensure appropriate implementation.

In recent times, several reviews have been completed investigating the utility of inertial sensor systems in the quantification of sensorimotor function. Hubble and colleagues [18] completed a systematic review of wearable sensor-based assessments; however this work specifically focused on applications in a Parkinsonian population. More recently, Gordt and colleagues [19] completed a review investigating the utility of wearable sensor-based training systems for improving balance, gait and functional performance. Importantly, this review focused on randomised control trials investigating the utility of such training systems and did not focus on sensor-based PC outcome measures. To date, no review has investigated the ability of inertial sensors to quantify sports science and medicine specific PC assessments. As such, the objective of this systematic review is to synthesise and evaluate studies which investigate the ability of inertial sensor systems to validly and reliably quantify PC performance, in sports science and medicine applications. This review will address three main aims: (1) provide a description of the wearable inertial sensor systems investigated to date; describing sensor type, mounting location, PC variable type and the assessment protocols investigated; (2) establish the reliability and validity of instrumented PC assessments; (3) investigate the use of sensor-based PC assessments in clinical sports medicine populations.

2 METHODS

This systematic review was registered with the International Prospective Register of Systematic Reviews on the 10/08/2018 (PROSPERO registration: CRD42018106363). The search strategy and study protocol are available at: https://www.crd.york.ac.uk/PROSPERO/display_record.php?RecordID=106363. This study was designed and conducted in accordance with the recommendations of the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) statement guidelines [20].

2.1 LITERATURE SEARCH STRATEGY AND STUDY SELECTION PROCESS

A literature search was conducted within the following three databases: PubMed, ScienceDirect and Embase. Papers focusing on the following were aggregated: PC, assessment and wearable inertial sensors. In compliance with Lipscomb and colleagues [21], Medical Subject Headings (MeSH) terms, title/abstract keywords, synonyms and spelling variations were used in several combinations, and modified depending on the database being searched. Articles published from inception to July 2018 were reviewed. Table 1 details the search strategy implemented in this systematic review. The search included peer-reviewed articles published in scientific journals and conference proceedings and only considered articles published in the English language. The conference website of any article obtained from a conference proceeding was checked to confirm a peer-review process. This review focused on the development, validation and reliability testing of inertial sensor-based PC assessments for sports science and medicine applications, and thus did not consider articles which focused solely on middle aged (aged 45-64 years) and older adult (aged 65+ years) populations, as defined by the World Health Organisation [22]. Due to the broad nature of PC and the resultant heterogeneity of the literature, this study considered discrete instrumented PC assessments where PC was the primary outcome of interest; thus, gait-based assessments were not considered for inclusion, as they were beyond the scope of this review. In cases where papers included middle-older adults and young populations, only the analysis focusing on the young population (aged <45 years) were included. Due to the relatively new nature of inertial sensor-based balance assessments in sports medicine, the grey literature was not searched, with only peer-reviewed scientific articles included.

Table 1 here

Fig. 1 outlines the article selection process, completed in line with the PRISMA guidelines [20]. The search process was conducted as follows; (1) a search of the aforementioned databases was completed, from inception to July 2018; (2) duplicates were removed; (3) titles and abstracts were screened based on the inclusion/ exclusion criteria (table 2); (4) remaining full texts were read, and final articles were selected based on the inclusion/ exclusion criteria.

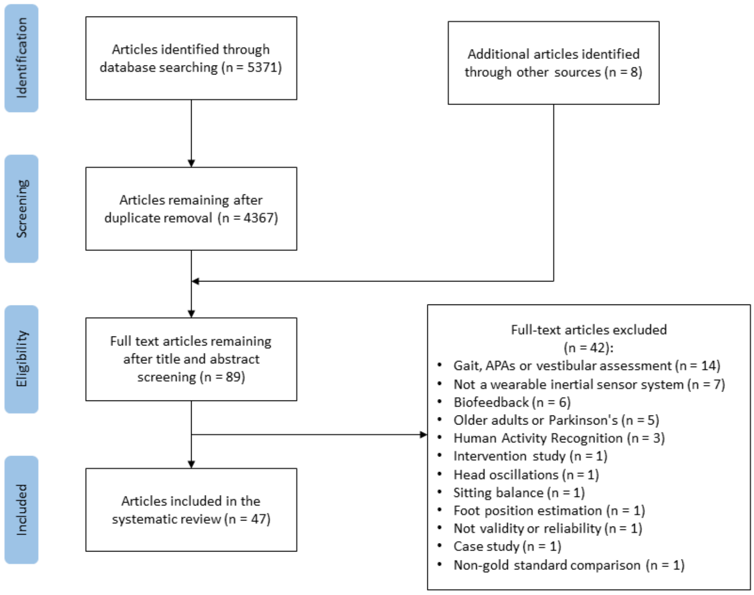

Fig 1: PRISMA study flow diagram

Table 2 here

2.2 DATA EXTRACTION

Data were extracted using a standardised template which allowed for the aggregation of the study design (validity and/or reliability), year of publication, participant demographic information, PC assessment type, sensor type, sensor parameters, sensor mounting location, number of sensors, sensor derived variable(s), results and conclusions. The studies were divided into two main categories: validity studies and reliability studies. Where articles focused on both reliability and validity, the appropriate data from the study in question was included in the relevant category. The included studies were then qualitatively summarised based on their aims, results and conclusions. Data extraction was completed separately by two authors (WJ and MOR) and cross-referenced for any discrepancies. Any inconsistencies were resolved by arbitration between WJ, MOR and a third author RA.

2.3 DATA ANALYSIS

A narrative synthesis of the data was completed by reporting the findings related to the sensor setup, type of assessments, type of sensor derived variable and validity and/or reliability statistics. The number of articles which investigated the various types of assessment, and if they focused on validity or reliability was reported. For the purpose of this study, validity was broadly broken down into four categories; concurrent validity, convergence validity, discriminant validity and predictive validity. Concurrent validity is defined as when a newly developed tool, such as an inertial sensor PC assessment, is compared to the 'gold standard'. Convergence validity is defined as when a newly developed tool, such as an inertial sensor PC assessment, is compared to a 'clinical standard' assessment. Discriminant validity can be defined as the ability of an assessment tool, such as an inertial sensor PC assessment, to differentiate groups or conditions that are known to be different. Predictive validity is defined as the ability of an assessment tool to predict a criterion measure. The reliability studies were broken down into two groups; intra-session and inter-session test-retest reliability. Intra-session reliability investigates the immediate test-retest reliability, related to the random variability of the measurement. Inter-session reliability is the reliability of the measure over a pre-defined time period (days, weeks, months), and accounts for the inherent natural variation and measurement error captured by the tool [23]. Due to the heterogeneous nature of the included studies, no quantitative meta-analysis could be completed for either the validity or reliability component of this review.

2.4 Assessment of study quality

The methodological quality of the included articles was assessed using a modified Downs and Black assessment scale [24]. The traditional Downs and Black scale consists of 27 questions, focused on reviewing the quality and risk of bias of randomised and non-randomised control trials. A modified version of the checklist is commonly used in the sports medicine field as a means to evaluate the quality of systematic reviews of observational cohort and case-control studies [21-23]. This modified scale appropriately utilises only the questions which are relevant to observational cohort studies and excludes the components specific to intervention studies, resulting in the use of 10 criteria (numbers 1-3, 6, 7, 10-12, 16, 18). Question 10 was modified to specify the inclusion of not only the p value probabilities, but also the reliability statistic scores, where appropriate. Two authors, (WJ and MOR) independently completed the quality assessment and data extraction of all eligible studies. In cases where WJ or MOR were authors of a paper included in this review, a third author (RA) assessed the quality to reduce the risk of bias. The quality ratings given by each author were cross-referenced, and any discrepancies were resolved by arbitration between WJ, MOR and RA. Once the reviewing authors reached a consensus, quality rating was summed, and each article was provided with a score out of 10.

3 Results

3.1 Database Search

The search strategy for this study identified 5371 potential articles for inclusion. After removal of 1004 duplicates, the total number of articles for title and abstract screening was 4367. Following title and abstract screening, an additional eight articles which were obtained through a reference list search, were screened and included, leaving 89 articles for full evaluation. Full review of the remaining articles resulted in the removal of 42 articles, resulting in a total of 47 articles eligible for the final dataset. Fig. 1 illustrates the various stages of the article screening process.

3.2 Paper Description and Quality

All 47 articles included in this review were observational studies, consisting of 32 cohort studies, 13 case-control studies and two prospective longitudinal studies. Broadly, 13 studies investigated the reliability of sensor-based measures of PC, while 43 studies focused on validity. Nine studies explored both the reliability and validity of inertial sensor-based measures of PC. All 32 cohort studies focused on healthy participants. Of the 11 case-control studies, eight investigated concussion, two focused on knee injuries, two investigated chronic ankle instability (CAI) and one examined lower back pain. A single prospective cohort study investigated the effect of concussion on PC performance in collegiate athletes, while the remaining prospective study evaluated the association between PC and general injury risk in collegiate athletes. Fig 2 illustrates the study design and number of studies investigating each clinical application.

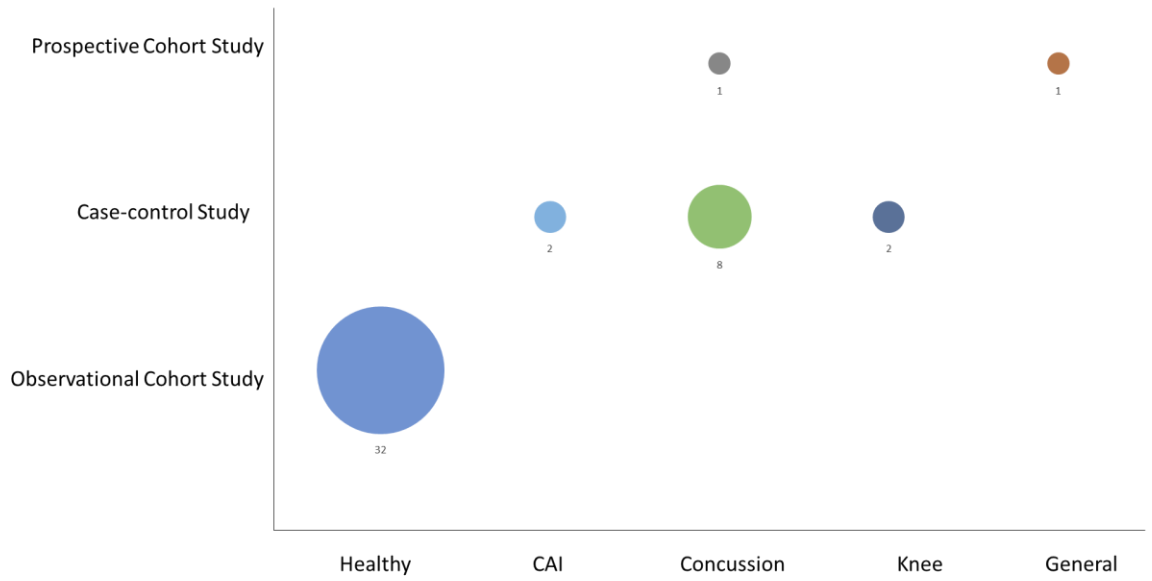

Fig 2: illustrates the design and number of included studies across the various populations. The study population is presented on the x-axis and the study type is presented on the y-axis. CAI – Chronic ankle instability

The quality of the included articles is detailed in appendix 1. Based on the quality ratings, 46/47 studies clearly described if the results were based on 'data dredging', 44/46 studies used appropriate statistical tests, while 40/47 studies provided estimates of random variability in the data for the main outcomes and presented the characteristics of the cohorts included. Importantly, no studies met the criteria for external validity (questions 11 and 12). Only 10/47 studies provided full details about the sensor calibration parameters and methods for deriving the inertial sensor-based PC variables (outcomes to be measured), while 22/47 provided an adequate description of the hypothesis and objective of the study.

3.3 Sensor Setup & Variables

Table 3 provides an overview of the sensor type, number, mounting location and variables used to describe PC performance. The most common sensor type used in the included articles was a tri-axial accelerometer (6 reliability, 17 validity), followed by tri-axial accelerometer/ gyroscope (4 reliability, 9 validity). Eight validity studies leveraged full inertial measurement units (IMU), a composite sensor consisting of a tri-axial accelerometer, tri-axial gyroscope and tri-axial magnetometer. A further five studies used bi-axial accelerometers, two using tri-axial gyroscopes and a single study using a tri-axial accelerometer/ magnetometer.

Table 3 here

Across all studies, a total of 25 distinctly different sensor mounting locations were explored. For the purpose of this review, the 25 mounting locations have been grouped into 10 general locations (Fig. 3); however, full details of the specific mounting locations are provided in table 3. Seven studies (7 validity, 1 reliability) investigated multiple sensor mounting locations, with two quantifying PC performance by combining data from both sensors. Only two studies compared multiple mounting locations directly, with Brown and colleagues [13] reporting that the head mounted tri-axial accelerometer/gyroscope had the best agreement with the clinical BESS, when compared to the sternum, waist, wrist and shank sensors. Additionally, Shah and colleagues [25] reporting that the thigh mounted tri-axial accelerometer possessed the greatest sensitivity compared to shank and waist tri-axial accelerometers, when differentiating static stance conditions.

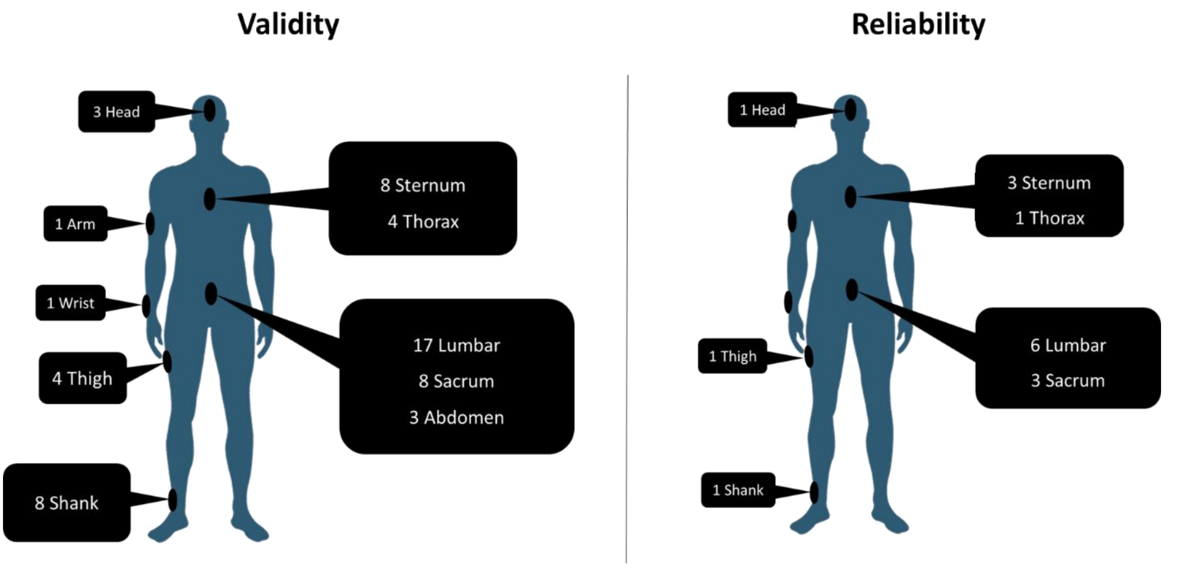

Fig 3: Illustrates the sensor mounting region utilised in the validity and reliability studies. The size of the rectangle is dependent on the number of studies which utilised that general sensor mounting location.

Similar to the sensor mounting locations, there was a large degree of heterogeneity in the variables used across the included studies. Fig. 4 illustrates the general type of inertial sensor derived variables, and the number of studies that included the variable type, with variables most commonly describing 'stability' through quantification of COM acceleration.

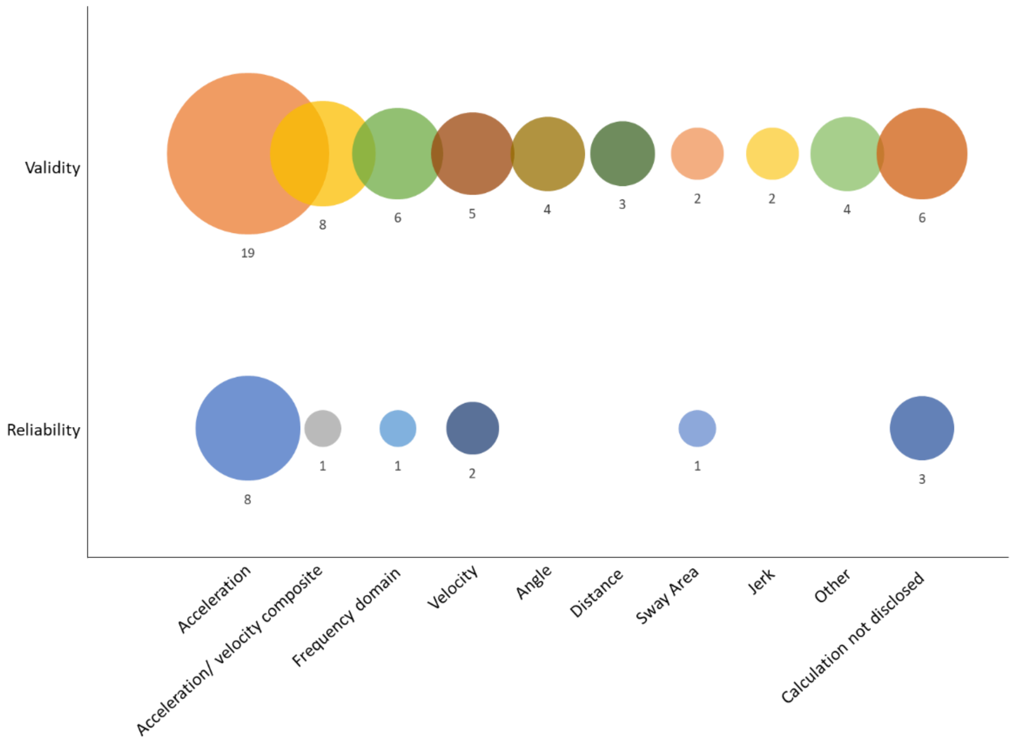

Fig. 4: Illustrates the number of studies which investigated the different types of PC variables. The size of the bubbles represents the number of studies included in each category. Variable type is presented on the x-axis, while study type is

presented on the y-axis. "Other" consists of studies which investigated centre of mass estimates, BESS errors, wavelet variables and classification methods.

3.4 Balance Assessment Protocols

Across the 47 studies investigated in this review, there were a total of 27 different PC assessment protocols used. Thirty-six studies investigated validity, while 11 studies focused on reliability of static PC assessments. The validity and reliability of dynamic PC assessments were investigated in nine and four studies, respectively. The most common methodologies involved using clinical assessment protocols, such as the (BESS)/modified BESS (mBESS) (15/46), sensory organisation test (SOT) (2/46), dynamic postural stability index (DPSI) (1/46), clinical test of sensory integration and balance (CTSIB) (1/46), star excursion balance test (SEBT) (2/46) and Y balance test (YBT) (2/46). A total of 20/46 studies leveraged combinations of static stance conditions (DLS, SLS, TS), eye conditions (EO, EC), stance surfaces (firm, foam), hearing conditions (ear defenders, no ear defenders), and cognitive tasks (cognitive task, no cognitive task). Table 4 details the different PC protocols included in the various reliability and validity studies.

Table 4 here

3.5 Measurement Validity

Forty-two of the included studies were identified as investigating the question of measurement validity. Fourteen of the 42 studies compared the inertial sensor PC assessment to a 'gold-standard' (concurrent validity), while six studies compared the system to a clinical standard assessment (convergence validity). A further 12 studies investigated discriminant validity in healthy adults (age, sex, stance position or condition comparison), 14 studies investigated the discriminant validity in sports medicine populations, while one study investigated the predictive validity. Table 5 details the type of validity investigated in each study.

3.5.1 Concurrent Validity

Of the 14 studies that investigated concurrent validity, ten utilised a force platform system as the gold standard [26-35], while five used 3D optoelectronic motion capture [12, 36-39]. One study used both motion capture and a force platform during the comparison [29]. Three studies investigated the concurrent validity of the BESS/mBESS protocol, with a correlation ranging from r = -0.32 to -0.64 [12, 35] and rho = 0.37 to 0.94, across the different stance conditions [34]. Two studies investigated the SOT protocol, reporting a r^2 = 0.15 to 0.92 and average error of between 5.8% and 10.42% [26, 33] across the different conditions. A further seven studies investigated the concurrent validity of various combinations of static stance positions [27, 30, 32, 36-39]. Of these seven studies, five provided adequate details pertaining to their results to allow comparison, with Abe et al. [36] not directly comparing the gold standard and inertial sensor variables and Seimetz et al. [32] not reporting any quantified statistical comparison.

3.5.2 Convergence Validity

Six studies investigated the convergence validity of inertial sensor-based PC assessments [12, 13, 16, 40-42]. Two studies were excluded from the validity synthesis as they investigated the convergence validity by comparing the measure to a second inertial sensor measurement [40, 41], thus not comparing the method of measurement to another accepted standard. The four remaining studies used a clinical standard comparator (ie BESS/mBESS errors) and reported varying correlation coefficients ranging from r = 0.01 to -0.77, rho = 0.16 to 0.94 and ICC = 0.68 to 0.94 [12, 13, 16, 42].

3.5.3 Discriminant Validity – Healthy population

Twelve studies assessed discriminant validity in healthy cohorts, assessing the ability of inertial sensor systems to discriminate between known stance, visual, hearing, fatigue or taping conditions [25, 28, 32, 41, 43-50].

3.5.4 Discriminant Validity – Sports Medicine Populations

Fourteen studies focused on establishing the discriminant validity in sports medicine clinical populations. Eight observational case-control studies focused on comparing concussed and healthy control groups [14, 15, 51-56], with 5/8 studies demonstrating statistically significant differences between concussion and healthy cohorts. A single prospective cohort study compared athletes healthy baseline and concussed measurements, reporting statistically significant differences between testing points in the DLS EC condition [57]. Two observational case-control studies focused on CAI, reporting conflicting findings when differentiating healthy and CAI [58, 59]. Two observational case-control studies compared healthy and knee injury populations [38, 60], while one study investigated differences in PC performance between a healthy and LBP population [47] demonstrating the capability of the inertial sensor system to discriminate the clinical groups.

3.5.5 Predictive Validity – Sports Medicine Populations

A single prospective cohort study investigated the association of inertial sensor-based PC assessment scores with sports related injuries in a cohort of collegiate American football players, reporting that players with poorer PC were at a 5.19 times greater-odds of sustaining a sports related injury [61].

Table 5 here

3.6 Measurement Reliability

Thirteen of the included studies were identified as assessing measurement reliability [16, 27, 28, 33, 40, 41, 45, 47, 58, 62-65]. Eleven of these studies investigated intra-session test retest reliability, demonstrating reliability ranging from poor to excellent (ICC = 0.03 to 0.97) [16, 27, 33, 40, 41, 45, 47, 58, 63-65], while five studies reported poor to excellent (ICC = 0.02 to 0.95) inter-session reliability [16, 28, 62-64]. Three studies investigated both intra and inter-session test-retest reliability [16, 63, 64]. Table 6 summarises the reliability studies included in this review.

Table 6 here

4 Discussion

The purpose of the systematic review was to examine the literature related to the validity and reliability of inertial sensor-based PC assessments, with a specific focus on sports science and medicine applications. In doing so, this review sought to synthesise the evidence related to the development of these systems, and investigate the evidence surrounding their use in sports medicine. The information provided in this review provides researchers and clinicians with a summary of the evidence on this topic, helping to guide research and enhance future sports medicine clinical practice.

4.1 Sensor Setup & Variables

The studies included within this systematic review employed a variety of sensor setups (locations, number of units, sensor types and variable types), as detailed in table 3. Twenty-eight studies leveraged a single inertial sensor type (accelerometer, gyroscope or magnetometer), whilst 19 used a combination of multiple sensor types. Furthermore, 10 general sensor mounting location groups (Fig 3) and 13 broad types of variables were explored across the studies (Fig.4). The most common sensor location/type utilised were lumbosacral accelerometers (16/47 studies). The likely explanation for the popularity of this methodology is the lumbosacral regions proximity to the COM, with previous reports demonstrating that it closely matches the acceleration of the COM [66, 67]. While seven studies investigated the use of multiple sensor mounting locations, only two studies with differing methodologies directly compared locations to determine which provided the most valuable measure of PC [13, 25]. As such, the authors draw no conclusions pertaining to optimal sensor mounting location.

Only 10/47 of the included studies provided an adequate overview of the outcomes to be measured (variables of interest), gave full information pertaining to the sources of data, and/or detailed the methods of assessment. Authors commonly failed to report the sensor parameters (sampling frequency and configured range) used in the data collection. Failure to report these methods is of importance, as changes in sensor parameters vastly alters the quality and accuracy of the collected data [68]. Additionally, authors frequently did not describe the methods for computing the inertial sensor PC variables, restricting reproducibility, slowing the progress towards sports medicine application.

When considering the use of such technology in research or sports medicine practice, the authors advise users to conduct a thorough needs analysis of the population and balance task in question. For example, when considering the use of an inertial sensor in the quantification of controlled static and dynamic PC tasks, a lumbar mounted tri-axial accelerometer and gyroscope, with a sampling frequency of 50-150Hz and calibration ranges of approximately ±2-4g and 500-1000 deg/sec respectively, is likely sufficient to capture sensorimotor control changes. However, this is largely dependent on the speed of the movement in question; thus, researchers should thoroughly evaluate if the task in question may require a wider calibration range or higher sampling frequency. Additionally, individuals should evaluate the utility of non-traditional descriptive statistics, such as non-linear and frequency domain analysis, as well as advanced machine learning classification and regression methods, as they may capture details in time series data which are not uncovered by traditional time-domain variables [69, 70]. Finally, researchers should consider comparing the reliability and validity of various sensor set-ups and variable selections, to facilitate the development of the optimal system. Article authors, reviewers and editors should ensure that all sensor parameters and signal processing methodologies are adequately reported. This will help to ensure that the best methodology is implemented when investigating the application of this technology in sports medicine clinical populations.

4.2 Measurement Validity

It is clear from this review that the application of inertial sensor technology in the assessment of PC for sports medicine applications is a new and evolving domain, with only three of the included studies published prior to 2007. Using the Downs and Black paper quality rating tool, it was determined that the median score was 6/10 (appendix 1). Generally, the included validity studies lacked information related to the derivation of outcome measure variable of interest (10/43), study hypothesis and objectives (22/43), and the external validity of the recruited cohort (question 11 and 12) (0/43). All other criteria were met by greater than 70% of the studies.

4.2.1 Concurrent Validity

Concurrent validity of the PC assessment protocols was investigated in 14/43 validity studies (Table 5). As outlined in section 4.1, there was a high degree of heterogeneity in the PC protocols and variables used between included validity studies. The most common standardised protocol examined was the BESS/mBESS (4/14 studies). The BESS/mBESS is a standardised static balance assessment protocol, commonly used following sports related concussion [71]. Importantly, three of these studies did not fully disclose the methods for calculating the variable of interest [27, 34, 35]. All four studies demonstrated significant correlations between the 'gold-standard' and the sensor derived variables, with higher correlations typically observed during more challenging stance positions such as SLS and TS [12, 27, 34, 35].

Two studies investigate the concurrent validity of the SOT [26, 33], a protocol which traditionally uses the Neurocom testing platform to 'conflict' or 'remove' one or more balance systems through stance surface or visual surrounding changes [72]. While research related to an inertial sensor instrumented SOT is limited to two studies, this early research indicates that an instrumented SOT may provide a valid measure of PC. However, this study protocol requires an additional platform capable of altering the balance state through the SOT conditions [72]. This is a major limitation of implementing the SOT protocol in clinical practice. Seven out of nine studies investigating static stance positions provided sufficient data to allow for comparisons of concurrent validity [27-31, 37, 38]. Four of these seven studies demonstrated a significant correlation between the lumbar/sternum and force platform variables [27-30]. Furthermore, a single study investigated the concurrent validity of inertial sensors in quantifying dynamic PC, demonstrating that the sensor-based DPSI accelerometer variables were highly correlated with the force platform measures [28].

The studies included in this review demonstrate that inertial sensor measures of PC demonstrate varying levels of concurrent validity, depending on the sensor set-up, balance protocol and variables of interest. These findings are expected as the inertial sensor variables and 'gold standard' measures are typically measuring somewhat different constructs. For example, force platform measures of PC are derived from COP excursion, while 3D motion capture systems provide a measure of COM excursion. In contrast, inertial sensor PC assessments typically leverage acceleration and/or velocity measures, not COP or COM measures. While it may be expected that there is a significant relationship between the two measurement types, one cannot expect there to be a perfect correlation or agreement. As such, the authors of this review contest that future research investigating the validity of inertial sensor-based PC assessments should focus on establishing the convergence, discriminant and predicative validity of the assessments, unless the researchers are attempting to directly estimate COP or COM based measures.

4.2.2 Convergence Validity

Four studies investigated the convergence validity of the sensor-based PC assessments by comparing the relationship between the sensor derived variables and clinical standard comparators. Three of these studies examined the relationship between the BESS/mBESS and the clinical error scores, reporting a strong correlation between the sensor and clinical measures. Brown et al. [13] developed a method for identifying BESS errors, demonstrating that the IMU system can identify errors with excellent agreement. While this method may aid in the development of an automated assessment, it would not allow for the capture of information pertaining to 'stability' or 'strategy', hypothesised to provide a more sensitive measure of performance [6], thus being restricted by similar limitations to the clinical method.

The studies included in this review which investigated the convergence validity of inertial sensor-based PC assessments primarily focused on the BESS/mBESS protocols. These studies demonstrated that there is a significant correlation between the traditional measures of the BESS/mBESS and the sensor-based variables; however, the traditional error scores do not account for 100% of the variance in the inertial sensor-based measures. This would suggest that the sensor-based measures may provide information that is not captured by the tradition clinical measures. The authors hypothesise that the inertial sensor-based measures of PC may provide a more sensitive measure of PC than the traditional clinical methods, such as the BESS/mBESS error score. However, to date, the clinical relevance of this increased sensitivity has yet to be thoroughly evaluated. Thus, investigations across various clinical domains should be conducted to determine the clinical relevance of this information.

4.2.3 Discriminant Validity - Healthy Populations

Two studies investigated the discriminant validity of the quantified BESS/mBESS by investigating its ability to differentiate age groups and/or sex, with both studies reporting significant differences [17, 73]. These findings are in line with previously published work which has demonstrated that males possess poorer PC than females, while PC improves with the transition from youth to young adulthood [74]. A further nine studies compared PC performance during different static stance conditions [29, 32, 41, 43, 44, 46-48, 75]. Two studies failed to quantitatively compare conditions [32, 43], while the seven remaining studies demonstrated significant differences between conditions, with more challenging stances causing increased instability [25, 29, 41, 44, 46, 48, 76]. As such, it is clear that the inertial sensor based measures of PC are capable of discriminating age group and/or sex differences, reproducing findings which have previously been observed in studies using 'gold-standard' comparisons [77].

Two studies investigated the ability of a lumbar mounted sensor to differentiate fatigue and non-fatigued dynamic PC performance during the YBT. Johnston et al. [49] demonstrated that a machine learning approach can classify PC performance with 61.9 - 71.4% accuracy, and differentiate YBT reach directions with 97.8% accuracy. Additionally, Johnston et al. [45] reported significant differences between fatigued and non-fatigued performance, capturing alterations in balance not detected by the traditional measure, suggesting greater sensitivity. Furthermore, a single low quality study demonstrated that a single shank mounted tri-axial gyroscope can capture significant differences in dynamic balance performance between taped, braced and normal conditions during the SEBT [50].

The studies presented above consistently demonstrate the ability of both instrumented static and dynamic PC assessments to validly differentiate between known groups and/or conditions in healthy populations. Across the seven static PC protocol studies and three dynamic PC assessment studies, it was seen that the inertial sensors could discriminate perfromance. However, the majority of the studies did not directly compare the discriminant ability of the sensor-based measures to the traditional clinical measures. As such, the authors advise that further discriminant validity studies should primarily focus on clinical populations, leverage high-quality and consistent study methodologies and ensure that there is a direct comparison of the sensitivity and specificity of the traditional and inertial sensor-based measures.

4.2.4 Discriminant Validity – Sports Medicine Populations

Only 15/42 validity studies completed to date have investigated the application of these systems in sports medicine clinical populations (Fig.2). These studies investigated a variety of sports medicine populations in a series of case-control (13/42) and prospective (2/42) studies; however, it is important to note that 6/14 studies leveraged small samples sizes, recruiting less than 30 injured participants.

4.2.4.1 Concussion

Eight observational case-control studies [14, 15, 51-56] and a single prospective cohort study [57] investigated the effect of concussion on the instrumented PC assessment scores. Four out of five BESS/mBESS case-control studies [15, 53-55] and a single CTSIB protocol [56] demonstrated that sensor derived measures were able to differentiate concussed and non-concussed subjects with a greater degree of accuracy than the traditional measures. Additionally, Bernstein and colleagues [57] prospective investigation reported that the sensor-based measure detected deteriorations in DLS EC balance performance in athletes, post-concussion. In contrast, two studies reported that the instrumented BAM [14] and DLS EO (with/without cognitive task) [51] did not differentiate concussed and healthy control groups. While it is not directly clear why these two studies demonstrated conflicting findings to the above studies, this may be due to differing study methodologies. For example, Furman et al. [14] leveraged an abdomen mounted bi-axial accelerometer in the quantification of the BAM, while the remainder of the concussion studies used a lumbosacral mounted inertial sensor. Additionally, the conflicting findings observed by Berkner et al. [51] may be explained by the protracted recruitment period and the heterogeneous population which would be recruited as a result of this.

In summary, these studies demonstrate that the instrumented static PC assessments may provide an increased level of accuracy in identifying concussion related deficits than the traditional clinical methods. However, no studies have investigated the role of instrumented dynamic PC assessments in concussed populations, which may challenge the sensorimotor subsystems more sufficiently than static assessments, highlighting subtle deficits [6]. As such, while there is sufficient evidence to suggest static inertial sensor PC assessments can validly discriminate concussed and healthy groups, only a single study prospectively evaluated this technology in the concussion use-case. As such, the authors advise that further research should focus on prospective studies, designed to establish the role that both static and dynamic PC systems may play in aiding concussion diagnosis and return to play decisions prior to their implementation into clinical practice.

4.2.4.2 Musculoskeletal Injuries

Six observational case-control studies investigated the discriminant validity of inertial sensor-based PC assessments in musculoskeletal injury populations, with two CAI studies [58, 59], two knee injury studies [38, 60], one LBP study [47] and one general sports injury study [61]. Both CAI studies leveraged differing dynamic PC assessments, reporting conflicting findings, with Brown et al. [58] demonstrating no significant difference between 'stable' and 'CAI' groups in the static stance protocol, while Martinez-Ramírez et al. [59] reported statistically significant differences between the groups using the instrumented SEBT. While there was no significant difference in the AP maximal acceleration magnitude between groups in the TTS task, there was a significant difference when considering the force plate derived measure of TTS. While the AP maximal acceleration magnitude variable failed to capture these changes, it is likely that acceleration magnitude max may not be the most appropriate variable of choice in this task, with different descriptive statistics such as RMS, range and variance potentially replicating the findings of the force plate. A further two studies reported positive findings related to the utility of sensor-based PC assessments in knee injury populations; however, both studies were significantly limited by the small sample sizes (n < 10). The single large LBP study reported that thigh and thoracic spine sensors can captured statistically significant differences between healthy and control groups, during a DLS EC task on a firm surface. Finally, a single small, prospective cohort study investigated the association between sensor-based PC variables, as measured by a thoracic spine mounted accelerometer during a unilateral forefoot squat and general sports related injuries. This study leveraged a small cohort of NCAA American football players, reporting that poor PC was associated with a 5.19 greater-odds of sustaining a sports related injury during the season. While the results presented in this study suggest that there may be value in sensor-based measures of PC in identifying athletes at an increased risk of injury, the small sample size and low study quality mean that these findings should be observed with caution.

The six studies which investigated the utility of sensor-based measures of PC in musculoskeletal injury contexts have demonstrated the feasibility and potential utility of such methods in the evaluation of athletes. However, the authors advise that further large high-quality prospective research is required to investigate the utility of such assessment protocols in various populations, prior to the introduction into clinical practice. Researchers should also investigate the role that both instrumented static and dynamic PC assessments may play in risk-factor screening, as well as across the different stages of the injury rehabilitation process.

4.3 Measurement Reliability

A total of 13/47 studies investigated the reliability of sensor-based PC assessments. Eleven of these assessed the intra-session test-retest reliability. A further five studies investigated the inter-session test-retest reliability. Similarly to the validity studies, a low proportion (3/13) provided full details related to the sources of data and methods used to calculate the sensor derived PC variables. Only 8/13 studies provided details related to study objectives and hypothesis. No reliability studies provided details related to the external validity of the studies (questions 11 and 12). The remaining criteria were met by at least 70% of included studies (appendix 1).

4.3.1 Intra-session Reliability

Eight studies [16, 27, 40, 41, 47, 63-65] investigated the intra-session reliability of instrumented static PC assessments. The reliability results reported by these eight studies varied considerably, ranging from poor – excellent (ICC = 0.03 to 0.97), depending on the stance condition, sensor mounting location, or variable of choice. Simon et al. [16] reported the highest intra-session test-retest reliability for the 95% ellipsoid volume variable derived from a lumbar mounted accelerometer and gyroscope (ICC = 0.97 [95% CI = 0.94 - 0.98]). Three studies leveraged the 'sway balance score' variable but failed to describe the methods related to calculating this variable. The intra-session reliability across these three studies ranged from ICC = 0.20 (95% CI 0.00 - 0.49) for DLS EO to ICC = 0.89 (95% CI 0.84 - 0.93) for a summative score across the difference stances. The remaining three studies investigated the intra-session reliability of dynamic PC assessments [33, 45, 58]. These studies reported the intra-session reliability to be excellent for the YBT (ICC = 0.76 - 92) [45], good-excellent for the TTS (ICC = 0.66 – 0.98) [58], and poor-excellent for the various SOT stance conditions (ICC = 0.16 – 0.80) [78].

The intra-session reliability results presented in this review demonstrate that various instrumented PC assessments can provide a reliable intra-session measure of PC. This is of importance as acceptable levels of intra-session reliability would imply the presence of a low degree of random measurement variance, suggesting it may have value in a clinical scenario. However, if these assessments are to be implemented into sports medicine practice, it is imperative that the inter-session test retest reliability of the tool is established to truly understand the degree of inherent variability, facilitating the development of clinically relevant thresholds of change.

4.3.2 Inter-session Reliability

The inter-session reliability studies demonstrated a high degree of heterogeneity in methodology and findings, limiting the synthesis and interpretability of these studies. Four of the five inter-session studies included in this review assessed static stance protocols [16, 62-64]. The static stance protocols demonstrated poor – excellent (ICC = 0.02 – 0.95), with SLS EC jerk demonstrating the highest reliability (ICC = 0.95 [95% CI not disclosed]) and TS EC possessing the poorest reliability (ICC = 0.02 [95% CI not disclosed]). A single study investigated the reliability of the DPSI PC assessment, reporting excellent inter-session reliability, ranging from ICC = 0.84 (95% CI 0.33 -0.96) for RMS acceleration in the AP plane and ICC = 0.92 (95% CI = 0.70 – 0.98) for resultant acceleration [28].

Importantly, only five studies included in this review investigated the inter-session test-retest reliability. This is of note as the typical purported use case for these PC systems in sports medicine applications would be to aid in the diagnosis and decision-making process following injuries. However, without establishing the inter-session reliability, it is not possible to fully understand the amount of inherent variability captured by the assessments over time and determine the minimal detectable change required for a change to be considered real [79, 80]. Furthermore, only one study provided an estimate of minimal detectable change, limiting the clinical utility of these tools. Simply reporting the intra-session reliability of a valid measurement tool is not enough, and the inter-session reliability and minimal detectable change of a system must be investigated and reported prior to implementation into clinical practice. Importantly, only three instrumented PC protocols have established both the inter-session and intra-session reliability [16, 63, 64]. As such, the authors would advise researchers and clinicians to ensure the test-retest reliability and minimal detectable change of instrumented PC protocols is fully investigated prior to clinical use.

4.4 Review Limitations

Despite the many strengths of this systematic review, it is imperative that readers consider several contextual factors when interpreting these findings. Firstly, this review only included articles which have been published in the English language, potentially resulting in the exclusion of studies. However, it is likely that authors of high-quality studies would pursue publication in high-impact journals, published in the English language to ensure optimal dissemination. As such, it is unlikely that this review excluded studies which may have significantly altered the conclusions. Secondly, the studies included in this review were broadly heterogeneous, with variable study quality, methodologies, sensor setups, PC variables and balance assessment protocols. This is a consequence of a novel and expanding research field that has yet to establish the optimal setup and assessment protocols. Importantly, this highlights a valuable strength of this review, as it identifies the large disparity between studies, synthesising the evidence and identifying protocols and methods that should be considered for future research and sports medicine clinical practice. Thirdly, this review focused solely on discrete static and dynamic PC assessments and did not consider gait-based stability assessments for inclusion. While this excludes one component of dynamic PC from the review, the introduction of another heterogeneous groups of studies would have significantly hampered the authors ability to synthesise and draw meaningful conclusions from the research. Finally, this review did not include studies which investigated the validity and reliability of sensor-based PC assessments in older adult populations, focusing exclusively on assessment protocols that may be applied to the sports medicine context. This was deemed appropriate by the authors due to the increased degree of heterogeneity that including older adult populations would introduce to the review.

4.5 Practical Implications & Future Research

With the decreasing cost of mobile technology and the increasing interest of its application in sports medicine, the number of commercially available systems purporting their benefits is likely to rise over the coming years. However, while researchers have taken the first step towards validating inertial sensor-based systems for PC assessments in sports medicine contexts, it is important to note that there are several critical steps that need to be taken prior to implementation into sports-medicine practice.

Researchers should focus on the instrumentation of established traditional sports medicine static and dynamic PC assessments. In doing so, researchers should consider initially leveraging multi-sensor setups, prior to establishing the most valuable mounting location, sensor type/combination and variables of interest. This process should be completed by first determining what variables can provide a reliable intra-session measure of PC. Based on these findings, the validity of the system should be investigated. If researchers are attempting to directly replicate PC variables given by ‘gold standard’ measures, then concurrent validity should be investigated. If not, researchers should first focus on establishing the convergence and discriminant validity of the system in both healthy and clinical populations. Once the validity of the sensor-setup and PC variables have been established, researchers should then investigate the inter-session test-retest reliability and report the minimal detectable change of the system to help establish clinically relevant thresholds. Finally, the clinical application of these system should be investigated in large, high-quality prospective cohort studies to investigate the role they may play in injury risk-factor identification, diagnosis and management.

4.6 Conclusion

Inertial sensor-based PC assessments for use in sports medicine applications are a burgeoning technology. Research to date has primarily focused on demonstrating the validity and reliability of such systems in laboratory environments. However, this research is broadly heterogeneous in nature, investigating a range of sensor setups, balance protocols (primarily static) and sensor variables, capable of describing PC 'stability' and 'strategy'. As such, research has failed to systematically evaluate the intra and inter-session reliability and appropriate validity of many sensor-based systems. Early case-control investigations have investigated the ability of this technology to discriminate a range of sports medicine clinical populations, with some promise being demonstrated in concussion applications. However, this research is significantly limited by small sample sizes, low study quality and the case-control study design. Additionally, this research has primarily focused on the use of static assessment protocols, potentially not capable of sufficiently challenging the sensorimotor systems in high functioning sporting populations. As such, further research is required to evaluate their utility in high-quality prospective studies prior to implementation into sports medicine clinical practice. Furthermore, it is imperative that researchers ensure that the inter-session reliability of these tools are robustly evaluated to ensure that minimal detectable changes can be determined and leveraged in clinical practice. Researchers and clinicians should ensure that they carry out a thorough evaluation of the technology in question, prior to implementation into research or clinical practice, to ensure that they are leveraging a truly valid and reliable measurement tool. Such steps will be critical in producing knowledge that will progress this technology from laboratory-based concept, to clinical-based reality.

5 REFERENCES

1. Hildebrandt C, Müller L, Zisch B, Huber R, Fink C, Raschner C. Functional assessments for decision-making regarding return to sports following ACL reconstruction. Part I: development of a new test battery. Knee Surgery, Sports Traumatology, Arthroscopy. 2015;23(5):1273-81.
2. Bell DR, Guskiewicz KM, Clark MA, Padua DA. Systematic review of the balance error scoring system. Sports health. 2011;3(3):287-95.
3. Stiffler MR, Bell DR, Sanfilippo JL, Hetzel SJ, Pickett KA, Heiderscheit BC. Star Excursion Balance Test Anterior Asymmetry Is Associated With Injury Status in Division I Collegiate Athletes. J Orthop Sports Phys Ther. 2017;47(5):339-46.
4. Woollacott MH, Tang P-F. Balance Control During Walking in the Older Adult: Research and Its Implications. Phys Ther. 1997;77(6):646-60.
5. Yim-Chiplis PK, Talbot LA. Defining and measuring balance in adults. Biol Res Nurs. 2000;1(4):321-31.
6. Johnston W, Coughlan GF, Caulfield B. Challenging concussed athletes: the future of balance assessment in concussion. QJM. 2016.
7. Johnston W, Doherty C, Büttner FC, Caulfield B. Wearable sensing and mobile devices: the future of post-concussion monitoring? Concussion (London, England). 2017;2(1):CNC28-CNC.
8. O'Reilly M, Caulfield B, Ward T, Johnston W, Doherty C. Wearable Inertial Sensor Systems for Lower Limb Exercise Detection and Evaluation: A Systematic Review. Sports Medicine. 2018;48(5):1-26.
9. Giggins O, Sweeney KT, Caulfield B. The use of inertial sensors for the classification of rehabilitation exercises. IEEE Eng Med Biol Soc; 2014; 2014. p. 2965-8.
10. Greene BR, Redmond SJ, Caulfield B. Fall Risk Assessment Through Automatic Combination of Clinical Fall Risk Factors and Body-Worn Sensor Data. IEEE J Biomed Health Inform. 2017;21(3):725-31.
11. Pantall A, Din SD, Rochester L. Longitudinal changes over thirty-six months in postural control dynamics and cognitive function in people with Parkinson's disease. Gait Posture. 2018;62:468-74.
12. Alberts JLT, A.; Hirsch, J.; Ozinga, S.; Dey, T.; Schindler, D. D.; Koop, M. M.; Burke, D.; Linder, S. M. Quantification of the Balance Error Scoring System with Mobile Technology. Med Sci Sports Exerc. 2015;47(10):2233-40.
13. Brown HJ, Siegmund GP, Guskiewicz KM, Van Den Doel K, Cretu EB, J. S. Development and validation of an objective balance error scoring system. Med Sci Sports Exerc. 2014;46(8):1610-6.
14. Furman GR, Lin CC, Bellanca JL, Marchetti GF, Collins MW, Whitney SL. Comparison of the balance accelerometer measure and balance error scoring system in adolescent concussions in sports. Am J Sports Med. 2013;41(6):1404-10.
15. King LA, Horak FB, Mancini M, et al. Instrumenting the balance error scoring system for use with patients reporting persistent balance problems after mild traumatic brain injury. Arch Phys Med Rehabil. 2014;95(2):353-9.
16. Simon M, Maerlender A, Metzger K, Decoster L, Hollingworth A, Valovich McLeod T. Reliability and Concurrent Validity of Select C3 Logix Test Components. Dev Neuropsychol. 2017;42(7-8):446-59.
17. Linder SM, Ozinga SJ, Koop MM, et al. Cleveland Clinic-Postural Stability Index Norms for the Balance Error Scoring System. Med Sci Sports Exerc. 2018.
18. Hubble RP, Naughton GA, Silburn PA, Cole MH. Wearable sensor use for assessing standing balance and walking stability in people with Parkinson's disease: a systematic review. PLoS One. 2015;10(4):e0123705.
19. Gordt K, Gerhardy T, Najafi B, Schwenk M. Effects of Wearable Sensor-Based Balance and Gait Training on Balance, Gait, and Functional Performance in Healthy and Patient Populations: A

Systematic Review and Meta-Analysis of Randomized Controlled Trials. Gerontology. 2018;64(1):74-89.
20. Moher D, Liberati A, Tetzlaff J, Altman DG, The PG. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLOS Medicine. 2009;6(7):e1000097.
21. Lipscomb CE. Medical subject headings (MeSH). Bull Med Libr Assoc. 2000;88(3):265.
22. Office UNS. Provisional Guidelines on Standard International Age Classifications: New York: United Nations; 1982.
23. Fisher ST. The intra-session and inter-session reliability of centre-of-pressure based measures of postural sway within a normal population: UNITEC Institute of Technology; 2010.
24. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. J Epidemiol Community Health. 1998;52(6):377-84.
25. Shah N, Aleong R, So I. Novel Use of a Smartphone to Measure Standing Balance. JMIR Rehabil Assist Technol. 2016;3(1):e4.
26. Alberts JL, Hirsch JR, Koop MM, et al. Using Accelerometer and Gyroscopic Measures to Quantify Postural Stability. J Athl Train. 2015;50(6):578-88.
27. Burghart M, Craig J, Radel J, Huisinga J. Reliability and validity of a mobile device application for use in sports-related concussion balance assessment. Current Research: Concussion. 2017;4(1):e1-e6.
28. Heebner NR, Akins JS, Lephart SM, Sell TC. Reliability and validity of an accelerometry based measure of static and dynamic postural stability in healthy and active individuals. Gait Posture. 2015;41(2):535-9.
29. Neville C, Ludlow C, Rieger B. Measuring postural stability with an inertial sensor: validity and sensitivity. Med Devices (Auckl). 2015;8:447-55.
30. Patterson JA, Amick RZ, Thummar T, Rogers ME. Validation of measures from the smartphone sway balance application: a pilot study. Int J Sports Phys Ther. 2014;9(2):135-9.
31. Rouis A, Rezzoug N, Gorce P. Validity of a low-cost wearable device for body sway parameter evaluation. Comput Methods Biomech Biomed Engin. 2014;17 Suppl 1:182-3.
32. Seimetz C, Tan D, Katayama R, Lockhart T. A comparison between methods of measuring postrual stability: force plates versus accelerometers. Biomed Sci Instrum. 2012;48:386-92.
33. Whitney SL, Roche JL, Marchetti GF, et al. A comparison of accelerometry and center of pressure measures during computerized dynamic posturography: a measure of balance. Gait Posture. 2011;33(4):594-9.
34. Dabbs NC, Sauls NM, Zayer A, & Chander H. Balance Performance in Collegiate Athletes: A Comparison of Balance Error Scoring System Measures. J Funct Morphol Kinesiol. 2017;2(3):26.
35. Mohamed ME, Neveen A. Abd-Elraouf, Khaled E. Ayad, Enas E. Abu-Taleb. Validity of using smart phone sway balance application in measuring dynamic balance. Int J Rehabil Res. 2017;6(4):31-7.
36. Abe Y, Sakamoto M, Nakazawa R, Shirakura K. Relationship between joint motion and acceleration during single-leg standing in healthy male adults. J Phys Ther Sci. 2015;27(4):1251-6.
37. Betker AL, Moussavi ZMK, Szturm T. Center of mass approximation and prediction as a function of body acceleration. IEEE Transactions on Biomedical Engineering. 2006;53(4):686-93.
38. Kim KJ, Agrawal V, Bennett C, Gaunaurd I, Feigenbaum L, Gailey R. Measurement of lower limb segmental excursion using inertial sensors during single limb stance. J Biomech. 2018;71(11).
39. Neville CL, C.; Rieger, B. Measuring postural stability with an inertial sensor: validity and sensitivity. Med Devices (Auckl). 2015;8:447-55.
40. Kosse NM, Caljouw S, Vervoort D, Vuillerme N, Lamoth CJ. Validity and Reliability of Gait and Postural Control Analysis Using the Tri-axial Accelerometer of the iPod Touch. Ann Biomed Eng. 2015;43(8):1935-46.

41. Salisbury JP, Keshav NU, Sossong AD, Sahin NT. Concussion Assessment With Smartglasses: Validation Study of Balance Measurement Toward a Lightweight, Multimodal, Field-Ready Platform. JMIR Mhealth Uhealth. 2018;6(1):e15.
42. Patterson JA, Amick RZ, Pandya PD, Hakansson N, Jorgensen MJ. Comparison of a mobile technology application with the balance error scoring system. Int J Athl Ther Train. 2014;19(3):4-7.
43. Bonnet SC, P.; Favre-Reguillon, F.; Guillemaud, R. Evaluation of postural stability by means of a single inertial sensor. IEEE Eng Med Biol Soc. 2004;3:2275-8.
44. Chiu YL, Tsai YJ, Lin CH, Hou YR, Sung WH. Evaluation of a smartphone-based assessment system in subjects with chronic ankle instability. Comput Methods Programs Biomed. 2017;139:191-5.
45. Johnston WOR, M.; Coughlan, G. F.; Caulfield, B. Inertial Sensor Technology Can Capture Changes in Dynamic Balance Control during the Y Balance Test. Digit Biomark. 2017;1(2):106-17.
46. Oliva Dominguez M, Bartual Magro J, Roquette Gaona J, Bartual Pastor J. Spectrum analysis in postural strategy on static tests in a healthy population. Acta Otorrinolaringol Esp. 2013;64(2):124-32.
47. Schelldorfer S, Ernst MJ, Rast FM, Bauer CM, Meichtry A, Kool J. Low back pain and postural control, effects of task difficulty on centre of pressure and spinal kinematics. Gait Posture. 2015;41(1):112-8.
48. Yvon C, Najuko-Mafemera A, Kanegaonkar R. The D+R Balance application: a novel method of assessing postural sway. J Laryngol Otol. 2015;129(8):773-8.
49. Johnston W, O'Reilly M, Dolan K, Reid N, Coughlan G, Caulfield B. Objective classification of dynamic balance using a single wearable sensor. 4th International Congress on Sport Sciences Research and Technology Support 2016, Porto, Portugal, 7-9 November 2016. 2016:15-24.
50. Budini K, Richards J, Cole T, et al. An exploration of the use of Inertial Measurement Units in the assessment of dynamic postural control of the knee and the effect of bracing and taping. Physiother Pract Res. 2018;39(2):91-8.
51. Berkner J, Meehan WP, Master CL, Howell DR. Gait and Quiet-Stance Performance Among Adolescents After Concussion-Symptom Resolution. J Athl Train. 2017;52(12):1089-95.
52. Borges A, Raab S, Lininger M. A comprehensive instrument for evaluating mild truamatic brain injury (MTBI)/ concussion in independent adults: a pilot study. Int J Sports Phys Ther. 2017;12(3):381-9.
53. Doherty C, Zhao L, Ryan J, Komaba Y, Inomata A, Caulfield B. Quantification of postural control deficits in patients with recent concussion: An inertial-sensor based approach. Clin Biomech (Bristol, Avon). 2017;42:79-84.
54. King LA, Mancini M, Fino PC, et al. Sensor-Based Balance Measures Outperform Modified Balance Error Scoring System in Identifying Acute Concussion. Ann Biomed Eng. 2017;45(9):2135-45.
55. Baracks J, Casa DJ, Covassin T, et al. Acute Sport-Related Concussion Screening for Collegiate Athletes Using an Instrumented Balance Assessment. J Athl Train. 2018;53(6):174-17.
56. Gera G, Chesnutt J, Mancini M, Horak FB, King LA. Inertial Sensor-Based Assessment of Central Sensory Integration for Balance After Mild Traumatic Brain Injury. Mil Med. 2018;183(suppl_1):327-32.
57. Bernstein JPK, Calamia M, Pratt J, Mullenix S. Assessing the effects of concussion using the C3Logix Test Battery: An exploratory study. Appl Neuropsychol Adult. 2018:1-8.
58. Brown CN, Mynark R. Balance deficits in recreational athletes with chronic ankle instability. J Athl Train. 2007;42(3):367-73.
59. Martínez-Ramírez AL, P.; Gómez, M.; Izquierdo, M. Wavelet analysis based on time-frequency information discriminate chronic ankle instability. Clin Biomech (Bristol, Avon). 2010;25(3):256-64.
60. Senanayake SM, Malik OA, Iskandar M, Zaheer D. 3-D kinematics and neuromuscular signals' integration for post ACL reconstruction recovery assessment. Conf Proc IEEE Eng Med Biol Soc. 2013;2013:7221-5.

61. Wilkerson GB, Gupta A, Colston MA. Mitigating Sports Injury Risks Using Internet of Things and Analytics Approaches. Risk Anal. 2018;38(7):1348-60.
62. Han S, Lee D, Lee S. A study on the reliability of measuring dynamic balance ability using a smartphone. J Phys Ther Sci. 2016;28(9):2515-8.
63. Williams JM, Dorey C, Clark S, Clark C. The within-day and between-day reliability of using sacral accelerations to quantify balance performance. Phys Ther Sport. 2016;17:45-50.
64. Amick RZ, Chaparro A, Patterson JA, Jorgensen MJ. Test-retest reliability of the sway balance mobile application. Journal MTM. 2015;4(2):40–7.
65. Dunn KL, Bay RC, Cardenas JF, Anastasi M, Williams RM, McLeod TV. Reliability of the Sway Balance Mobile Application: A Retrospective Analysis. Int J Athl Ther Train. 2017;23(2):69-72.
66. Moe-Nilssen R. A new method for evaluating motor control in gait under real-life environmental conditions. Part 1: The instrument. Clin Biomech (Bristol, Avon). 1998;13(4-5):320-7.
67. Zijlstra W, Hof AL. Assessment of spatio-temporal gait parameters from trunk accelerations during human walking. Gait Posture. 2003;18(2):1-10.
68. Mitschke C, Kiesewetter P, Milani TL. The Effect of the Accelerometer Operating Range on Biomechanical Parameters: Stride Length, Velocity, and Peak Tibial Acceleration during Running. Sensors (Basel). 2018;18(1).
69. Stergiou N, Decker LM. Human movement variability, nonlinear dynamics, and pathology: is there a connection? Hum Movement Sci. 2011;30(5):869-88.
70. Bravi A, Longtin A, Seely AJ. Review and classification of variability analysis techniques with clinical applications. Biomed Eng Online. 2011;10:10-90.
71. McCrory P, Meeuwisse W, Dvorak J, et al. Consensus statement on concussion in sport—the 5th international conference on concussion in sport held in Berlin, October 2016. Br J Sports Med. 2017;51:838-47.
72. Broglio SP, Ferrara MS, Sopiarz K, Kelly MS. Reliable change of the sensory organization test. Clin J Sport Med. 2008;18(2):148-54.
73. Anderson SL, Gatens D, Glatts C, Russo SA. Normative Data Set of SWAY Balance Mobile Assessment in Pediatric Athletes. Clin J Sport Med. 2017;EPub.
74. Paniccia M, Wilson KE, Hunt A, et al. Postural Stability in Healthy Child and Youth Athletes: The Effect of Age, Sex, and Concussion-Related Factors on Performance. Sports Health. 2017;10(2):175-82.
75. Shah NA, R.; So, I. Novel Use of a Smartphone to Measure Standing Balance. JMIR Rehabil Assist Technol. 2016;3(1):e4.
76. Schelldorfer SE, M. J.; Rast, F. M.; Bauer, C. M.; Meichtry, A.; Kool, J. Low back pain and postural control, effects of task difficulty on centre of pressure and spinal kinematics. Gait Posture. 2015;41(1):112-8.
77. Prieto TE, Myklebust JB, Hoffmann RG, Lovett EG, Myklebust BM. Measures of postural steadiness: differences between healthy young and elderly adults. IEEE Trans Biomed Eng. 1996;43(9):956-66.
78. Whitney SLR, J. L.; Marchetti, G. F.; Lin, C. C.; Steed, D. P.; Furman, G. R.; Musolino, M. C.; Redfern, M. S. A comparison of accelerometry and center of pressure measures during computerized dynamic posturography: a measure of balance. Gait Posture. 2011;33(4):594-9.
79. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J Strength Cond Res. 2005;19(1):231-40.
80. Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. Sports Med. 1998;26(4):217-38.
81. Simon MM, A.; Metzger, K.; Decoster, L.; Hollingworth, A.; Valovich McLeod, T. Reliability and Concurrent Validity of Select C3 Logix Test Components. Dev Neuropsychol. 2017;42(7-8):446-59.
82. Betker ALM, Z. M. K.; Szturm, T. Center of mass approximation and prediction as a function of body acceleration. IEEE Transactions on Biomedical Engineering. 2006;53(4):686-93.

83. Kosse NMC, S.; Vervoort, D.; Vuillerme, N.; Lamoth, C. J. Validity and Reliability of Gait and Postural Control Analysis Using the Tri-axial Accelerometer of the iPod Touch. Ann Biomed Eng. 2015;43(8):1935-46.
84. Chiu YLT, Y. J.; Lin, C. H.; Hou, Y. R.; Sung, W. H. Evaluation of a smartphone-based assessment system in subjects with chronic ankle instability. Comput Methods Programs Biomed. 2017;139:191-5.
85. Burghart MC, J.; Radel, J.; Huisinga, J. Reliability and validity of a mobile device application for use in sports-related concussion balance assessment. Current Research: Concussion. 2017;4(1):e1-e6.
86. Heebner NRA, J. S.; Lephart, S. M.; Sell, T. C. Reliability and validity of an accelerometry based measure of static and dynamic postural stability in healthy and active individuals. Gait Posture. 2015;41(2):535-9.
87. Senanayake SMM, O. A.; Iskandar, M.; Zaheer, D. 3-D kinematics and neuromuscular signals' integration for post ACL reconstruction recovery assessment. Conf Proc IEEE Eng Med Biol Soc. 2013;2013:7221-5.
88. Williams JMD, C.; Clark, S.; Clark, C. The within-day and between-day reliability of using sacral accelerations to quantify balance performance. Phys Ther Sport. 2016;17:45-50.

***Table 1**: The literature search strategy utilised in this review*

Balance	"Balance" OR "Postural Stability" OR "Postural Control" OR "Stability" OR "Postural"
	AND
Inertial measurement units	“inertial sensor” OR “gyroscope” OR “IMU” OR “inertial measurement units” OR “wearable” OR “accelerometer” OR “sensor system” OR “sensor network” OR “magnetometer” OR “MEMS” OR “smartphone” OR “mobile” OR “wireless”
	AND
Assessment	"Assessment" OR "monitoring" OR "quantification" OR "quantifying" OR "quantify" OR "tracking" OR "outcome measure" OR "evaluation" OR "analysis"
	AND NOT
	“robot” OR “exoskeleton” OR "Ph" OR "chromatography" OR "blood" OR "chemistry" OR "physics" OR "chemical" OR "plasma"

Table 2: Inclusion and exclusion criteria for the review process

Inclusion Criteria

1. Studies contain an inertial sensor system for PC assessment
2. The inertial sensor system is intended for quantifying PC performance during a discrete balance protocol
3. Studies that focus on validity and/or reliability of discrete inertial sensor-based PC assessments
4. Articles were written in English

Exclusion Criteria

1. Systematic review & literature reviews
2. Books & other non-peer reviewed literature
3. Studies investigating human activity recognition
4. Studies investigating fall identification
5. Sensing modality used was not a wearable accelerometer, gyroscope or combination
6. Study investigates gait analysis
7. Study investigates exercise and sport technique analysis
8. Study concerns non-human, animal subjects
9. Studies that investigate postural alignment
10. Studies that focused on traditional clinical measures of balance only.
11. Studies that utilised a non-gold standard comparator
12. Studies that investigate electromyography only
13. Studies that focus on postural activity in muscles.
14. Studies that focus on anticipatory postural adjustments
15. Studies which focused on older adult populations
16. Studies limited to sitting balance assessment
17. Intervention studies

Table 3: An overview of the sensor setups and PC variables utilised in each study.

Articles	Sensor Type	Number of sensors	Specific Sensor Location(s)	Sensor Calibration	Sensor Variable(s)
Abe et al. (2015) [36]	Tri-axial accelerometer	3	Forehead; 3rd lumbar vertebra; above lateral malleolus	SF = 50Hz; accelerometer range ND	RMS acceleration; acceleration median frequency
Alberts et al. (2015a) [26]	Tri-axial accelerometer and gyroscope	1	Sacrum	SF = 100Hz; accelerometer range ± 2g; gyroscope ± 250 ◦/s	COG sway angle derived from accelerometer and gyroscope
Alberts et al. (2015b) [12]	Tri-axial accelerometer and gyroscope	1	Sacrum	SF = 100Hz; accelerometer ± 2g; gyroscope ± 250 ◦/s	Acceleration and gyroscope peak-to-peak normalised path length, root mean squared distance; 95% sway ellipse
Amick et al. (2015) [64]	Tri-axial accelerometer	1	Held to sternum	ND	'Sway balance score' – no calculation details provided
Anderson et al. (2017) [73]	Tri-axial accelerometer	1	Held to sternum	ND	'Sway balance score' – no calculation details provided
Baracks et al. (2018) [55]	IMU	1	3rd/4th lumbar vertebra	ND	RMS sway; 95%; ellipse sway area.
Berkner et al. (2017) [51]	Tri-axial accelerometer and gyroscope	1	Lumbosacral junction	SF = 128 Hz; accelerometer and gyroscope range ND	RMS sway; 95% sway ellipse; RMS sway.
Bernstein et al. 2018 [57]	Tri-axial accelerometer and gyroscope	1	Sacrum	ND	95% sway ellipse
Betkner et al. (2006) [37]	Tri-axial accelerometer	2	2nd thoracic vertebra; below lateral knee joint.	SF = 500Hz; accelerometer range ND	COM estimate
Bonnet et al. (2004) [43]	Tri-axial accelerometer and magnetometer	1	Lateral thorax, level of the brachial plexus	SF = 150Hz; accelerometer ± 1/2g; magnetometer range ND	Azimuth angle change
Borges et al. (2017) [52]	Tri-axial accelerometer and gyroscope	1	Waist	ND	COG sway angle derived from accelerometer and gyroscope
Brown et al. (2007) [58]	Tri-axial accelerometer	1	Anterior tibial crest	ND	AP maximal acceleration magnitude

Brown et al. (2014) [13]	Tri-axial accelerometer and gyroscope	1	Forehead; sternum; anterior waist (below navel); right/ left wrist; right/ left shin	SF = 102.4Hz; accelerometer and gyroscope range ND	IMU estimated BESS errors
Budini et al. (2018) [50]	Tri-axial gyroscope	2	4cm superior to the lateral shank	SF = 148Hz; gyroscope range ND	Range of angular velocity
Burghart et al. (2017) [27]	Tri-axial accelerometer	1	Held to sternum	SF = 10Hz; accelerometer range ND	'Sway balance score' – no calculation details provided
Chiu et al. (2017) [44]	Bi-axial accelerometer	1	Anterior shank (half the distance from the lateral malleolus to fibular head	SF =10Hz; accelerometer range ND	Mean acceleration
Dabbs et al. (2018) [34]	Tri-axial accelerometer	1	Held to sternum	ND	'Sway balance score' – no calculation details provided
Doherty et al. (2017) [53]	Tri-axial accelerometer and gyroscope	1	Sacrum	SF = 102.4Hz; gyroscope ± 1000d/s; accelerometer ± 8g	95% sway ellipse
Dunn et al. (2017) [65]	Tri-axial accelerometer	1	Sternum	ND	'Sway balance score' – no calculation details provided
Furman et al. (2013) [14]	Bi-axial accelerometer (AP and ML axes)	1	Anterior waist	SF = 100Hz; accelerometer ± 1.2g	Balance accelerometery measure (group normalised AP normalised path length)
Gera et al. (2018) [56]	IMU	1	5th lumbar vertebra	ND	95% sway ellipse
Han et al. (2016) [62]	Tri-axial accelerometer and gyroscope	1	3rd/4th lumbar vertebra	ND	Acceleration and gyroscope magnitude
Heebner et al. (2015) [28]	Tri-axial accelerometer	1	5th lumbar vertebra	SF = 1000Hz; accelerometer ± 16g	RMS acceleration (full 10 seconds of static stance); RMS acceleration (three second window beginning at the time of peak vertical acceleration)
Johnston et al. (2016) [49]	IMU	1	4th lumbar vertebra	SF = 102.4Hz; accelerometer ±2g; gyroscope ±500 ◦/s; magnetometer ±1 gauss	Random forest classification algorithm

Johnston et al. (2017) [45]	Tri-axial accelerometer and gyroscope	1	4th lumbar vertebra	SF = 102.4Hz; accelerometer ±2g; gyroscope ± 500◦/s; magnetometer ±1 gauss	95% sway ellipse
Kim et al. (2018) [38]	IMU	5	4 cm below the knee joint line bilateral shanks; 4 cm above the lateral knee joint line; sacrum	SF = 50Hz; accelerometer ± 16g; gyroscope ± 2000 ◦/s; magnetometer ± 0.025 gauss	Region of limb stability
King et al. (2017) [54]	IMU	1	5th lumbar vertebra	ND	RMS, Total power, mean distance, range, 95% sway ellipse, 95 power frequency, path length, total sway area, 95% circle sway area, jerk, high frequency power, centroidal frequency, mean velocity, normalised jerk, frequency dispersion.
King et al. (2014) [15]	Bi-axial accelerometer (AP and ML axes)	1	5th lumbar vertebra	SF = 120Hz; ND	RMS Acceleration
Kosse et al. (2015) [40]	Bi-axial accelerometer (AP and ML axes)	1	3rd lumbar vertebra	SF = 88-92Hz, accelerometer range ND	RMS and Median frequency acceleration; Sway Area
Linder et al. (2018) [17]	IMU	1	Sacrum	ND	Normalised path length
Martínez-Ramírez et al. (2010) [59]	IMU	1	3rd lumbar vertebra	ND	Daubechies 4th wavelet, sum of the coefficients of details 1 and 2 of the pitch, roll and yaw signals
Mohamed et al. (2016) [35]	Tri-axial accelerometer	1	Held to sternum	ND	'Sway balance score' – no calculation details provided
Neville et al. (2015) [29]	Tri-axial accelerometer	1	5th lumbar vertebra	SF = 250 Hz; accelerometer ± 1.7 g	RMS acceleration magnitude; centroidal frequency
Oliva Dominguez et al. (2013) [46]	Tri-axial gyroscope	1	Lumbar spine – specific location ND	ND	Angular velocity spectral density at frequencies of 1.4, 2.5, 3.7, 4.9, 6.1, 7.2, 8.4, 9.6, 10.7, 11.9, 13.1, 14.3, 15.4, 16.8, 18.2 and 19.3 Hz.

Patterson et al. (2014) [42]	Tri-axial accelerometer	1	Held to sternum	ND	'Sway balance score' – no calculation details provided
Patterson et al. (2014) [30]	Tri-axial accelerometer	1	Held to sternum	ND	'Sway balance score' – no calculation details provided
Rouis et al. (2014) [31]	Tri-axial accelerometer	1	5th lumbar vertebra	SF = 50Hz; accelerometer ± 2g	Acceleration, range, mean, sway length, sway area, mean velocity, mean frequency, total power, median frequency, 95% frequency.
Salisbury et al. (2018) [41]	Tri-axial accelerometer	2	Head (glasses) & comparison sensor anterior waist	SF = 50Hz; accelerometer range ND	Normalised path length; total sway magnitude
Schelldorfer et al. (2015) [47]	IMU	4	Right thigh; 2nd sacral vertebra; 1st lumbar vertebra; 1st thoracic vertebra.	SF = 20Hz; range ND	Mean absolute sway deviation; sway velocity
Seimetz et al. (2012) [32]	Tri-axial accelerometer	1	Sternum	ND	Mean Sway velocity
Senanayake et al. (2013) [60]	Tri-axial accelerometer and gyroscope	4	Bilateral shank & thigh	SF = 128Hz; range ND	Knee joint flexion/extension, abduction/adduction and internal/external rotation angle
Shah et al. (2016) [25]	Tri-axial accelerometer	3	Above the talocrural joint line; above superior midline of the patella; level of the subject's umbilicus.	SF = 14-15Hz; range ND	Mean Acceleration magnitude
Simon et al. (2017) [81]	Tri-axial accelerometer and gyroscope	1	Lumbar spine – specific location ND	ND	95% sway ellipse
Whitney et al. (2011) [78]	Bi-axial accelerometer (AP and ML axes)	1	Pelvis (ASIS-PSIS)	SF = 100Hz; accelerometer ± 1.2g	Normalised path length; total acceleration RMS and Peak-to-peak

Wilkerson et al. (2018) [61]	Tri-axial accelerometer	1	Thoracic spine (between scapula)	SF = 100Hz; accelerometer range ND	Jerk magnitude RMS
Williams et al. (2016) [63]	Tri-axial accelerometer and gyroscope	1	Sacrum	SF = 16Hz; range ND	Sway path length; sway jerk; RMS acceleration magnitude
Yvon et al. (2015) [48]	Tri-axial accelerometer	1	Lateral left upper arm	ND	Area of an ellipse with two standard deviations in the anteroposterior and lateral planes about a mean point

IMU inertial measurement unit, *RMS* root mean squared, *COM* centre of mass, *COG* centre of gravity, *ND* not disclosed, *SF* sampling frequency, *ASIS* anterior superior iliac spine, *PSIS* posterior superior iliac spine, *AP* antero-posterior, *ML* medio-lateral, *TP* transverse plane

Table 4: Balance protocols utilised across the different validity and reliability studies.

Balance Protocol	Measurement Validity	Measurement Reliability
Defined Balance Protocols		
Balance assessment measure	[41]	[41]
Modified balance error scoring system/ Balance error scoring system	[12-17, 53-55, 57, 73]	[52, 65, 73, 81]
Biodex system modified balance error scoring system	[34, 35]	
Clinical test of sensory integration and balance	[56]	
Sensory organisation test	[26, 33]	[33]
Y Balance Test	[45, 49]	[45]
Star excursion balance test	[50, 59]	
Dynamic postural stability index	[28]	[28]
Dynamic push test	[43]	
Tandem Static Stance Variations		
Tandem stance eyes open firm	[25, 27, 29, 42]	[27, 63, 64]
Tandem stance eyes closed firm	[29]	[63]
Tandem stance eyes open foam	[29, 48]	
Tandem stance eyes closed foam	[29, 48]	
Tandem stance eyes open foam ear defenders	[48]	
Tandem stance eyes closed foam, ear defenders	[48]	
Tandem stance firm with cognitive task eyes open	[40]	[40]
Tandem stance firm with cognitive task eyes closed	[40]	[40]
Double Leg Static Stance Variations		
Double leg stance eyes open firm	[25, 27, 29, 32, 37, 42, 43, 46-48]	[27, 47, 63, 64]
Double leg stance eyes closed firm	[29, 32, 43, 46-48, 82]	[47, 63]
Double leg stance eyes open foam	[25, 29, 32, 46, 47]	[47]
Double leg stance eyes closed foam	[29, 32, 46, 47]	[47]
Double leg stance firm with cognitive task eyes open	[40, 51]	[40]
Double leg stance firm with cognitive task eyes open	[40]	[83]
Double leg, eyes open, ear defenders	[48]	
Double leg, eyes closed, ear defenders	[48]	
Tasadna Yoga Pose	[31]	
Tibial nerve electrical stimulation time to stabilisation task	[58]	[58]
Single leg Static Stance Variations		
Single leg stance eyes open firm	[25, 27, 30, 36, 38, 43, 60, 84]	[27, 63, 64]
Single leg stance eyes closed firm	[29, 43, 44, 60]	[63]
Single leg stance eyes open foam	[25]	
Single leg stance eyes closed foam	[25]	
Biodex balance system double leg stance eyes open		[62]
Biodex balance system double leg stance eyes closed		[62]
Single leg unilateral forefoot squat	[61]	

Table 5: description of the validity studies included in the review.

Study	Sensor Location(s)	Sample	Population	Validity Type	Comparison	Findings
Abe et al. (2015) [36]	Forehead; 3rd lumbar vertebra; above lateral malleolus	20 healthy males	Healthy youth	Concurrent - gold standard	3D motion capture – EO SLS	There is no significant association between AP/ML head, lumbar or shank acceleration and hip/ankle joint movement.
Alberts et al. (2015a) [26]	Sacrum	49 (22 male; 27 female)	Healthy young adults	Concurrent - gold standard	Force platform – SOT	The accelerometer and gyroscope derived sway equilibrium score was able to track the NeuroCon centre of gravity with an average error ranging from 5.87% to 10.42% across SOT conditions.
Alberts et al. (2015b) [12]	Sacrum	32 (14 male; 18 female)	Healthy youth & young adults	Concurrent - gold standard; Convergence - clinical standard	3D motion capture - BESS	Accelerometer and gyroscope derived variables were significantly correlated (Spearman's correlation = 0.37-0.94) with 3D motion capture. 95% sway ellipse was more sensitive to change than the traditional BESS errors. BESS conditions 2,3 and 6 were significantly correlated with 95% sway ellipse (Spearman's correlation = 0.51-0.55). No significant correlation between conditions 4 & 5 (Spearman's correlation = 0.16 & 0.36 respectively).
Anderson et al. (2017) [73]	Held to sternum	466 (240 male; 226 female)	Healthy youth & adolescents	Discriminant - known groups	Between age group and sex comparisons - mBESS	SWAY balance scores improve with age (4 age groups) across all stance conditions. Females demonstrated statistically significantly lower SWAY score than males during the SLS task.
Baracks et al. (2018) [55]	3rd/4th Lumbar vertebra	93 (63 male; 30 female). 48 case; 45 control. Group sex ND.	Healthy collegiate athlete (control) & concussed collegiate athletes (case)	Discriminant - known groups	Concussion vs healthy - mBESS	Comparison of the concussed and healthy groups demonstrated that there was a significant difference between groups when considering the inertial sensor derived RMS sway and 95% sway ellipse. ROC analysis demonstrated that a change of 0.5 SD or greater resulted in the greatest sensitivity when classifying concussed and non-

						concussed individuals, with a sensitivity = 54% and specificity = 71% for the RMS, and sensitivity = 52% and specificity = 80% for the 95% sway ellipse.
Berkner et al. (2017) [51]	Lumbosacral junction	44 controls (19 male; 25 female); 37 case (17 male; 20 female)	Healthy control adolescents and symptomatic concussed (injured within 21 days)	Discriminant - known groups	Concussion vs healthy - DLS with cognitive task, EO.	No significant differences were observed between concussed and healthy groups at either testing time point during the DLS task (EO), with or without the cognitive task.
Bernstein et al. (2018) [57]	Sacrum	57 (32 male; 25 female)	A cohort of collegiate athletes	Discriminant – known condition	Healthy baseline vs post-concussion score - BESS	Statistically significant increase in DLS EC 95% ellipsoid volume following concussion, when compared to the baseline measure. There was no significant difference observed for between time points for the other stance conditions.
Betkner et al. (2006) [37]	2nd thoracic vertebra; bellow lateral knee joint.	16 (11 male; 5 female)	Healthy young adults	Concurrent - gold standard	3D motion capture – DLS, EO, firm; DLS, EC, firm	The genetic sum of sines model was able to estimate the centre of mass trajectory with a normalised error of between 9.4-10.7%. The neural network demonstrated and average normalised error of 8.4-12.0%, while the adaptive network based fuzzy inference system had an average normalised error ranging from 9.9-12.0% across the four stance conditions.
Bonnet et al. (2004) [43]	Lateral thorax, level of the brachial plexus	1 (sex ND)	Healthy young adult	Discriminant - known condition	Stance condition - DLS EC firm, SLS EO SLS EC & dynamic push test	Visual analysis of the signal demonstrated that the accelerometer derived azimuth angle can differentiate the different balance conditions.
Borges et al. (2017) [52]	Waist	14 (9 male; 5 female) case/ control group size not disclosed	Healthy young adults (control) & concussion (participants with cognitive	Discriminant - known groups	Concussion vs healthy – mBESS	The authors did not report the results for the balance component of the testing protocol. There was a significant difference between concussed and healthy groups when considering the summative scores (dizziness, symptoms,

			impairment 14 days or more post-injury with persistent symptoms (case)			cognitive assessment, balance, vestibular/ ocular screening, king-devick test and C3 Logix battery.
Brown et al. (2007) [58]	Anterior tibial crest	20 case (10 male; 10 female); 20 control (10 male; 10 female)	Young healthy adults (control) & chronic ankle instability (case)	Discriminant - known groups	Chronic ankle instability vs healthy control - Time to stabilisation	There was no significant difference between the "stable" and chronic ankle instability groups when considering the AP maximal acceleration magnitude, despite significant differences in the time to stabilisation in the AP direction between groups.
Brown et al. (2014) [13]	Forehead; sternum; anterior waist (below navel); right/ left wrist; right/ left shin	30 (15 male; 15 female)	Young healthy subjects	Convergence - clinical standard	Comparison of BESS error identification accuracy	All five IMUs (accelerometer and gyroscope) demonstrated excellent agreement (ICC3,1 = 0.94) with the clinical raters, while a single forehead mounted IMU (accelerometer only) was able to identify BESS errors with excellent agreement to the clinical raters (ICC3,1 = 0.92) and predict individual BESS scores (ICC3,1 = 0.90). The IMU system demonstrated an agreement of ICC3,1 = 0.89 when considering the foam surfaces only, while the firm surface conditions demonstrated lower agreement (ICC3,1 = 0.68).
Budini et al. (2018)[50]	Lateral shank	24 (13 male; 11 female)	Young healthy subjects	Discriminant – known condition	Taping and bracing vs no taping and bracing	Statistically significant difference (medium effect size) for gyroscope range between conditions in the coronal ($p = 0.003$) and sagittal plane ($p = 0.009$) between conditions. There was a significant difference between the clinical SEBT score in the anterior ($p = 0.028$) and posteromedial ($p = 0.010$) reach directions. Pairwise comparisons showed that taping and bracing conditions significantly reduced range of angular velocity in the sagittal plane by 8% and

						6%, respectively. Coronal plane angular velocity was reduced by 9% and 7%, respectively. The traditional reach distance scores were only reduced by 1% across the different reach directions with taping and bracing.
Burghart et al. (2017) [85]	Held to sternum	27 (12 male; 15 female)	Young healthy adults	Concurrent - gold standard	Force platform – DLS EO & EC firm; TS EO & EC firm; SLS EO firm	The correlation (Pearson product moment) between the force platform COP variables and the inertial sensor derived sway area, RMS and velocity ranged from -0.22 to 0.43, -0.2 to -0.42 and -0.18 to -0.49, respectively. Single-leg stance and TS EO consistently demonstrated the highest correlation with the force platform COP variable, across all three IMU derived variables.
Chiu et al. (2017) [44]	Anterior shank (half the distance from the lateral malleolus to fibular head)	15 (6 male; 9 female)	Young healthy adults	Discriminant - known groups and known condition	Chronic ankle instability vs healthy control - SLS, EO and EC	There was a statistically significant difference between the chronic ankle instability and healthy limb, across all stance conditions, EO and EC. Additionally, there was a significant difference between EO and EC conditions during SLS (both healthy and chronic instability limb).
Dabbs et al. (2018) [34]	Held to sternum	184 (sex not disclosed)	Healthy collegiate athletes	Concurrent - gold standard	Biodex balance system modified balance error scoring system – force platform	Pearson product moment correlation between the SWAY balance score and the Biodex balance system was as follows: Pearson's correlation = -0.32 (DLS, EO, firm); Pearson's correlation = -0.32 (TS, EO, firm); Pearson's correlation = -0.64 to -0.69 (SLS, EO, firm); Pearson's correlation = -0.70 (total score).
Doherty et al. (2017) [53]	Sacrum	15 case (11 male; 4 females); 15 control (11 male; 4 females)	Young healthy adults (control) & Concussion (case)	Discriminant - known groups	Concussion vs healthy - mBESS	The sacrum mounted IMU derived 95% sway ellipse was able to identify statistically significant differences between the healthy control and concussed groups during the DLS position only.

Furman et al. (2013) [14]	Anterior waist	27 healthy (8 male; 19 female); 10 acute concussion (9 male; 1 female); 33 subacute (18 male; 15 female)	Young healthy adult (control) & acute and subacute concussion (case)	Discriminant - known groups	Concussion vs healthy – BAM	The waist mounted inertial sensor instrumented BAM did not capture statistically significant differences between the healthy control and concussion groups, across all stance conditions. Conversely, the traditional BESS (tandem (firm/ foam), double leg (foam)) demonstrated statistically significant differences between groups.
Gera et al. (2018) [56]	5th Lumbar vertebra	38 case (25 male; 13 female); 81 control (44 male; 37 female)	Healthy collegiate athlete (control) & concussed collegiate athletes (case)	Discriminant - known groups	Concussion vs healthy - CTSIB	The concussed group demonstrated a significantly larger 95% sway area when compared to the control group for three of the four conditions: EO firm, EC firm and EC foam.
Heebner et al. (2015)[86]	5th Lumbar vertebra	10 (all male)	Young healthy adult	Concurrent - gold standard; Discriminant - known condition	Concurrent: Dynamic postural stability index – force platform. Construct: comparison of stance positions	Spearman's rank correlation between the force platform and accelerometer measures was statistically significant for the SLS EO (Spearman's correlation = 0.63) and EC (0.73) conditions (ML acceleration), DPSI-AP (Spearman's correlation = 0.68) and DPSI-ML (Spearman's correlation = 0.70) conditions (vertical acceleration) and DPSI-ML (Spearman's correlation = 0.59) (resultant acceleration). Correlation for all other variables was non-significant and ranged from Spearman's correlation = -0.54 to 0.56. All RMS acceleration variable was able to capture statistically significant differences between the static and dynamic tasks, similar to the force platform. Only ML RMS acceleration was able to capture differences between static conditions, despite the differences identified by the force platform.

Johnston et al. (2016) [49]	4th Lumbar vertebra	15 (7 male; 8 female)	Young healthy adults	Discriminant - known condition	Fatigue vs non-fatigue state YBT; different reach directions.	The random forest classification algorithm for the lumbar mounted IMU was capable of differentiating the three reach directions in the pre-fatigue baseline measures with an accuracy of 97.80%, sensitivity of 97.86 ± 0.89% and specificity of 98.90 ± 0.56%; Classification accuracy of fatigued vs non-fatigued balance performance ranged from 61.90%-71.43%, sensitivity of 61.90%-69.04% and specificity of 61.90%- 78.57% depending on which reach direction was chosen.
Johnston et al. (2017)[45]	4th Lumbar vertebra	15 (7 male; 8 female)	Young healthy adult	Discriminant - known condition	Construct: Fatigue vs non-fatigue state YBT. Concurrent: normalised reach distance	The IMU derived 95% sway ellipse demonstrated statistically significant differences (Cohens D = 0.59 to 1.03) in dynamic balance performance during all three YBT reach directions, similar to the traditional normalised reach distances. The IMU 95% sway ellipse detected statistically significant differences up to 20 minutes post-fatigue for all three reach directions (Cohens D = 0.77 to 0.96), despite non-significant differences demonstrated by the traditional reach distances (Cohens D = -0.06 to – 0.25).
Kim et al. (2018)[38]	4 cm below the knee joint line bilateral shanks; 4 cm above the lateral knee joint line; sacrum	5 healthy (2 male; 3 female); 5 injured (3 male; 2 female)	Healthy collegiate athletes (control); Knee injury collegiate athletes (case)	Concurrent - gold standard; Discriminant - known groups;	SLS EO, EO – 3D motion capture	The RMSE between the 3D motion capture ROLS and the IMU derived variables are as follows: thigh area = 0.23cm^2, thigh ML = 0.09cm, thigh AP = 0.11cm, shank area = 0.18 cm^2, ML = 0.10cm, AP = 0.11cm. Pearson correlation between the two systems ranged from Pearson's correlation = 0.82 to 0.93. There was no significant difference in sacral excursion area between healthy and injured limbs. There was a significant difference between ROLS index between healthy and injured limbs.

King et al. (2014) [15]	5th Lumbar vertebra	13 case (3 boys; 10 girls); 13 controls (3 boys; 13 girls.	Young health adolescents (control)& concussion (case)	Discriminant - known groups	Construct: Concussion vs healthy – BESS and mBESS Concurrent: BESS & mBESS	The RMS acceleration derived from the lumbar worn inertial demonstrated statistically significant differences between healthy controls and concussed individuals during the BESS (p = 0.04) and mBESS (P = 0.01), despite no significant differences identified using the traditional BESS and mBESS errors. The instrumented BESS and mBESS demonstrated an AUC of 0.70 (CI = 0.5 to 0.9) 0.81 (CI = 0.6 to 1.0), respectively. The traditional BESS and mBESS reach distances demonstrated an AUC of 0.63 (CI = 0.4 to 0.9) and 0.64 (CI = 0.4 – 0.9), respectively.
King et al. (2017) [54]	5th Lumbar vertebra	52 case (35 male; 17 female); 76 control (38 male; 38 female)	Young healthy college athletes (control) & concussed college athletes (case)	Discriminant - known groups	Construct: Concussion vs healthy – mBESS Concurrent: mBESS	All but one IMU derived variables (95% power frequency) demonstrated statistically significant differences between the healthy and concussed athletes. There was no significant difference between the healthy and injured athletes when considering the traditional mBESS errors. DLS RMS ML was the most valuable variable and had an AUC of 0.73 (CI = 0.6 to 0.8), while the traditional mBESS errors had an AUC of 0.61 (CI = 0.5 to 0.7).
Kosse et al. (2015) [40]	3rd Lumbar vertebra	22 (11 male; 11 female)	Young healthy adults	Convergence - non-gold standard (accelerometer)	iPod vs accelerometer - DLS with cognitive task, EO; TS with cognitive task, EO; DLS with cognitive task, EO; TS with cognitive task, EO.	The iPod derived variables demonstrated excellent agreement with the accelerometer variables across all stance positions (ICC2,1 = 0.84 to 0.99). The RMS variable demonstrated the lowest RPC% of the balance variables, ranging from 8.4% to 14%, while the median frequency ranged from 32% to 45%.
Linder et al. (2018)[17]	Sacrum	6762 (youth (age 5-13): males (n=360), females	Youth, adolescent and collegiate athletes	Discriminant - known groups	Age group and sex group comparison - BESS	Multivariate comparisons across age groups demonstrated a significant effect of age and sex across all cohorts (partial $eta^2 = 0.04$), with younger cohorts demonstrating poorer postural

		(n=246), high school (age 14-18): males (n=3743), females (n=1673), and college (age 19-23): males (n=497), females (n=243))				control than older cohorts. Females exhibited significantly better balance scores when compared to males in the youth and high school cohorts.
Martínez-Ramírez et al. (2010) [59]	3rd Lumbar vertebra	13 case (6 male; 7 female); 12 control (male 7; 5 female)	Young healthy adults (control); chronic ankle instability (case)	Discriminant - known groups	Construct: chronic ankle instability vs healthy control. Concurrent: normalised reach distance	The CAI group demonstrated a statistically significantly greater peak amplitude of the approximation of the yaw signal than the healthy control group. No significant differences were observed across the other reach directions or IMU derived balance variables. No statistically significant difference was found between the CAI and healthy group when considering the traditional SEBT normalised reach distances, for all three directions. Additionally, no significant differences were observed between groups when considering the equivalent force platform derived variables.
Mohamed et al. (2016) [35]	Held to sternum	30 (13 male; 17 female)	Young healthy adults	Concurrent - gold standard	Biodex balance system Modified balance error scoring system – force platform	Pearson product moment correlation between the SWAY balance score and the Biodex balance system was as follows: Pearson’s correlation = -0.42 (DLS, EO, firm); Pearson’s correlation = -0.35 (semi-TS, EO, firm); Pearson’s correlation = -0.61 (SLS, EO, firm).
Neville et al. (2015) [29]	5th Lumbar vertebra	10 (3 males; 7 females)	Young healthy adults	Concurrent - gold standard; Discriminant - known condition	TS EO/ EC firm; TS EO/ EC foam. DLS EO/ EC DLS EO/ EC foam – force platform/ 3D motion capture	The lumbar worn inertial sensor demonstrated a statistically significant correlation with the gold standard force platform (Pearson’s correlation = 0.79) and the 3D motion capture rigid-body movement of the L4-L5 segment (Pearson’s

						correlation = 0.88). There was a significant main effect for balance condition for the RMS and centroidal frequency variables. There was a significant three-way interaction between stance, surface and eye condition, indicating that the inertial sensor derived variables can differentiate between the different balance conditions. A significant difference was observed between the DLS, EO condition and all other conditions, except TS, EO foam pad.
Oliva Dominguez et al. (2013) [46]	Lumbar spine	10 (10 male; 10 female)	Young healthy adults	Discriminant - known condition	DLS EO/ EC DLS EO/ EC foam.	The lumbar worn IMU derived pitch and roll spectral densities (2.5Hz and 3.7Hz) were able to identify differences between all stance conditions except DLS EO vs DLS EC and DLS EO foam vs DLS EC foam.
Patterson et al. (2014) [42]	Held to sternum	21 (7 male; 14 female)	Young healthy adults	Convergence – clinical standard	BESS vs inertial sensor BESS	A strong inverse correlation was found between the traditional total BESS score and the inertial sensor derived SWAY balance score (Pearson's correlation = -0.77).
Patterson et al. (2014) [30]	Held to sternum	30 (13 male; 17 female)	Young healthy adults	Concurrent - gold standard	SLS EO – force platform	There was a significant correlation (Pearson's correlation = 0.63) between the inertial sensor derived SWAY balance score and the Biodex balance system stability score, with a mean difference of 0.03 ±0.7. Paired samples t- Test revealed no significant difference between the mean sway measures for the subjects when measured by each device (p = 0.818).
Rouis et al. (2014) [31]	5th Lumbar vertebra	15 (6 male; 9 female)	Young healthy adults	Concurrent - gold standard	Tasadna Yoga Pose – force platform	Correlation between the lumbar worn accelerometer and force platform ranged from 0.18 to 0.91 for during the EO stance, and -0.07 to 0.59 during the EC stance. RMS acceleration demonstrated the highest correlation in the EO stance, while acceleration range demonstrated the highest correlation during the EC stance. Median frequency and 95% power frequency demonstrated the lowest correlation for both

						stance conditions (Pearson's correlation = -0.07 to 0.39).
Salisbury et al. (2018) [41]	Head (glasses) & comparison sensor anterior waist	42 (26 male; 16 female)	Young healthy adults	Discriminant - known condition; Convergence - non-gold standard (waist accelerometer).	Balance Assessment Measure – glasses mounted accelerometer. DLS firm/foam EO and EC. TS, EO and EC.	The head (glasses) and waist worn tri-axial accelerometer NPL AP and total were significantly correlated (Spearman's correlation = 0.85 and 0.90 respectively). NPL AP/ total demonstrated significant differences between all stance conditions. DLS EO demonstrated the lowest AP and total NPL, while DLS EC foam and TS EC demonstrated the highest AP and total NPL.
Schelldorfer et al. (2015) [47]	Right thigh; 1st Lumbar vertebra; 1st thoracic vertebra.	57 lower back pain (26 male; 31 female); 22 healthy control (14 male; 8 female)	Young healthy adult (control) & Lower back pain (case)	Discriminant - known condition; Discriminant - known groups	DLS EO/ EC DLS EO/ EC foam	There was a statistically significant difference between the stance surface condition (firm vs foam) across all three sensor locations. There was a statistically significant difference between the healthy control and lower back pain groups when considering the ML sway position derived from the T1 and thigh mounted sensor. There was no significant group x stance condition interaction for either variables/ sensor location.
Seimetz et al. (2012) [32]	Sternum	5 (2 males; 3 females)	young healthy adults	Concurrent – gold standard. Discriminant - known condition	DLS EO/ EC DLS EO/ EC foam – force platform	No quantifiable statistical methods were implemented to investigate the concurrent and construct validity. Visual analysis suggests that the foam-based stance conditions resulted in a greater mean sway velocity, similar to the force platform derived mean COP velocity. Differences between eye conditions during the foam stance were not as clear using the accelerometer as they were using the force platform.
Senanayaka et al (2013) [87]	Shank; thigh	8 case (6 males; 2 female); 4 control (3 male; 1 female)	Young healthy adults (control) & anterior cruciate ligament knee	Discriminant – known groups	Anterior cruciate ligament injury vs healthy control - SLS, EO; EC	This study did not investigate the accuracy of the inertial sensor derived balance kinematics alone in classifying recovery of subjects after ACL. When considering the balance kinematics together with EMG signals, the classification

			injured adults (case)			accuracy was 94.44% for the eyes open task, and 95.83% for the eyes closed task.
Shah et al. (2016) [25]	Above the talocrural joint line; above superior midline of the patella; level of the subject's umbilicus.	50 (21 male; 27 female)	Young healthy adults	Discriminant - known condition	SLS EO, EC (firm/foam); TS EO; DLS EO, foam	Statistically significant 2-way interaction of stance condition by sensor mounting location. Post-hoc tests indicated higher acceleration magnitude for exercises of greater difficulty. The results revealed the knee as the location most sensitive for the detection of differences in acceleration between stance conditions.
Simon et al. (2017) [16]	Lumbar spine	38 (18 male; 20 female)	Healthy Collegiate athletes	Convergence – clinical standard	BESS vs inertial sensor instrumented BESS.	Pearson product moment correlation between the inertial sensor derived 95% sway ellipse and the traditional BESS errors was Pearson's correlation = 0.44 (single leg EC firm), Pearson's correlation = 0.63 (TS, EC, firm), Pearson's correlation =0.01 (SLS, EC, foam), Pearson's correlation =0.61 (TS, EC, foam) and Pearson's correlation =0.41 (BESS total). These results demonstrate that the TS conditions demonstrated the highest correlation.
Whitney et al. (2011)[33]	Pelvis (ASIS-PSIS)	81 (30 male; 51 female)	Young healthy adults	Concurrent – gold standard	Sensory organisation test – force platform	Linear regression analysis demonstrated a significant association between the force platform and accelerometer derived variables across all variables (both single trials and averages) except RMS and peak-to-peak when comparing the average of the force platform trial to the first accelerometer trial. R^2 values ranged from 0.18 to 0.92 (NPL), 0.15 to 0.64 (RMS) and 0.21 to 0.71 (P2P). For all variables, the R^2 increased with the increase in difficulty of the SOT task.
Wilkerson et al. (2018) [61]	Thoracic spine (between scapula)	45 (all male)	Cohort of collegiate American football players	Predictive	Prospective evaluation of injury risk - unilateral forefoot squat	Logistic regression analysis demonstrated that individuals who possessed poorer balance performance during the unilateral forefoot squat were at a 5.19 greater-odds of sustaining a sports related injury. When considering this together

						with the sports fitness index, the odds ratio increased to 8.64.
Yvon et al. (2015) [48]	Lateral left upper arm	50 (13 male; 37 female)	young healthy adults	Discriminant - known condition	TS EO/EC foam; TS EO/ EC foam, ear defenders. DLS EO/ EC firm; DLS EO/ EC firm ear defenders.	Comparison of balance conditions demonstrated that participants demonstrated significantly greater area of sway ellipse when participants had their EC and feet in TS position vs feet together; standing with their EC on foam vs on the floor; and standing with their EC on foam with feet in the tandem position vs on the floor with feet together. There was no significant difference when comparing balance performance in a normal and sound proofed room.

AP antero-posterior, *ML* medio-lateral, *TP* Transverse plane, *EO* eyes open, *EC* eyes closed, *BESS* Balance error scoring system, *mBESS* modified balance error scoring system, *BAM* Balance assessment measure, *SOT* Sensory organisation test, *RMS* root mean squared, *DLS* Double leg stance, *SLS* Single leg stance, *TS* Tandem stance, *ROLS* region of limb stability, *AUC* Area under the curve, *ICC* Intraclass correlation coefficient, *NPL* Normalised path length, *ND* Not disclosed

Table 6: description of the reliability studies included in the review

Study	Setup	Sample	Population	Reliability Type	Time-frame	Findings
Amick et al. (2015) [64]	Held to sternum	24 (15 male; 9 female)	Young healthy adults	inter and intra session	3 test sessions – 7 days apart	Repeated measures ANOVA revealed no significant mean differences between SWAY balance scores of the experimental trials (F(5,115) = 0.673; p = 0.65). **Intersession ICC** = 0.61 to 0.76 I**ntrasession ICC** = 0.47 to 0.78. Minimal difference to be considered real of 15.
Brown et al. (2007) [58]	Anterior tibial crest	20 case (10 male; 10 female); 20 control (10 male; 10 female)	Young healthy adults (control) & chronic ankle instability (case)	Intrasession	1 test session	Healthy ICC = 0.66 (CI = ND); CAI ICC = 0.98 (CI = ND).
Burghart et al. (2017) [27]	Held to sternum	27 (12 male; 15 female)	Young healthy adults	Intrasession	1 test session	DLS EO ICC = 0.41 (CI = 0.03-0.69); DLS EC ICC = 0.45 (CI = 0.08 - 0.71); TS EO ICC = 0.21 (CI = -0.20 - 0.55); TS EC ICC = 0.36 (CI = -0.06 - 0.67). SLS EO ICC = 36 (-0.06 – 0.67)
Dunn et al. (2017) [65]	Sternum	18 youth (all male); 69 high school (all male); 63 collegiate athletes (all male)	Youth, high school and collegiate athletes (all male)	Intrasession	3 (5 min rest period	**SWAY balance ICC youth:** total sway = 0.66 (CI 0.24 – 0.86); DLS EO firm = 0.10 (CI = 0.00 – 0.64); SLS EO firm left 0.47 (0.00 – 0.79); SLS EO firm right = 0.74 (CI = 0.44 – 0.89); TS EO firm right = 0.33 (CI = 0.00 – 0.73); TS EO firm left = 0.42 (0.00 – 0.77) **SWAY balance ICC high school:** total sway = 0.89 (CI 0.84 – 0.93); DLS EO firm = 0.48 (CI = 0.23 – 0.66); SLS EO firm left 0.72 (0.58 – 0.82); SLS EO firm right = 0.72 (CI = 0.59 – 0.82); TS EO firm right = 0.65 (CI = 0.47 – 0.77); TS EO firm left = 0.65 (0.45 – 0.76). **SWAY balance ICC collegiate:** total sway = 0.83 (CI 0.74 – 0.89); DLS EO firm = 0.20 (CI = 0.00 – 0.49); SLS EO firm left = 0.71 (0.56 – 0.82); SLS EO firm right = 0.75 (CI = 0.62 – 0.84); TS EO firm right = 0.45 (CI = 0.17 – 0.65); TS EO firm left = 0.45 (0.17 – 0.65).

Han et al. (2016) [62]	3rd/4th lumbar vertebra	30 (sex not disclosed)	Young healthy adults	Intersession	2 test sessions – 1 day apart	DLS – Gyroscope magnitude ICC = 0.7 (CI = ND) (EO), 0.6 (CI = ND) (EC); Acceleration magnitude ICC = 0.8 (CI = ND) (EO), 0.9 (CI = ND) (EC)
Heebner et al. (2015) [28]	5th lumbar vertebra	10 (all male)	Young healthy adult	Intersession	2 test sessions – 2 days apart	RMS acceleration AP ICC = 0.84 (CI = 0.33 – 0.96) RMS acceleration ML ICC = 0.84 (CI = 0.36 – 0.96) RMS acceleration vertical ICC = 0.89 (CI = 0.57 – 0.97) RMS acceleration resultant ICC = 0.92 (CI = 0.70 – 0.98)
Johnston et al. (2017) [45]	4th lumbar vertebra	15 (7 male; 8 female)	Young healthy adult	Intrasession	1 test session	95% sway ellipse anterior ICC = 0.76 (CI = 0.43 - 0.92) 95% sway ellipse posteromedial ICC = 0.89 (CI = 0.72 - 0.96) 95% sway ellipse posterolateral ICC = 0.92 (CI = 0.80 - 0.97)
Kosse et al. (2015) [40]	3rd lumbar vertebra	22 (11 male;11 female)	Young healthy adults	Intrasession	1 test session	DLS EO ICC**:** RMS AP = 0.86 (CI = 0.66 – 0.94**);** RMS ML **=** 0.74 (CI = 0.39 – 0.89); MPF AP = 0.39 (CI = -0.36 - 0.74); MPF ML = 0.86 (CI = 0.66 – 0.94); sway area = 0.57 (CI = -0.04 - 0.82). Semi-TS EO ICC: RMS AP = 0.80 (CI = 0.49 – 0.92); RMS ML = 0.74 (CI = 0.38 – 0.89); MPF AP = 0.77 (CI = -0.45 - 0.90); MPF ML = 0.92 (CI = 0.81 – 0.97); sway area = 0.55 (CI = 0.00 to 0.81).
Salisbury et al. (2018) [41]	Head (glasses) & comparison sensor anterior waist	42 (26 male; 16 female	Young healthy adults	Intrasession	1 test session	AP NPL ICC = 0.85 (CI = 0.81-0.88); Total NPL ICC = 0.87 (CI = 0.83 – 0.90).
Schelldorfer et al. (2015) [47]	Right thigh; 2nd sacral vertebra; 1st lumbar vertebra; 1st thoracic vertebra.	57 lower back pain (26 male; 31 female); 22 healthy control (14 male; 8 female)	Young healthy adult (control) & Lower back pain (case)	Intrasession	1 test session	The ICCs of the three repetitions ranged from 0.38 to 0.86 for asymptomatic controls and 0.43 to 0.83 for LBP patients, with higher values for sway velocity. No ICC scores were provided for specific groups or variables.
Simon et al. (2017) [16]	Lumbar spine – specific location not disclosed	38 (18 male; 20 female)	Healthy Collegiate athletes	inter and intra session	2 test sessions – 7 days apart	**Intrasession 95% sway ellipse:** DL EO firm = 0.9 (CI = 0.81 – 0.95); DL EO foam = 0.91 (CI = 0.83 – 0.95); SL EO firm = 0.80 (CI =0.61 - 0.90); SL EO foam = 0.83 (CI = 0.67 – 0.91); TS EO firm = 0.86 (CI = 0.74 – 0.93); TS foam = 0.67 (CI = 0.35 -0.82); all stances = 0.97 (CI = 0.94 - 0.98).

						Intersession 95% sway ellipse: DL EO firm = 0.83 (CI = 0.68 – 0.91); DL EO foam = 0.90 (CI = 0.83 – 0.91); SL EO firm = 0.72 (CI =0.46 - 0.85); SL EO foam = 0.84 (CI = 0.69 – 0.91); TS EO firm = 0.79 (CI = 0.60 – 0.89); TS foam = 0.74 (CI = 0.49 -0.86); all stances = 0.90 (CI = 0.82 - 0.95).
Whitney et al. (2011) [33]	Pelvis (ASIS-PSIS)	81 (30 male; 51 female)	Young healthy adults	Intrasession	1 test session	NPL ICC ranged from 0.67 to 0.80 across the 6 stance conditions. RMS ranged from 0.16 to 0.66 across the 6 stance conditions. P2P ranged from 0.47-0.79 across the stance conditions. The more challenging stance conditions generally demonstrated higher reliability.
Williams et al. (2016) [63]	Sacrum	30 (12 male; 18 female)	Young Healthy adults	Interrater and intrarater	Intrasession – 1 session Intersession: 2 test sessions - >1 day apart	**Intrasession reliability ICC:** PL ranged from 0.43 - 0.80 across the 8 stance conditions. Jerk ranged from 0.03 to 0.73 across the 8 stance conditions. RMS ranged from 0.26 to 0.77 across the 8 stance conditions. **Intersession reliability ICC:** PL ranged from 0.44 - 0.90 across the 8 stance conditions. Jerk ranged from 0.02 to 0.95 across the 8 stance conditions. RMS ranged from 0.3 to 0.71 across the 8 stance conditions.

AP antero-posterior, *ML* medio-lateral, *TP* Transverse plane, *EO* eyes open, *EC* eyes closed, *BESS* Balance error scoring system, *mBESS* modified balance error scoring system, *RMS* root mean squared, *DLS* Double leg stance, *SLS* Single leg stance, *TS* Tandem stance, *ICC* Intraclass correlation coefficient, *NPL* Normalised path length, *ND* not disclosed

Appendix 1: Paper quality assessment based on the Downs and Black criteria.

Article	1	2	3	4	5	6	7	8	9	10	Total
Abe et al. (2015) [36]	0	0	1	1	1	1	0	0	1	1	6
Alberts et al. (2015a) [26]	0	1	1	1	1	1	0	0	1	1	7
Alberts et al. (2015b) [12]	0	1	1	1	1	1	0	0	1	1	7
Amick et al. (2015) [64]	0	0	1	1	1	1	0	0	1	1	6
Anderson et al. (2017) [73]	1	0	1	1	1	1	0	0	1	1	7
Baracks et al. (2018) [55]	1	0	1	1	1	1	0	0	1	1	7
Berkner et al. (2017) [51]	1	0	1	1	1	1	0	0	1	1	7
Bernstein et al. (2018) [57]	1	0	0	1	1	1	0	0	1	1	6
Betkner et al. (2006) [37]	0	0	1	0	1	1	0	0	1	1	5
Bonnet et al. (2004) [43]	0	1	0	1	0	0	0	0	0	0	2
Borges et al. (2017) [52]	1	0	1	1	0	1	0	0	1	1	6
Brown et al. (2007) [58]	1	0	1	0	1	1	0	0	1	1	6
Brown et al. (2014) [13]	1	1	1	0	1	1	0	0	1	1	7
Budini et al. (2018) [50]	0	0	1	0	1	1	0	0	1	1	5
Burghart et al. (2017) [27]	1	0	1	1	1	0	0	0	1	1	6
Chiu et al. (2017) [44]	0	0	1	1	1	1	0	0	1	1	6
Dabbs et al. (2018) [34]	0	0	1	1	1	1	0	0	1	1	6
Doherty et al. (2017) [53]	1	1	1	1	1	1	0	0	1	1	8
Dunn et al. (2017) [65]	0	0	0	1	1	1	0	0	1	1	5
Furman et al. (2013) [14]	0	0	1	1	1	0	0	0	1	1	5
Gera et al. (2018) [56]	1	0	1	1	1	1	0	0	1	1	7
Han et al. (2016) [62]	0	0	1	0	1	1	0	0	1	1	5
Heebner et al. (2015) [28]	1	1	1	1	1	1	0	0	1	1	8
Johnston et al. (2016) [49]	1	1	1	1	1	1	0	0	1	1	8
Johnston et al. (2017) [45]	1	1	1	1	1	1	0	0	1	1	8
Kim et al. (2018) [38]	1	0	1	1	1	1	0	0	1	1	7
King et al. (2014) [15]	1	0	1	1	1	1	0	0	1	1	7
King et al. (2017) [54]	1	0	1	1	1	1	0	0	1	1	7
Kosse et al. (2015) [40]	1	0	1	1	1	1	0	0	1	1	7

Linder et al. (2018) [17]	1	0	1	1	1	1	0	0	1	1	7
Martínez-Ramírez et al. (2010) [59]	0	0	1	1	1	1	0	0	1	1	6
Mohamed et al. (2016) [35]	0	0	1	1	1	1	0	0	1	1	6
Neville et al. (2015) [29]	1	1	1	1	1	1	0	0	1	1	8
Oliva Dominguez et al. (2013) [46]	0	0	1	1	0	1	0	0	1	1	5
Patterson et al. (2014) [42]	0	0	1	1	1	0	0	0	1	1	5
Patterson et al. (2014) [30]	0	0	1	1	1	1	0	0	1	0	5
Rouis et al. (2014) [31]	0	0	0	1	1	1	0	0	1	1	5
Salisbury et al. (2018) [41]	1	0	1	1	1	1	0	0	1	1	7
Schelldorfer et al. (2015) [47]	1	0	1	0	0	1	0	0	1	1	5
Seimetz et al. (2012) [32]	0	0	0	0	0	0	0	0	1	0	1
Senanayake et al. (2013) [60]	0	0	0	1	0	0	0	0	1	1	3
Shah et al. (2016) [25]	0	0	1	1	1	1	0	0	1	1	6
Simon et al. (2017) [16]	1	0	1	1	1	1	0	0	1	1	7
Whitney et al. (2011) [33]	0	1	1	1	1	0	0	0	1	1	6
Wilkerson et al. (2018) [61]	0	0	0	1	0	0	0	0	1	1	3
Williams et al. (2016) [88]	0	0	1	1	1	1	0	0	1	1	6
Yvon et al. (2015) [48]	0	0	1	0	1	1	0	0	1	1	5

Items legend: 1. Is the hypothesis/aim/objective of the study clearly described?; 2. Are the main outcomes to be measured clearly described in the Introduction or Methods section?; 3. Are the characteristics of the patients included in the study clearly described?; 4. Are the main findings of the study clearly described?; 5. Does the study provide estimates of the random variability in the data for the main outcomes?; 6. Have actual probability or reliability values been reported (e.g. 0.035 rather than <0.05) for the main outcomes except where the probability value is less than 0.001?); 7. Were the subjects asked to participate in the study representative of the entire population from which they were recruited?; 8. Were those subjects who were prepared to participate representative of the entire population from which they were recruited?; 9. If any of the results of the study were based on "data dredging", was this made clear?; 10. Were the statistical tests used to assess the main outcomes appropriate