# Persistent Effects of Playing Football and Associated (Subconcussive) Head Trauma on Brain Structure and Function – A Systematic Review of the Literature

<table>
<thead>
<tr>
<th>Journal:</th>
<th><em>British Journal of Sports Medicine</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID:</td>
<td>bjsports-2016-096593.R1</td>
</tr>
<tr>
<td>Article Type:</td>
<td>Systematic Review</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>08-Sep-2016</td>
</tr>
</tbody>
</table>
| Complete List of Authors: | Tarnutzer, Alexander; University Hospital Zurich, Neurology; Schulthess Clinic, Swiss Concussion Center  
Straumann, Dominik; University Hospital Zurich, Neurology; Schulthess Clinic  
Brugger, Peter; University Hospital Zurich, Neurology  
Feddermann-Demont, Nina; University Hospital Zurich, Neurology; Schulthess Clinic |
| Keywords: | Football, Neurology, Sporting injuries, Trauma |
PERSISTENT EFFECTS OF PLAYING FOOTBALL AND ASSOCIATED (SUBCONCUSSIVE) HEAD TRAUMA ON BRAIN STRUCTURE AND FUNCTION – A SYSTEMATIC REVIEW OF THE LITERATURE

Tarnutzer AA M.D. (1,2), Straumann D M.D. (1,2), Brugger P PhD. (1), Feddermann-Demont N M.D. (1,2)

1) Department of Neurology, University Hospital Zurich and University of Zurich, Zurich, Switzerland
2) Swiss Concussion Center, Schulthess Clinic, Zurich, Switzerland

Short title: persistent effects of football on the brain

Corresponding author:
Alexander A. Tarnutzer, MD
Department of Neurology, University Hospital Zurich
Frauenklinikstr. 26, 8091 Zurich, Switzerland
Phone: 0041 44 255 11 11
Fax: 0041 44 255 43 80
Email: alexander.tarnutzer@access.uzh.ch

Statistics:
word count for the text (excluding legends, references & abstract): 4908
word count for the abstract: 250
character count for the title (including spaces): 153
number of figures: 2
number of tables: 5
supplementary files: 4
ABSTRACT

Aim/objective: There is ongoing controversy about persistent neurological deficits in active and former football (soccer) players. We reviewed the literature for associations between football activities (including heading/head-injuries) and decline in brain structure/function.

Design: Systematic literature review


Eligibility criteria for selecting studies: Original studies reporting on football-related persistent effects on brain structure/function. Results from neurocognitive testing, neuroimaging and electroencephalography were compared to controls and/or correlated with heading-frequency and/or head-injuries. Methodological quality was rated for risk-of-bias including appropriateness of controls, correction for multiple statistical testing, and assessment of heading-frequency and head-injuries.

Results: Thirty studies with 1691 players were included. Those 57% (8/14) of case-control studies reporting persistent neurocognitive impairment had higher odds for inappropriate control of type-1 errors (OR=17.35 [95%-CI=10.61-28.36]) and for inappropriate selection of controls (OR=1.72 [1.22-2.43]) than studies observing no impairment. Studies reporting a correlation between heading-frequency and neurocognitive deficits (6/17) had lower quality of heading-assessment (OR=14.20 [9.01-22.39]) than studies reporting no such correlation. In 7/13 studies (54%) the number of head-injuries correlated with the degree of neurocognitive impairment. Abnormalities on neuroimaging (6/8 studies) were associated with subclinical neurocognitive deficits in 3/4 studies.

Summary/conclusions: Various methodological shortcomings limit the evidence for persistent effects of football play on brain structure/function. Sources of bias include low-quality assessment of heading-frequency, inappropriate control for type-1 errors and inappropriate selection of controls. Combining neuroimaging-techniques with neurocognitive
testing in prospective studies seems most promising to further clarify on the impact of football on the brain.

**Bullet statements**

*What is already known?*

- Repetitive head injuries and heading the ball were suggested to be linked to persistent neurocognitive impairments and structural brain abnormalities in advanced neuroimaging in both professional and amateur football players.
- However, there is ongoing controversy to what extent these findings are real or rather result from methodological limitations.
- A systematic assessment of existing studies using prospectively defined criteria is therefore needed to improve our understanding of persistent effects of football on the brain and to better estimate the role of methodological limitations in previous studies.

*What are the new findings?*

- While the majority of studies, addressing the effect of football play and football-related injuries on neurocognitive functions, reported significant impairment in at least one domain, methodological shortcomings were found to be more frequent in studies with reportedly significant findings.
- Evidence for a correlation between heading frequency and neurocognitive deficits was weak and likely biased by inaccurate heading-frequency estimates.
- Although the rate of football-related head injuries was reportedly higher in women than men, females were under-represented in studies that reported neurocognitive impairment compared to those studies that did not identify deficits, which may be
related to the fact that none of these studies included (retired) female professional football players.

- Combining neuroimaging and neurocognitive testing in prospective longitudinal and cross-sectional studies in male and female players to link structural and functional deficits seems most promising to further clarify associations between football play and brain abnormalities.

Key words: football (soccer), heading, head injury, neurocognitive testing, systematic review
INTRODUCTION

Concussions (i.e., a subtype of mild traumatic brain injury (mTBI) without structural abnormalities on conventional CT or MRI)\(^1\) represent 1-5% of all football-(soccer-) related injuries.\(^2\) While most players return to play within 7-10 days, head-trauma related symptoms may last for weeks to months in 10-15% and even persist in selected cases.\(^6\)

Neurodegenerative disorders (such as Alzheimer disease) have been reported in retired professional football players and in athletes from other contact sports as rugby and American football.\(^7\)\(^8\) A postulated association between football play and chronic traumatic encephalopathy, however, remains controversial,\(^9\) and the effect of football-related concussions is not well understood.

Likewise, the impact of purposeful heading the ball to play and guide its direction – unique to football – on the brain has been debated.\(^10\)\(^11\) On average, players head the ball 1-16 times during a competitive football match,\(^12\)\(^16\) accumulating over a season to several hundred headings\(^17\)\(^19\) and to many thousand headings during a professional football career. This has raised concerns that heading may – similar to boxers receiving punches to the head - pose players at increased risk for “subconcussive” trauma,\(^20\)\(^24\) potentially resulting in neuronal damage similar to that in repetitive concussions but not accompanied by overt symptoms.\(^20\)\(^23\)

These considerations have led to uncertainty in football players and their (medical) attendants,\(^1\) albeit such a link is far from being established and the impact of parameters such as heading technique, player’s age and playing position remain unclear. Nonetheless, with raising concerns and facing a concussion litigation, the football federation of the USA issued in November 2015 a ban for heading in children aged 10 years or less and limited heading in children aged 11-13 years.\(^26\) Concerns that the maturing brain could be especially vulnerable to subconcussive head injury may have supported this decision.

Research interest in associations between concussion, heading and persistent changes of the human brain has grown substantially. At the end of the last century, a series of case-
control studies indicated persistent neurocognitive impairments in Dutch professional\textsuperscript{16,19} and amateur\textsuperscript{27} football players. These studies were the basis for further investigations addressing functional, structural and metabolic brain changes in football players. While some studies confirmed neurocognitive abnormalities compared to controls,\textsuperscript{28,29} others found no such evidence.\textsuperscript{14,15,30,31} Likewise, associations between neurocognitive deficits and heading-frequency were reported by some,\textsuperscript{16,17,19,28,29} but not by others.\textsuperscript{14,15,30-34}

In accordance with neuroimaging for TBI,\textsuperscript{35} different protocols were applied to study structural (diffusion-tensor imaging [DTI],\textsuperscript{36} voxel-based MR-morphometry [VBM]) and metabolic (functional MRI [fMRI], magnetic-resonance spectroscopy [MRS]) brain changes in football players and to correlate with neurocognitive function. In two small case-control studies, memory impairment was linked to cortical thinning in former professionals\textsuperscript{37} and to diffuse white-matter abnormalities in amateur players,\textsuperscript{17} while recently in a prospective case-control study over five years in professional players no such link could be drawn.\textsuperscript{38}

In summary, whether or not football play is linked to persistent changes in brain function/structure remains controversial. Against this background, we aimed to systematically review the literature on associations between football play and persistent changes in brain function/structure and the impact of heading-frequency and concussive head-injuries. Assessing study quality and identifying methodological limitations using standardized tools for reporting risk-of-bias was a special focus.\textsuperscript{10}
MATERIAL AND METHODS

Data sources and searches

A literature search (MEDLINE, Embase, PsycINFO, CINAHL, Cochrane CRCT, SportDiscus, Cochrane DSR=4) was performed (02/08/2016) to identify articles reporting on associations between football play and especially heading and football-related head-injuries and persistent structural/functional changes of the brain. The MEDLINE (OVID) search strategy was translated for each database, and is reported in supplement 1.

“Persistent” changes were defined as changes that were still recognizable >6 months after a potential impact or linked to exposure to football play for >6 months or a full season. We adhered to the time-frame usually applied for the persistent post-concussive syndrome, albeit no consensus-based definition of this term exists.

We also performed a manual search of reference lists from eligible articles. We did not seek to identify research abstracts from meeting proceedings or unpublished studies, nor non-English language studies. Retrospective or prospective studies with five or more participants were eligible. This review complies with PRISMA guidelines.

Study selection

We identified 2191 citations for screening and included 30 studies for quantitative synthesis (Figure 1) based on abstract and for selected studies full-text review by two experienced neurologists (NFD, AAT). Articles were selected using pre-determined criteria (supplement 1).

/* Figure 1 about here */
Data extraction and quality assessment

Reports on neurocognitive testing, neuroimaging, postural control and
electroencephalography (EEG) were considered. Data extraction was performed by AAT and
confirmed by NFD. When extracting data from selected studies, we assessed the type of
study, the type of diagnostic tests performed, the frequency of heading and head-injuries, and
the level of play, distinguishing between youth, high-school/college (including inter-scholar),
university, amateur, active professional and former professional players. In studies reporting
on neurocognitive testing, all tests applied were retrieved and assigned to the category that
best described the domain of neurocognitive function evaluated. Categories were: abstract
reasoning, attention, (verbal) creativity and divergent thinking, decision-making, executive
functions, intelligence, language and language-associated functions, memory/learning, mood,
motor skills, and visuospatial skills.

A standardized risk-of-bias assessment was performed using the Newcastle-Ottawa
Scale (NOS). Its use for non-randomized case-control studies and observational studies has
been promoted by the Cochrane Collaboration. The NOS requires rating the selection,
comparability and exposure/outcome for a total of nine items. Study quality was rated as
“good”, “fair” or “poor” (see and supplement 2). Studies rated as “poor” were excluded. The
NOS included an assessment of the response rate when asked to participate. Studies with a
low (<50%) participation rate, with participation rates differing >10% between football
players and controls or studies that did not report response rates were rated as high-risk for
selection bias. Whenever non-football playing controls were available (n=21 studies), their
suitability was rated. Only control groups that were age- and gender-matched and that
participated in non-contact sports with a comparable physical-activity profile (e.g.,
swimming, track, tennis) were considered “appropriate” or low-risk for bias. A distinct (i.e.,
lower/higher) physical-activity profile may introduce a bias and observed differences may be
attributed falsely to effects of football play. Controls falling short of these criteria were considered “inappropriate” or high-risk of bias.

Based on previously described methodological limitations, we further assessed the quality of included studies regarding assessment of heading-frequency and history of head-injuries and control for type-1 errors.\textsuperscript{10} Rutherford identified insufficient control for type-1 errors as a potential source for false-positive statistical differences.\textsuperscript{10} To follow-up on this limitation, we assessed methods for avoiding type-1 errors. Only studies that reported sufficient controlling for multiple testing (e.g., by applying Bonferroni correction) were considered “appropriate” or low-risk for type-1 errors, while studies falling short of these criteria were considered “inappropriate” or high-risk. Both heading-frequency and head-injuries may be overstated or understated by players, posing them at risk for recall-bias. Therefore, only studies that prospectively collected data on heading-frequency (e.g. by an independent observer) were considered low-risk of bias, while studies falling short of these requirements (e.g., relied on self-reported numbers, a heading-exposure-index\textsuperscript{44}) were considered high-risk. We did not require loss of consciousness for making the diagnosis of concussion and relied on the original study authors’ assessment.

**Data synthesis and analysis**

Excel 2011 (Microsoft Coorp., Redmont, USA) and Matlab 7.0 (The MathWorks, Nantuck, USA) were used for data analyses. Statistical analyses were performed using two-sample t-tests (with Bonferroni correction) and odds ratios (ORs) including 95%-confidence-intervals (CIs).
RESULTS

From the 32 studies included for qualitative synthesis (Figure 1), one was excluded because of “poor” quality on NOS and one was removed because of duplicity of data. Among the 30 studies included for quantitative synthesis (n=1691, 22.4% females), only six were prospective. Twenty-three studies (76.7%, n=1518) reported on results of neurocognitive testing, while data on neuroimaging were provided in eight studies (26.7%, n=143). Information on EEG (6.7%, n=106) was available from two; postural stability (3.3%, n=15) was provided in one study (Tables 1 and 2). Four studies reported on more than one modality (supplement 3). NOS ratings were “good” and “fair”, respectively, in 15 studies each (supplement 2). Key domains for neurocognitive testing of (sub)concussive brain injury (attention, executive functions, memory) were assessed by 18 of 23 studies, while in the remaining five studies one (n=3) or two (n=2) key domains were missing.

Case-control studies reporting on neurocognitive testing

Fourteen studies compared neurocognitive test results in football players (n=581) with those from controls (n=348). On average, 8.7±5.8 tests covering 4.4±1.9 categories were administered. Eight studies (57.1%) reported significantly lower results for the football players than for the controls in at least one test (2.8±2.7 tests, average±1SD). Most frequently, deficits of attention, executive function and memory were noted (Table 3 and Figure 2).

/* Figure 2 about here */

Studies reporting neurocognitive deficits had a higher rate of inappropriate control of type-1 errors (OR=17.35 [95%-CI=10.61-28.36]) and higher odds for inappropriate controls (OR=1.72 [1.22-2.43]) than studies not reporting any differences. The fraction of female players was about the same for both groups (OR=0.82 [0.47-1.47]). The fraction of younger
players (youth/high-school/college) compared to more elderly players was larger in studies with negative findings than in those with significant deficits (OR=1.92 [1.38-2.68]) (Table 3).

**Impact of heading on neurocognitive functions**

A potential link between heading-exposure and performance on neurocognitive testing was analyzed in 17 studies (n=1173, 26.4% females). On average 8.5±5.1 neurocognitive tests covering 4.4±1.8 categories were obtained (Table 3). Two thirds of the studies did not find any relation between heading-frequency and neurocognitive test-performance, while six studies (35%) reported a correlation in 2.0±1.1 tests. Deficits of attention, memory and intelligence were most frequent (Table 4). The quality of heading-frequency assessment was lower for those studies reporting a link than for those studies without such a link (OR=14.2 [9.0-22.4]). The rate of inappropriate control of type-1 errors was similar in studies confirming or discarding such a link (OR=0.74 [0.53-1.04]). The fraction of players with more extensive exposure (amateurs, university, (former) professionals) was significantly higher amongst those studies reporting neurocognitive deficits than for those studies that did not observe deficits (OR=2.15 [1.63-2.85]).

**Impact of head-injuries on neurocognitive functions**

A potential association between previous head-injuries and neurocognitive deficits was investigated in 13 studies (n=1103, 26.6% females) (Table 3). On average, 11.3±4.9 neurocognitive tests covering 5.2±1.4 categories were obtained. Seven studies (54%) reported a correlation, with abnormalities noted in 2.4±1.8 tests. Deficits of visuospatial functions, decision-making, attention and executive function were most frequent (Table 4). The rate of inappropriate control of type-1 errors was lower in studies with positive findings compared to studies with negative findings (OR=0.20 [0.15-0.26]).
Data on previous head-injuries were available in ten studies with average numbers of concussions ranging between 1.0 and 2.1, with the most recent event between 6-8 months and several years ago. The assessment of previous head-injuries was based on players’ reports in all but two studies. Only football-related concussions were considered in 5/7 studies with positive findings and in 3/6 studies with negative findings, while the remaining five studies included other, non-football-related concussions as well or did not further specify.

**Neuroimaging studies**

We identified eight studies (n=143, 15.4% females) using imaging modalities focusing on brain structure (conventional MRI, VBM, DTI) or brain metabolism (fMRI, MRS). DTI (2 studies, n=49), conventional MRI (2 studies, n=44) and VBM (2 studies, n=25) were most frequently applied (Table 5). All studies used a case-control design with selection of controls rated as “appropriate” in six (75%). Most players were professionals (active=56; former=26) or amateurs (n=37). Only two studies were prospective. In one prospective study no conventional-MRI changes could be depicted in professional players (observation period=5 years). Prospectively observing female high-school players over one season using fMRI, significant reductions in frontotemporal cerebrovascular reactivity persisting up to 4-5 months after the season had ended were reported, resembling the pattern described in mTBI. Retrospectively, in former professionals, VBM demonstrated cortical thinning in the right inferolateral-parietal, temporal and occipital cortex and MRS showed higher choline and myo-inositol levels in the posterior-cingulate gyrus. In professional players, DTI indicated widespread white-matter abnormalities (albeit no changes in fractional anisotropy), but conventional MRI did not demonstrate changes related to the years of football-participation. In college-football players, VBM showed decreased grey-matter density and volume within the anterior-temporal cortex.
Four studies linked neuroimaging with neurocognitive data.\textsuperscript{17, 37, 38, 51} Cortical thinning was associated with worse performance on 1/6 tests (Rey-Osterrieth complex-figure long-delay recall),\textsuperscript{37} Glutathion-levels were linked to inferior results in 1/4 tests (trail-making-test B)\textsuperscript{51} and lower levels of fractional-anisotropy in parieto-occipital areas were associated with 1/6 tests (poorer memory).\textsuperscript{17} In the only prospective study, neither changes in neurocognitive performance nor in conventional MRI could be depicted over an observation period of five years.\textsuperscript{38}

A link between heading-exposure and structural/metabolic neuroimaging-changes was investigated in five studies,\textsuperscript{17, 37, 44, 51, 54} Lifetime estimates of heading-numbers were inversely correlated with cortical thickness in the right parietal/occipital lobes\textsuperscript{37} and with myo-inositol and glutathione levels.\textsuperscript{51} Fractional-anisotropy levels in temporo-parietal white-matter were inversely correlated with the annual number of headings.\textsuperscript{17} A high cumulative head-acceleration exposure was linked to more profound reductions in cerebrovascular reactivity, outlasting the end of the season by 4-5 months before returning to baseline by month eight.\textsuperscript{54} No correlation between career heading-exposure and abnormalities on conventional MRI were reported in another study.\textsuperscript{44} The potential impact of remote head-injuries on brain structure was examined in two studies, both demonstrating no association.\textsuperscript{17, 44}

**EEG-studies**

Two studies used EEG in active (n=69)\textsuperscript{55} and former (n=37)\textsuperscript{56} professional male players. In both studies standard EEG-recordings were examined by a clinical neurophysiologist and EEGs were classified as “normal”, “slightly abnormal” or “abnormal” based on background-activity and alpha-activity. EEG-ratings in the players were compared to those in age-matched men of “various occupations”. With information on matched physical activities lacking in the controls, their quality was rated “inappropriate”. The rate of EEGs considered normal was lower in active and former players compared to controls. Amongst the
active players, all abnormal EEGs were observed in players who considered themselves as non-headers. Among former players, there were no EEG-differences between headers and non-headers.

**Postural stability**

One study (n=15) reported on balance, using the balance-error-scoring-system (BESS). This study described no significant differences between players and controls.
DISCUSSION

With the recently issued ban for heading in child-football players in the US, the ongoing debate about potential persistent effects of football and football-related (subconcussive) trauma on brain function received increased attention and caused uncertainty amongst football players, medical staff and media. Given the worldwide popularity of football, football-related health-issues may have far-reaching implications that have to be balanced and compared to benefits due to regular activity. This emphasizes the need to intensify hypothesis-driven research and the study of associations between football-play and persistent structural/functional changes of the brain.

Under-representation of female players

For most aspects evaluated, female players were in a minority, consistent with reportedly lower numbers of active female football players. While no conclusions could be drawn on football-related changes in neuroimaging and EEG, women were under-represented in studies that reported neurocognitive impairment compared to those not observing such deficits. This observation was unexpected since the rate of football-related head-injuries was reportedly higher in women. Of note, none of the studies reported on (former) professional female players. Also, for studies reporting on neurocognitive testing, female players were over-represented in lower levels of play (youth, high-school/college) compared to higher levels (university, amateur, (former) professional) (OR=28.57 [19.25-42.41]). These observations suggest that cumulative exposure to football play or cumulative intensity has been lower in female players, not reaching levels that may be necessary to result in brain abnormalities. Future studies should pay special attention on functional/structural brain changes in female players with more extensive football exposure.
Neurocognitive testing in (former) football players

Applied in 77% of studies, neurocognitive testing remains the most common approach to investigate potential associations between football-play and changes in brain function. Most studies dealt with effects of heading (74%) and head-injuries (57%). Over 60 different neurocognitive tests were used, most of them only in few studies. With all three key domains (attention, executive function and memory) assessed by 78% of studies, risk for false-negative results due to inappropriate selection of neurocognitive test domains seems low. Even for the most frequently used tests in those domains considered most important in patients with (sub)concussive brain injury (supplement 4), the fraction of abnormal test results was low (0.29±0.18). This suggests that reported neurocognitive impairments were rather subtle and their detection may have depended on study-specific parameters as age, gender, level of play and selection of controls. Also, among all tests administered in a given study, those with abnormal outcome were infrequent (fraction=0.21±0.27). In 19/23 studies more than one neurocognitive test was applied to evaluate a single category (e.g. TMT-B and Stroop for executive functions). Noteworthy, in 42% of these studies discrepant test results in a given category were noted. This affected 37% (17/46) of all categories in studies that received multiple testing. These results suggest that changes are subtle and may get identified only by some tests. Moreover, these inconsistencies underline the importance of standardized neurocognitive testing in football.

Persistent neurocognitive changes in (former) players compared to controls

Fifty-seven percent of case-control studies reported persistent associations between football and neurocognitive impairment focusing on attention, executive function and memory. These categories are primarily mediated by the frontal and temporal lobes and are typically involved in mTBI. Based on the quality-assessment performed, several confounders must be considered to put the significance of these observations in context.
Probably of the most far-reaching implication is the finding that the rate of appropriate control for type-1 errors was smaller among studies with abnormal test results (OR=17.35 [10.61-28.36]). These studies thus bear an increased risk for false-positive test results. Choosing the right controls is essential. Including control subjects without matching the profile of physical activity, might point to global effects of physical-activity rather than to football-related changes.\textsuperscript{14} While in our review controls were judged as “appropriate” in 67% of case-control studies, the odds for inappropriate controls were higher for studies reporting neurocognitive deficits (OR=1.72 [1.22-2.43]). This indicates that inappropriate selection of control subjects may represent a serious source of bias. As a consequence, caution is warranted when interpreting impairment in neurocognitive testing based on the existing literature. Noteworthy, with the fraction of younger players being larger in studies with negative findings than in those with significant deficits, this might suggest that the exposure duration was simply not long or intense enough to cause a significant effect, further limiting conclusions.

\textit{No clear evidence for heading-related persistent impairment of neurocognitive function}

Six of 17 studies that correlated heading-frequency with neurocognitive deficits reported a link (mostly for attention, executive functions, and memory), but these studies also contained more methodological limitations than those reporting no link. Most importantly, the assessment-quality of heading-frequency was lower in studies with positive findings (OR=14.2 [9.0-22.4]). Self-reported heading-frequencies tend to be higher than those obtained by more reliable approaches,\textsuperscript{68} indicating potential risk of reporting bias. This emphasizes the need for prospective observer-based assessments of heading-frequency in future studies. Furthermore, studies so far remained incomplete in providing an accurate estimate of heading-exposure, since heading during practice sessions was not considered and other variables such as heading-technique and ball-properties and -velocity were not available. Studies focusing on former players with heading-exposure decades ago\textsuperscript{45 56 57} may also be biased by differences in

\url{https://mc.manuscriptcentral.com/bjsm}
the properties of the ball – heavy leather balls, in common use until the mid-1970s and even heavier on wet undergrounds, should not be compared with the more lightweight balls used thereafter.

The ratio of more senior to more junior players was higher in studies reporting significant impact of heading than in those with negative findings (OR=2.15 [1.63-2.85]). For more senior players the lifetime heading-number is higher and the duration of exposure is longer. Current studies therefore cannot exclude that in more senior players neurocognitive deficits may eventually arise due to a decreased cerebral reserve-capacity in view of accumulated (sub)clinical head-trauma. Of note, in retired professional UK-football players no signs for accelerated cognitive decline were found. Furthermore, female players were under-represented in studies that noted a link between heading-frequency and neurocognitive impairments. A lower heading-frequency in female players and the skewed distribution of female players towards lower levels of play may explain this seemingly “protective effect” of female gender on heading-related neurocognitive impairments.

Based on our review, no firm conclusions on an association between heading-frequency and accelerated neurocognitive decline can be drawn. Methodological limitations identified more often in studies with positive findings emphasize caution in linking heading to persistent neurocognitive deficits. This extends conclusions of a previous review on acute and persistent effects of concussions and heading in football. Whether or not significant differences in neurocognition arise with increasing heading-exposure, awaits further clarification, especially in former professional players with extensive exposure and better control for head-injury.

Repetitive concussions may be linked to persistent cognitive impairment

A negative impact of (repetitive) head-injuries to brain function has been reported for other contact sports as American football, rugby or boxing. For football (soccer)
play, Barnes and colleagues estimated a 50%-risk that a professional male player will suffer a concussion within a 10-year period, while the corresponding figure for female players was 22%. Neuropathological changes associated with mTBI are axonal injuries with a focus on the orbito-frontal and temporal-polar zones. Cognitive functions affected most are delayed memory, executive functions, language and attention. Based on our review, 54% of studies addressing the influence of head-injuries on neurocognitive test-performance found a link in one or more categories. Seventy-one percent of studies with positive findings restricted their analysis to football-related concussions, which indicates a low-risk that football-unrelated concussions biased the results. Furthermore, the rate of studies with inappropriate control of type-1 errors was even lower amongst studies with positive findings compared to those with negative findings.

Distinguishing between effects secondary to heading and head-trauma, however, may be difficult or even impossible for several reasons: First, up to 50% of concussions are not reported by players or team physicians. Second, 89% of studies controlled only for recent (3-6 months) head-trauma or did not take head-trauma into account at all. Most studies relied on self-reporting of head-injuries, introducing risk of recall bias. Even more importantly, lack of control for head-injuries bears the risk that effects of heading and head-trauma are mixed as players with higher heading-frequencies tend to experience head-injuries more frequently. Third, definitions for football-related head-injuries have been applied differently, potentially resulting in inconsistencies between studies. Assuming that effects of heading and head-injury may have been intermingled, one would expect overestimating the link between heading-frequency and neurocognitive deficits. In fact, the opposite was true; refuting the assumption that head-injury related effects have significantly biased results in the assessment of heading-related persistent effects. In summary, the link between head-injuries and persistent neurocognitive impairment was moderate only and its impact may have been over-estimated due to several methodological shortcomings.
Structural and metabolic changes on neuroimaging

Neuroimaging in football players was driven by the hypothesis that repetitive subconcussive head-trauma result in structural/metabolic changes similar to those known from mTBI. Along with the anterior-posterior gradient in brain vulnerability, anterior regions may also be linked to executive and neurocognitive deficits in mTBI.

With a limited number of study subjects (n=143) and studies (n=8), different neuroimaging-modalities and levels of play, all but two studies reported significant brain changes in football players. These changes were localized preferentially to frontal and anterior-temporal regions. Structural abnormalities located in parieto-occipital areas, i.e., in areas opposite the presumed point of heading impact, may be explained in analogy to the principle of contre-coup injury.

In three studies with 63 players, changes depicted on neuroimaging could be linked to subclinical neurocognitive deficits, suggesting that these changes may be functionally relevant. While these studies used advanced neuroimaging, no changes could be observed on conventional MRI in the only prospective study (five-year observation period). Whereas there is extensive experience in the interpretation of standard MR-sequences, changes in advanced imaging as DTI, VBM and MR-spectroscopy are much more difficult to put into clinical context due to their relatively recent use.

A correlation between heading-exposure and persistent changes on neuroimaging was observed in four out of five studies, albeit reversible within eight months after season end in the only prospective study. While these findings suggest a possible link between heading-frequency and neuro-degeneration, the heading-exposure assessment was low-quality in four out of five studies, posing them at increased risk for recall bias. Correction for multiple statistical testing was reported by Svaldi and in one study by Koerte, but not in another, indicating possible increased risk of false-positive correlations. In the study by
Lipton, a link was observed only for players with more than 885-1500 headings per year, suggesting that heading below this threshold may be safe.\textsuperscript{17} Overall, taking the methodological limitations and the small sample sizes into consideration, support for a link between heading-frequency and persistent structural brain changes seems weak. Our review did not find any evidence for an association between (repetitive) concussions and structural brain changes on neuroimaging, albeit small sample sizes limit conclusions.\textsuperscript{17,44}

**Low-quality of EEG-based studies**

Both studies included reported higher incidences of EEG-abnormalities in male professional players than in controls.\textsuperscript{55,56} With only 106 players analyzed, conclusions on persistent effects of football on EEG can be considered preliminary only. Caution is warranted as significant limitations were identified: control groups were rated “inappropriate” and the authors did not control for remote head-injury. The clinical implications of the higher incidence of EEG-abnormalities remain unclear, especially since no information on the location of EEG-abnormalities was provided. Both studies did not provide any evidence for a link between heading and EEG-changes as in the group of active football players all abnormal EEGs were observed in players who considered themselves as non-headers\textsuperscript{55}. In the group of former players no EEG-differences were detected comparing headers and non-headers.\textsuperscript{56} Nonetheless, the authors concluded that the reason for the EEG-abnormalities was most likely repeated minor head-trauma. Considering the limitations listed above, evidence in support of this conclusion is unconvincing.

**Limitations**

Studies varied in the criteria for reporting previous concussions and we were lacking details to validate the ratings. Self-reporting was standard in most studies, posing them at
high-risk for recall bias. Therefore the impact of previous concussions on brain structure/function may have been over- or underestimated in our review. The impact of heading depends on many parameters. However, only scarce information about player’s position, circumstances of heading (purposeful, passively hit) and heading-skills could be retrieved. This limits the assessment of a possible association between heading and structural/functional brain changes. Lack of systematically controlling for timing and time lags between exposure and testing potentially weakens reported associations. Likewise failure to assess and/or correct for potential confounders as concussions unrelated to football, different physical-activity profiles, medical conditions (hypertension, overweight, diabetes) and life-style (e.g., alcohol consumption, smoking) may have biased associations between (former) football play and structural/functional brain changes.

A pooled analysis of individual study data (i.e., a meta-analysis) was not possible due to the heterogeneity in study design, data collection and data analysis. This may have resulted in the incorrect detection or masking of high-quality studies due to a higher number of low-quality studies.

We did not identify any studies that compared the diagnostic accuracy of different neurocognitive-testing procedures that assessed the same neurocognitive domains. Moreover, single studies obtaining more than one neurocognitive test for a specific domain often demonstrated discrepancies. This limits any recommendation on specific neurocognitive tests and emphasizes the need for prospective, controlled studies comparing the diagnostic accuracy of neurocognitive tests.

CONCLUSIONS

There is weak to non-existent evidence from the medical literature for football-related persistent functional and structural brain deficits and a putative role of (repetitive) head trauma in the development of neurocognitive impairment. Comparison of included studies
was limited by various methodological shortcomings. Case-control studies reporting neurocognitive deficits in football players significantly more often included inappropriate controls and control of type-1 errors than studies reporting no deficits. Furthermore, no clear link between heading-frequency and neurocognitive deficits could be established and low-quality assessment of heading-frequency was identified as the major confounder. Of special interest were studies that combined different modalities: while in four out of five neuroimaging studies structural and metabolic deficits could be correlated with heading-exposure, the clinical and preventive implications of these findings remain inconclusive, as most studies used a low-quality assessment of heading-frequency. In three out of four small case-control studies a link between neuroimaging abnormalities and subclinical neurocognitive deficits could be established, suggesting that these morphological and metabolic changes might be functionally relevant.

Further studies combining functional and structural modalities in larger numbers of football players with long-lasting football-exposure appear most promising to shed more light on a potential link between football play and brain structure/function. Such studies should include neurocognitive testing of attention, executive functions and memory as well of objective tests of cervical, vestibular and ocular motor function. They should also be prospective to appropriately control for confounders as history of head-injuries, heading-frequency and medical conditions. Further validation and head-to-head comparison is required to provide the basis of more standardized testing batteries to improve the quality of studies and to allow for better comparability between studies. A longitudinal and cross-sectional study design will help to determine whether identified subclinical structural and functional abnormalities eventually progress to clinically relevant, symptomatic deficits or rather resolve again, especially after finishing exposure to football play.
### TABLES
<table>
<thead>
<tr>
<th>Citation, year</th>
<th>Study design</th>
<th>Level of play</th>
<th>control group</th>
<th>Sample size (test / control group) (n), gender</th>
<th>Observation / exposure duration to competitive football play (mean±1SD)</th>
<th>Focus of study</th>
<th>NOS rating (&quot;good&quot;, “fair”, or “poor”)</th>
<th>controls appropriate / appropriate control of type-1 errors / response rate (%)</th>
<th>High-quality assessment of heading-frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abreu et al. 1990</td>
<td>Retrospective, case-control</td>
<td>College</td>
<td>College tennis players</td>
<td>31 / 31, m</td>
<td>≥2 seasons (college level)</td>
<td>NCP football players vs. controls</td>
<td>Fair</td>
<td>Yes / no / 62/220 (28%)</td>
<td>N/A</td>
</tr>
<tr>
<td>Adams et al. 2007</td>
<td>Retrospective, case-control</td>
<td>College</td>
<td>non-football playing College students</td>
<td>10 / 10, m</td>
<td>≥8 years</td>
<td>Grey-matter density and volume (VBM)</td>
<td>Fair</td>
<td>No / yes / NR</td>
<td>N/A</td>
</tr>
<tr>
<td>Downs et al. 2002</td>
<td>Retrospective, case-control</td>
<td>College / pro</td>
<td>College swimmers</td>
<td>32 / 29, f=m</td>
<td>≥4 years (collegiate/national level)</td>
<td>NCP football players vs. controls, impact of heading</td>
<td>Fair</td>
<td>No / no / NR</td>
<td>No (Heading-exposure-index)</td>
</tr>
<tr>
<td>Ellemberger et al. 2007</td>
<td>Retrospective, cross-sectional</td>
<td>University</td>
<td>N/A</td>
<td>22 / N/A, f</td>
<td>6-8 months since concussion, 15 years of play</td>
<td>Impact of concussion on NCP</td>
<td>Fair</td>
<td>N/A / yes / NR</td>
<td>N/A</td>
</tr>
<tr>
<td>Forbes et al. 2014</td>
<td>Prospective, cohort</td>
<td>Inter-scholastic</td>
<td>N/A</td>
<td>210 / N/A, f</td>
<td>27±21 months since (last) concussion, 9 years of play</td>
<td>Impact of concussion on NCP</td>
<td>Good</td>
<td>NA / unclear / NR</td>
<td>Yes</td>
</tr>
<tr>
<td>Gusiewicz et al. 2002</td>
<td>Retrospective, case-control</td>
<td>Collegiate</td>
<td>non-football contact sports, (n=96), no sports controls (n=53)</td>
<td>91 / 149, f=m</td>
<td>≥5 years</td>
<td>NCP football players vs. controls, impact of concussion</td>
<td>Fair</td>
<td>No / yes / NR</td>
<td>N/A</td>
</tr>
<tr>
<td>Janda et al. 2002</td>
<td>Prospective, cohort</td>
<td>Youth</td>
<td>N/A</td>
<td>57 / NA, f=m</td>
<td>9 months (prospective)</td>
<td>Impact of heading on NCP</td>
<td>Good</td>
<td>N/A / unclear / NR</td>
<td>Yes</td>
</tr>
<tr>
<td>Jordan et al. 1996</td>
<td>Retrospective, case-control</td>
<td>Active pro</td>
<td>elite track athletes</td>
<td>20 / 20, m</td>
<td>17.7±3.1 years</td>
<td>Structural abnormalities (MRI)</td>
<td>Good</td>
<td>Yes / no / 20/25 (80%)</td>
<td>N/A</td>
</tr>
<tr>
<td>Kemp et al. 2016</td>
<td>Prospective, case-control</td>
<td>Active pro</td>
<td>Various non-contact sports athletes from local university</td>
<td>32 / 33, m</td>
<td>≥5 years of competitive football play</td>
<td>NCP football players vs. controls, structural abnormalities (MRI)</td>
<td>Fair</td>
<td>Yes / yes / NR</td>
<td>N/A</td>
</tr>
<tr>
<td>Koerte et al. 2015</td>
<td>Retrospective, case-control</td>
<td>Former pro</td>
<td>Former non-contact pro athletes</td>
<td>11 / 14, m</td>
<td>Trained since childhood, ≥1 year as pro</td>
<td>Brain biochemistry (MRS), correlation with NCP</td>
<td>Good</td>
<td>Yes / no / NR</td>
<td>No (self-reported)</td>
</tr>
<tr>
<td>Koerte et al. 2017</td>
<td>Retrospective, case-control</td>
<td>Former pro</td>
<td>Former non-contact pro athletes</td>
<td>15 / 15, m</td>
<td>Trained since childhood, 4.7±3.6 years as pro</td>
<td>Grey-matter thickness (VBM), correlation with NCP</td>
<td>Good</td>
<td>Yes / no (NC1), yes (imaging) / NR</td>
<td>No (self-reported)</td>
</tr>
<tr>
<td>Kontos et al. 2011</td>
<td>Prospective, cross-sectional</td>
<td>Youth</td>
<td>N/A</td>
<td>63 / NA, f=m</td>
<td>Btw. 6.7±3.0 and 8.5±4.6 years</td>
<td>Impact of heading on NCP</td>
<td>Good</td>
<td>N/A / yes / NR</td>
<td>Yes</td>
</tr>
<tr>
<td>Lipton et al. 2013</td>
<td>Retrospective, cross-sectional</td>
<td>Amateur</td>
<td>N/A</td>
<td>37 / N/A, f=m</td>
<td>&gt;5 years</td>
<td>White matter microstructure (DTI), correlation with NCP</td>
<td>Good</td>
<td>N/A / yes / NR</td>
<td>No (self-reported)</td>
</tr>
<tr>
<td>Matser et al. 1996</td>
<td>Prospective, case-control</td>
<td>Active pro</td>
<td>elite non-contact sports athletes</td>
<td>53 / 27, m</td>
<td>≥4 years amateur (median=12y) and ≥1 year pro (median=5y)</td>
<td>NCP, impact of heading/head injuries</td>
<td>Good</td>
<td>Yes / no / NR</td>
<td>No (self-reported)</td>
</tr>
<tr>
<td>Matser et al. 1997</td>
<td>Retrospective, case-control</td>
<td>Amateur</td>
<td>amateur athletes (swimming/track)</td>
<td>33 / 27, m</td>
<td>≥5 years (avg=17y)</td>
<td>NCP football players vs. controls, impact of head injuries</td>
<td>Good</td>
<td>Yes / no / 33/33 (100%)</td>
<td>N/A</td>
</tr>
<tr>
<td>Matser et al. 2001</td>
<td>Retrospective, case-control</td>
<td>Active pro</td>
<td>N/A</td>
<td>84 / N/A, m</td>
<td>≥1 year pro play (median=4y)</td>
<td>Impact of heading and head injuries on NCP</td>
<td>Good</td>
<td>N/A / no / NR</td>
<td>No (self-reported)</td>
</tr>
<tr>
<td>Rutherford et al. 2005</td>
<td>Retrospective, case-control</td>
<td>University</td>
<td>Rugby players (n=17), non-contact sports (n=24)</td>
<td>22 / 41, m</td>
<td>≥5 years</td>
<td>NCP football players vs. controls, impact of heading/head injuries</td>
<td>Good</td>
<td>Yes / yes / 109/156 (70%)</td>
<td>Yes</td>
</tr>
<tr>
<td>Rutherford et al. 2009</td>
<td>Retrospective, case-control</td>
<td>University</td>
<td>Rugby players (n=30), non-contact sports (n=22)</td>
<td>25 / 52, m</td>
<td>≥5 years (avg=6.1±0.1y)</td>
<td>NCP football players vs. controls, impact of heading/head injuries</td>
<td>Good</td>
<td>Yes / yes / 165/254 (65%)</td>
<td>Yes</td>
</tr>
<tr>
<td>Salinas et al. 2009</td>
<td>Prospective, cross-sectional</td>
<td>Youth</td>
<td>N/A</td>
<td>49 / N/A, f=m</td>
<td>3.1±1.9 years</td>
<td>Impact of heading on NCP</td>
<td>Fair</td>
<td>N/A / unclear / NR</td>
<td>Yes</td>
</tr>
<tr>
<td>Stephens et al. 2010</td>
<td>Retrospective, case-control</td>
<td>High school</td>
<td>Rugby players (n=27), non-contact sports (n=16)</td>
<td>54 / 43, m</td>
<td>3.4±1.0 years</td>
<td>NCP football players vs. controls, impact of heading/head injuries</td>
<td>Fair</td>
<td>Yes / yes / 193 / 337 (57%)</td>
<td>Yes</td>
</tr>
<tr>
<td>Study</td>
<td>Design</td>
<td>Cohort Type</td>
<td>N/A</td>
<td>Age at highest league</td>
<td>Impact of heading/head injuries on NCP</td>
<td>Abbreviations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------------------------</td>
<td>-----------------</td>
<td>--------------</td>
<td>-----</td>
<td>-----------------------</td>
<td>---------------------------------------</td>
<td>---------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Straume et al. 2005</td>
<td>Retrospective,</td>
<td>Active pro</td>
<td>N/A</td>
<td>271 / NA, m</td>
<td>Impact of concussion on NCP</td>
<td>Good / N/A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>cross-sectional</td>
<td></td>
<td></td>
<td>Btw: 2.8±2.4 and 3.3±2.4 years at highest league</td>
<td>Good / N/A / no / 271/300 (90%)</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Straume et al. 2009</td>
<td>Prospective,</td>
<td>Active pro</td>
<td>N/A</td>
<td>144 / N/A, m</td>
<td>Impact of concussion on NCP</td>
<td>Good / N/A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>cohort</td>
<td></td>
<td></td>
<td>1 year (i.e., 1 season)</td>
<td>Good / N/A / yes / NR</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Svaldi et al. 2016</td>
<td>Prospective,</td>
<td>High school</td>
<td></td>
<td>14 / 12, f</td>
<td>Cerebrovascular reactivity related to head acceleration (heading)</td>
<td>Good / N/A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>case-control</td>
<td></td>
<td></td>
<td>Btw. 2.8±2.4 and 3.3±2.4 years at highest league</td>
<td>No / yes / NR</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>High school</td>
<td></td>
<td>Btw. 2.8±2.4 and 3.3±2.4 years at highest league</td>
<td>Yes (head acceleration event monitoring)</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tysvaer and Storli 1989</td>
<td>Retrospective,</td>
<td>Active pro</td>
<td>men with various occupation</td>
<td>69 / 69, m</td>
<td>EEG disturbances</td>
<td>Fair / no / unclear / NR</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>case-control</td>
<td></td>
<td></td>
<td>Btw. 2.8±2.4 and 3.3±2.4 years at highest league</td>
<td>No (self-reported)</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tysvaer et al. 1989</td>
<td>Retrospective,</td>
<td>Former pro</td>
<td>men from ‘different occupational groups’</td>
<td>37 / 37, m</td>
<td>EEG disturbances</td>
<td>Fair / no / unclear / NR</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>case-control</td>
<td></td>
<td></td>
<td>Btw. 2.8±2.4 and 3.3±2.4 years at highest league</td>
<td>No (self-reported)</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tysvaer et al. 1989</td>
<td>Retrospective,</td>
<td>Former pro</td>
<td>Men from ‘different occupational groups’</td>
<td>37 / 20, m</td>
<td>EEG disturbances</td>
<td>Fair / no / unclear / NR</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>case-control</td>
<td></td>
<td></td>
<td>Btw. 2.8±2.4 and 3.3±2.4 years at highest league</td>
<td>No (self-reported)</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vann Jones et al. 2014</td>
<td>Retrospective,</td>
<td>Former pro</td>
<td>UK-based MCI prevalence study</td>
<td>92 / N/A, m</td>
<td>EEG disturbances</td>
<td>Fair / no / unclear / NR</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>case-control</td>
<td></td>
<td></td>
<td>Btw. 2.8±2.4 and 3.3±2.4 years at highest league</td>
<td>No (heading-exposure-index)</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Witol et al. 2002</td>
<td>Retrospective,</td>
<td>Interscholastic, amateur/pro</td>
<td>non-soccer playing controls</td>
<td>60 / 12, m</td>
<td>EEG disturbances</td>
<td>Fair / no / unclear / NR</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>case-control</td>
<td></td>
<td></td>
<td>Btw. 2.8±2.4 and 3.3±2.4 years at highest league</td>
<td>No (heading-exposure-index)</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zhang et al. 2013</td>
<td>Retrospective,</td>
<td>High school</td>
<td></td>
<td>12 / 12, f</td>
<td>EEG disturbances</td>
<td>Fair / no / unclear / NR</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>case-control</td>
<td></td>
<td></td>
<td>Btw. 2.8±2.4 and 3.3±2.4 years at highest league</td>
<td>No (heading-exposure-index)</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations: avg=average; btw=between; CT=computed tomography; DTI=diffusion tensor imaging; EEG=electroencephalography; MRI=magnetic resonance imaging; MRS=magnetic resonance spectroscopy; N/A=not available; NCP=neurocognitive performance; NOS=Newcastle-Ottawa Scale, NR=not reported; pro=professional; VBM=voxel-based morphometry.

* These studies did not provide years of play, but games played. Assuming a total of 30-60 games per season, in all cases more than one season of professional play can be assumed.
Table 2: summary information about included studies (n=30)

<table>
<thead>
<tr>
<th></th>
<th>Studies (n, [%])</th>
<th>Female football players (n [%])</th>
<th>Male football players (n [%])</th>
<th>All football players (n [%])</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Control of type-1 errors</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inappropriate / unclear</td>
<td>18 [60.0]</td>
<td>275 [72.6]</td>
<td>898 [68.4]</td>
<td>1173 [69.4]</td>
</tr>
<tr>
<td>Total</td>
<td>30 [100.0]</td>
<td>379 [100.0]</td>
<td>1312 [100.0]</td>
<td>1691 [100.0]</td>
</tr>
<tr>
<td><strong>Selection of controls</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA</td>
<td>10 [33.3]</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Total</td>
<td>30 [100.0]</td>
<td>108 [100.0]</td>
<td>413 [100.0]</td>
<td>521 [100.0]</td>
</tr>
<tr>
<td><strong>Response rate</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High (&gt;50%)</td>
<td>7 [23.3]</td>
<td>0 [0.0]</td>
<td>462 [35.2]</td>
<td>462 [27.3]</td>
</tr>
<tr>
<td>Low (≤50%)</td>
<td>3 [10.0]</td>
<td>0 [0.0]</td>
<td>135 [10.3]</td>
<td>135 [8.0]</td>
</tr>
<tr>
<td>Not reported</td>
<td>20 [66.7]</td>
<td>379 [100.0]</td>
<td>715 [54.5]</td>
<td>1094 [64.7]</td>
</tr>
<tr>
<td>Total</td>
<td>30 [100.0]</td>
<td>379 [100.0]</td>
<td>1312 [100.0]</td>
<td>1691 [100.0]</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Football players</td>
<td>N/A</td>
<td>379 [22.4]</td>
<td>1312 [77.6]</td>
<td>1691 [100.0]</td>
</tr>
<tr>
<td>Control subjects</td>
<td>N/A</td>
<td>108 [20.7]</td>
<td>413 [79.3]</td>
<td>521 [100.0]</td>
</tr>
<tr>
<td><strong>Category</strong>†</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NCT</td>
<td>23</td>
<td>365</td>
<td>1153</td>
<td>1751</td>
</tr>
<tr>
<td>Case-control studies</td>
<td>14 [47.1]</td>
<td>56</td>
<td>525</td>
<td>512</td>
</tr>
<tr>
<td>Impact of heading</td>
<td>17 [29.7]</td>
<td>310</td>
<td>863</td>
<td>1459</td>
</tr>
<tr>
<td>Impact of head-injuries</td>
<td>13 [29.7]</td>
<td>188</td>
<td>810</td>
<td>980</td>
</tr>
<tr>
<td>Neuroimaging</td>
<td>13 [44.8]</td>
<td>22</td>
<td>121</td>
<td>138</td>
</tr>
<tr>
<td>EEG</td>
<td>2 [6.7]</td>
<td>0</td>
<td>106</td>
<td>106</td>
</tr>
<tr>
<td>Postural control</td>
<td>1 [3.3]</td>
<td>0</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td><strong>Level of play</strong>‡</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Youth</td>
<td>3 [10.0]</td>
<td>69 [18.2]</td>
<td>100 [7.6]</td>
<td>169 [10.0]</td>
</tr>
<tr>
<td>Professional</td>
<td>8 [26.7]</td>
<td>0 [0.0]</td>
<td>677 [51.6]</td>
<td>677 [40.0]</td>
</tr>
<tr>
<td>Total</td>
<td>30 [100.0]</td>
<td>379 [100.0]</td>
<td>1312 [100.0]</td>
<td>1691 [100.0]</td>
</tr>
</tbody>
</table>

Abbreviations: EEG=electroencephalography; N/A=not available; NCT=neurocognitive testing.

* Fraction of subjects that agreed to participate after being invited to participate in the study. This includes both football players and control subjects.

† Some studies provided testing for more than one modality (e.g. neurocognitive testing and neuroimaging or neuroimaging and balance testing), resulting in a total study number larger than 31. Within neurocognitive testing, also some studies provided case-control data as well as a correlation analysis for e.g. heading-frequency and neurocognitive deficits in the football players.

‡ Level of play as reported in the original studies. Age-range for youth football players was 10-13 and 13-18. The category “High-school / college” includes the “interscholastic” as well.
### Table 3: Studies reporting on neurocognitive testing (NCT) in football players

<table>
<thead>
<tr>
<th>Case-control studies</th>
<th>Studies addressing the impact of heading-frequency</th>
<th>Studies addressing the impact of head-injuries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Studies with significant differences in NCT†</td>
<td>Studies without sign. differences in NCT</td>
<td>All studies</td>
</tr>
<tr>
<td>Studies without sign. impact on NCT</td>
<td>Studies with significant impact on NCT</td>
<td>All studies</td>
</tr>
<tr>
<td>OR=17.3 [10.61-28.36]†</td>
<td>OR=0.20 [0.15-0.26] †</td>
<td>OR=0.74 [0.53-1.04] †</td>
</tr>
</tbody>
</table>

#### Control type-1 errors (n studies / n subjects [%])

- **Appropriate**: 1 / 22 [8.3], 4 / 194 [61.2], 5 / 216 [37.2], 2 / 59 [20.5], 3 / 142 [16.0], 5 / 201 [17.1], 4 / 242 [58.7], 3 / 153 [22.1], 7 / 395 [35.8]
- **Inappropriate/unclear**: 7 / 242 [17.1], 9 / 365 [24.8], 4 / 229 [79.5], 8 / 743 [84.0], 12 / 972 [82.9]
- **OR=0.20 [0.15-0.26] †**

#### Selection of controls (n studies / n subjects [%])

- **Appropriate**: 5 / 155 [58.2], 5 / 225 [71.0], 10 / 380 [65.4]
- **Inappropriate/unclear**: 3 / 109 [41.3], 1 / 92 [29.0], 4 / 201 [34.6]
- **OR=1.72 [1.22-2.43] †**

#### Response rate (n studies)

- **High (>50%)**: 3, 2, 5
- **Low (≤50%)**: 0, 2, 2
- **Not reported**: 7, 6, 11
- **OR=1.20 [0.87-1.63] ††**

#### Gender (n subjects [%])

- **Females**: 23 / 59 [20.5], 5 / 225 [16.0], 10/ 380 [65.4]
- **Males**: 241 / 59 [91.3], 284 / 225 [84.0], 525 / 380 [72.4]
- **Total**: 264 / 59 [100.0], 317 / 225 [100.0], 581 / 380 [100.0]
- **OR=0.20 [0.15-0.26] ††**

#### Sample size (mean±SD)

- **Football players**: 33±17, 53±32, 42±26
- **Controls**: 21±7, 36±34, 27±22
- **OR=0.74 [0.53-1.04] ††**

#### NCTs (mean±SD)

- **4.6±2.0, 4.0±2.0, 4.4±1.9**
- **OR=0.82 [0.47-1.44] ††**

#### Level of play (n studies / n subjects [%])

- **High school / college**: 3 / 104 [39.4], 5 / 176 [55.5], 6 / 280 [48.2]
- **University**: 1 / 22 [8.3], 1 / 25 [25.0], 2 / 47 [8.1]
- **Amateur**: 1 / 33 [12.5], 0 / 0 [0.0], 1 / 33 [3.7]
- **Professional**: 1 / 53 [20.1], 1 / 24 [7.6], 2 / 77 [13.5]
- **Former professional**: 2 / 52 [19.7], 1 / 92 [9.2], 3 / 144 [24.8]
- **All**: 8 / 264 [100.0], 6 / 317 [100.0], 14 / 581 [100.0]
- **OR=1.92 [1.38-2.68] ††**

---

https://mc.manuscriptcentral.com/bjsm
Abbreviations: CI=confidence interval; NCT=neurocognitive testing; NR=not reported; OR=odds ratio including 95% confidence intervals.

* Studies may have provided data both for neurocognition in football in general and specifically related to heading-frequency and / or the number of head-injuries.

† In these studies statistically significant (p<0.05) differences (between cases and controls or related to heading-frequency or the number of head-injuries) in at least one neuropsychological test were reported.

‡ Odds ratio (OR; including 95% confidence intervals in brackets) comparing fractions of a given condition (e.g., appropriateness of type-1 errors or gender) between studies with or without significant differences in NCT.

§ Controls were considered appropriate if the following criteria were met: 1) age- and gender-matched, 2) physical activities of comparable intensity but without body contact, i.e., swimming, running or cycling, 3) no additional (recreational) exposure to football or other contact sports.

¶ This includes both football players and control subjects.

|| Two-sample t-tests using Bonferroni correction for multiple testing.

** Odds ratio for high-quality heading-assessment (prospective recorded or combined prospective and retrospective approach) vs. low-quality heading-assessment (self-reported heading rates or calculation of heading-exposure-risk based on player’s position).

†† Odds ratio between studies with or without significant differences in NCT comparing younger (youth, high school, college, interscholastic) and more elderly (university, amateur, professional, former professional) football players.
Table 4: Distribution of NCT categories (in alphabetical order) and percentage of abnormal tests*

<table>
<thead>
<tr>
<th>Category</th>
<th>Case-control studies (n=14)</th>
<th>Studies reporting on heading-frequency (n=17)</th>
<th>Studies reporting on the number of head-injuries (n=13)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Studies with sign. effect / all studies reporting abnormal NCT in at least one category (n=8) [%]</td>
<td>Studies with sign. effect / all studies reporting abnormal NCT in at least one category (n=6) [%]</td>
<td>Studies with sign. effect / all studies reporting abnormal NCT in at least one category (n=7) [%]</td>
</tr>
<tr>
<td>Abstract reasoning</td>
<td>0/3 [0]</td>
<td>0/2 [0]</td>
<td>0/3 [0]</td>
</tr>
<tr>
<td>Creativity</td>
<td>0/2 [0]</td>
<td>0/2 [0]</td>
<td>0/2 [0]</td>
</tr>
<tr>
<td>Decision making</td>
<td>0/0 [0]</td>
<td>0/1 [0]</td>
<td>2/2 [100]</td>
</tr>
<tr>
<td>Intelligence</td>
<td>1/6 [17]</td>
<td>1/6 [17]</td>
<td>0/3 [0]</td>
</tr>
<tr>
<td>Language</td>
<td>0/4 [0]</td>
<td>0/6 [0]</td>
<td>0/6 [0]</td>
</tr>
<tr>
<td>Motor skills</td>
<td>1/3 [33]</td>
<td>1/3 [33]</td>
<td>0/3 [0]</td>
</tr>
<tr>
<td>Visuospatial functions</td>
<td>2/7 [29]</td>
<td>2/5 [40]</td>
<td>3/8 [38]</td>
</tr>
</tbody>
</table>

Abbreviations: NCT = neurocognitive testing; sign.=significant.

* Studies may have provided data both for neurocognition in football in general and specifically related to heading-frequency and / or number of head-injuries. Categories with ≥50% of studies with abnormal NCT results are bold.
Table 5: Persistent effects of football on the brain: neuroimaging studies (n=8)

<table>
<thead>
<tr>
<th>Effect confirmed (n, [%])</th>
<th>No effect (n, [%])</th>
<th>All (n, [%])</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 (75.0)</td>
<td>2 (25.0)</td>
<td>8 (100.0)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Selection of controls (n studies / n subjects [%]) *</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Appropriate</td>
<td>4 / 52 (83.9)</td>
</tr>
<tr>
<td>Inappropriate / unclear</td>
<td>1 / 10 (10.1)</td>
</tr>
<tr>
<td>N/A</td>
<td>1 / N/A</td>
</tr>
<tr>
<td>Total</td>
<td>6 / 62 (100.0)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gender football players (n subjects [%])</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Females</td>
<td>22 (22.2)</td>
</tr>
<tr>
<td>Males</td>
<td>77 (77.8)</td>
</tr>
<tr>
<td>Total</td>
<td>99 (100.0)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gender controls (n subjects [%])</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Females</td>
<td>12 (19.4)</td>
</tr>
<tr>
<td>Males</td>
<td>50 (80.6)</td>
</tr>
<tr>
<td>Total</td>
<td>62 (100.0)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample size (mean±1SD)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Football players</td>
<td>16.5±10.2</td>
</tr>
<tr>
<td>Controls</td>
<td>12.4±2.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type of imaging (n studies / n football players [%])</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventional MRI</td>
<td>0 / 0 (0.0)</td>
</tr>
<tr>
<td>Diffusion-tensor imaging (DTI)</td>
<td>2 / 49 (49.5)</td>
</tr>
<tr>
<td>Voxel-based MR-morphometry (VBM)</td>
<td>2 / 25 (25.3)</td>
</tr>
<tr>
<td>MR-spectroscopy (MRS)</td>
<td>1 / 11 (11.1)</td>
</tr>
<tr>
<td>Functional MRI (fMRI)</td>
<td>1 / 14 (14.1)</td>
</tr>
<tr>
<td>Total</td>
<td>6 / 99 (100.0)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level of play (n studies / n football players [%])</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Youth</td>
<td>0 / 0 (0.0)</td>
</tr>
<tr>
<td>High school / college</td>
<td>2 / 24 (24.2)</td>
</tr>
<tr>
<td>University</td>
<td>0 / 0 (0.0)</td>
</tr>
<tr>
<td>Amateur</td>
<td>1 / 37 (37.4)</td>
</tr>
<tr>
<td>Professional</td>
<td>1 / 12 (12.1)</td>
</tr>
<tr>
<td>Former professional</td>
<td>2 / 26 (26.3)</td>
</tr>
<tr>
<td>Total</td>
<td>6 / 99 (100.0)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Correlating neuroimaging with NCT (n studies)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Structural and NCT changes correlate</td>
<td>38 (38)</td>
</tr>
<tr>
<td>No correlation</td>
<td>0</td>
</tr>
<tr>
<td>N/A</td>
<td>58 (58)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Correlating neuroimaging with heading-frequency (n studies)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Changes on neuroimaging and heading-frequency correlate</td>
<td>44 (44)</td>
</tr>
<tr>
<td>No correlation</td>
<td>0</td>
</tr>
<tr>
<td>N/A</td>
<td>54 (54)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Correlating neuroimaging with the number of head-injuries (n studies)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Changes on neuroimaging and number of head-injuries correlate</td>
<td>0</td>
</tr>
<tr>
<td>No correlation</td>
<td>1</td>
</tr>
<tr>
<td>N/A</td>
<td>54 (54)</td>
</tr>
</tbody>
</table>

Abbreviations: NCT=neurocognitive testing; N/A=not available

* Controls were considered appropriate if the following criteria were met: 1) age- and gender-matched, 2) physical activities of comparable intensity but without body contact, i.e., swimming, running or cycling, 3) no additional (recreational) exposure to football or other contact sports.
FIGURE LEGENDS

Figure 1
Flow chart depicting the selection process of identified articles. * One study was excluded because of duplicity of data, another study was excluded because of a “poor” risk-of-bias rating (Newcastle-Ottawa Scale).

Figure 2
Spider plots illustrating in how many studies the different neurocognitive categories were evaluated (in red) and how often an abnormal test result was retrieved in this category (in blue). The number of studies is provided along the intersection of the web. Separate plots are provided for all case-control studies (panel A), for all studies investigating the impact of heading on neurocognitive tests (panel B) and for all studies reporting on the impact of head injuries on cognition (panel C). Irrespective of the study categorization, attention, executive functions, memory, and visuospatial functions were the cognitive domains most frequently tested and also most frequently impaired.
REQUIRED STATEMENTS

Acknowledgements: We thank Dr. K. Alix Hayden (Libraries & Cultural Resources University of Calgary, Calgary, Canada) for optimizing the search strategy and performing the literature search.

Funding sources for this study: none.

Conflict of interest:
Dr. Tarnutzer reports no conflict of interest.
Prof. Brugger reports no conflict of interest.
Prof. Straumann reports no conflict of interest.
Dr. Feddermann-Demont reports no conflict of interest.
REFERENCES


https://mc.manuscriptcentral.com/bjsm
Figure 2
203x61mm (300 x 300 DPI)
Supplementary file 1: Detailed search and selection strategy

The search strategy was designed by Dr. Tarnutzer and Dr. K. A. Hayden. We searched seven databases (MEDLINE, Embase, PsycINFO, CINAHL, Cochrane CRCT, SportDiscus, Cochrane DSR = 4) for English-language articles. For MEDLINE, the following search strategy was used. This search strategy was then translated for each database.

Database(s): Epub Ahead of Print, In-Process & Other Non-Indexed Citations, Ovid MEDLINE(R) Daily and Ovid MEDLINE(R) 1946 to Present

Search Strategy:
# Searches
1 soccer.mp.
2 football.mp.
3 exp soccer/
4 or/1-3
5 concuss*.mp.
6 traumatic brain injur*.mp.
7 heading.mp.
8 head acceleration.mp.
9 sub-concuss*.mp.
10 subconcuss*.mp.
11 head impact*.mp.
12 exp Post-Concussion Syndrome/ or exp Brain Concussion/
13 or/5-12
14 4 and 13
15 limit 14 to English language

Our search was updated through August 2, 2016. We also performed a manual search of reference lists from eligible articles. Research abstracts from meeting proceedings or unpublished studies or non-English language studies were not considered. Where appropriate, we attempted to contact authors regarding study details. There was no review protocol.

All identified articles were subject to title and abstract screening by two independent reviewers (AAT, NFD). Articles were selected using pre-determined criteria. Reviewers excluded papers that lacked original player data, were not related to football (soccer), did not
Tarnutzer et al. 2016: Persistent effects of playing football and associated (subconcussive) head trauma on brain structure and function – a systematic review of the literature

report any symptom data about heading or head trauma, did not report any symptom data about persistent effects (i.e., duration of exposure or observation period < 6 months for football play or after the impact), reported on less than five cases. Note that studies reporting on persistent effects of heading and head-related injuries in children (i.e., patients aged less than 18 years) were also potentially eligible.

Full-text screening was applied to all abstracts considered eligible by at least one reviewer (i.e., labeled “yes” or “maybe” in the abstract review). The two independent reviewers identified whether full-text manuscripts were eligible and provided a reason for exclusion. Discrepancies in selection status and reasons for exclusion were settled between the two reviewers by discussion and adjunction of a third reviewer if needed.

Information abstracted from each article included study type, number of former/active football players with long-term deficits related to heading / head injuries, inclusion criteria, and study site. For each study, we extracted which diagnostic tests were used in the evaluation of the former/active football players and in what fraction of players this test was positive and negative.

Data were handled in EndNote X 7.5 (Thomson Reuters, NY) and Microsoft Excel 2011 (Redmond, WA).

Search Results

Our search identified 2191 unique citations, of which 2079 (94.9%) were excluded at the abstract level (see flow diagram in main manuscript for details in the search strategy). We did not require concordance on the reasons for abstract exclusion, but, of concordant codings (80.1%, n=1670), exclusions were for the following reasons: 39% lacked original patient data; 28% were not related to football; 8% did not record any symptom data about persistent effects; 2% were abstracts only; 1% were not in English; 1% did not indicate that any
Tarnutzer et al. 2016: Persistent effects of playing football and associated (subconcussive) head trauma on brain structure and function – a systematic review of the literature

symptom data about heading or head trauma had been recorded; and 1% reported on less than five cases.

We sought to examine 112 full articles. After initial screening, there were 11 disagreements on study inclusion (Cohen’s kappa 0.77), and 4 disagreements on the reason for exclusion. These were settled by adjudication and discussion between the two reviewers. After final full-text review, 80 articles were excluded. The most common reason for exclusion was failure to provide data related to football (43%, n=34); other reasons for exclusion were as follows: paper was not about persistent effects (36%, n=29); were abstracts only (13%, n=10); was not about head / head-related injuries (5%, n=4); contained less than 5 cases (3%, n=2); and lacked original patient data (1%, n=1). A review of the bibliography of the selected 30 articles identified another 2 articles that met the inclusion criteria. In total 32 articles were eligible. Eligible articles represented 1.5% of the total (n=2191) articles. One article presented preliminary data from another selected study and was therefore excluded as well. One article was excluded due to “poor” quality in the Newcastle-Ottawa Scale (see supplement 3). The finally selected 30 articles for quantitative synthesis involved 1691 football players (active or former).

References


Tarnutzer et al. 2016: Persistent effects of playing football and associated (subconcussive) head trauma on brain structure and function – a systematic review of the literature

Supplementary file 2: Risk-of-bias assessment using the Newcastle-Ottawa Scale (NOS)

We applied the Newcastle-Ottawa Scale to perform a standardized risk-of-bias assessment.\(^1\) The use of the NOS for non-randomized case-control studies and observational studies has been promoted by the Cochrane Collaboration.\(^2\) The NOS addresses the following three aspects: 1) the selection of the study group(s), 2) the comparability of the study group(s) and 3) the ascertainment of either the exposure (for case-control studies and for cross-sectional studies) or the outcome (for cohort studies) of interest. While the NOS provides a frame for a standardized risk of bias assessment, it requires completion of several items according to the specifications of the studies applied to. For our review, the assessment of history of non-football associated head injuries and the response rate (cases and if available, also the controls) were incorporated in the NOS. Full rating instructions are attached at the end of this supplement.

While the NOS was designed for non-randomized case-control studies and for observational cohort studies, no rating of cross-sectional studies has been originally considered. Others have therefore previously modified the NOS such as it can be applied to cross-sectional studies as well.\(^3\)

Overall, the NOS rates nine items and for each item that meets criteria for low risk of bias, a star is given. The quality of individual studies was rated according to the thresholds proposed by McPheeters and co-workers in appendix G as either “good” (≥7 stars), “fair” (3-6 stars) or “poor” (≤2 stars).\(^4\) However, to give equal weight to each of the nine items, we used the total number of stars as cut-off, not the subscore in each of the three domains as originally proposed by McPheeters and colleagues. Studies rated as “poor” were excluded from the systematic review.

From the 31 studies selected after full-text search and after removal of one study due to duplicity of data,\(^5\) 15 studies were rated as “good” quality and 15 as “fair”, while 1 study
Tarnutzer et al. 2016: Persistent effects of playing football and associated (subconcussive) head trauma on brain structure and function – a systematic review of the literature

was judged as “poor”. Detailed ratings can be found in tables 1, 2 and 3 for case-control studies, cohort studies and cross-sectional studies.
Tarnutzer et al. 2016: Persistent effects of playing football and associated (subconcussive) head trauma on brain structure and function – a systematic review of the literature

Table 1: NOS ratings for case-control studies

<table>
<thead>
<tr>
<th>Quality assessment criteria</th>
<th>Koerte 2015&lt;sup&gt;6&lt;/sup&gt;</th>
<th>Koerte 2015&lt;sup&gt;7&lt;/sup&gt;</th>
<th>Vann Jones 2014&lt;sup&gt;8&lt;/sup&gt;</th>
<th>Zhang 2013&lt;sup&gt;9&lt;/sup&gt;</th>
<th>Koerte 2012&lt;sup&gt;10&lt;/sup&gt;</th>
<th>Stephens 2010&lt;sup&gt;11&lt;/sup&gt;</th>
<th>Rutherford 2009&lt;sup&gt;12&lt;/sup&gt;</th>
<th>Rutherford 2005&lt;sup&gt;13&lt;/sup&gt;</th>
<th>Winol 2002&lt;sup&gt;14&lt;/sup&gt;</th>
<th>Downs 2002&lt;sup&gt;15&lt;/sup&gt;</th>
<th>Guskiewics 2002&lt;sup&gt;16&lt;/sup&gt;</th>
<th>Matser 1999&lt;sup&gt;17&lt;/sup&gt;</th>
<th>Matser 1998&lt;sup&gt;18&lt;/sup&gt;</th>
<th>Jordan 1996&lt;sup&gt;19&lt;/sup&gt;</th>
<th>Tysvaer 1991&lt;sup&gt;20&lt;/sup&gt;</th>
<th>Sortland 1989&lt;sup&gt;21&lt;/sup&gt;</th>
<th>Tysvaer 1989&lt;sup&gt;22&lt;/sup&gt;</th>
<th>Tysvaer 1989&lt;sup&gt;23&lt;/sup&gt;</th>
<th>Abreau 1990&lt;sup&gt;24&lt;/sup&gt;</th>
<th>Adams 2007&lt;sup&gt;25&lt;/sup&gt;</th>
<th>Kemp 2016&lt;sup&gt;26&lt;/sup&gt;</th>
<th>Svaldi 2016&lt;sup&gt;27&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Is the case definition adequate?</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Representativeness of the cases</td>
<td>-</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Selection of controls</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Definition of controls</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Comparability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study controls for age and sex?</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Study controls for history of non-football associated head injuries?</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Exposure</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ascertainment of exposure</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Same method of ascertainment for cases and controls?</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Non-Response rate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total score</td>
<td>7</td>
<td>7</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>4</td>
<td>1</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>6</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>
Tarnutzer et al. 2016: Persistent effects of playing football and associated (subconcussive) head trauma on brain structure and function – a systematic review of the literature

Table 2: NOS ratings for cohort studies

<table>
<thead>
<tr>
<th>Quality assessment criteria</th>
<th>Acceptable (*)</th>
<th>Forbes 201428</th>
<th>Straume 200929</th>
<th>Janda 200230</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Representativeness of the exposed cohort</td>
<td>Representative of average football player in community (recruitment from at least one entire football team or training group)</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Selection of the non-exposed cohort</td>
<td>Drawn from the same community as the exposed cohort</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Ascertainment of exposure</td>
<td>Secure record, structured interview or standardized questionnaire</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Demonstration that outcome of interest was not present at start of study</td>
<td>yes</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Comparability</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study controls for age and sex?</td>
<td>yes</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Study controls for history of non-football associated head injuries</td>
<td>yes</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Outcome</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessment of outcome?</td>
<td>Structured interview blinded to exposure status or use of computer-based automatized testing (e.g. IMPACT, CogState)</td>
<td>*</td>
<td>*</td>
<td>-</td>
</tr>
<tr>
<td>Was follow-up long enough for outcomes to occur?</td>
<td>Observation period or duration since exposure ≥6 months or one season</td>
<td>*</td>
<td>-</td>
<td>*</td>
</tr>
<tr>
<td>Adequacy of follow up of cohorts</td>
<td>Follow-up rate ≥50%</td>
<td>*</td>
<td>*</td>
<td>-</td>
</tr>
<tr>
<td>Total score</td>
<td></td>
<td>9</td>
<td>8</td>
<td>7</td>
</tr>
</tbody>
</table>
Tarnutzer et al. 2016: Persistent effects of playing football and associated (subconcussive) head trauma on brain structure and function – a systematic review of the literature

Table 3: NOS ratings for cross-sectional studies (adapted from the rating scale for cohort studies)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Representativeness of the exposed cases</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Representativeness of non-exposed cases</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Ascertainment of exposure</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Selection of outcome parameters clearly specified in methods</td>
<td>yes</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Comparability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study controls for age and sex?</td>
<td>yes</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Study controls for history of non-football associated head injuries</td>
<td>yes</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Outcome</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessment of outcome?</td>
<td>Structured interview where blind to exposure status or use of computer-based automatized testing (e.g. IMPACT, CogState)</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Was follow-up long enough for outcomes to occur?</td>
<td>Observation period or duration since exposure was at least 6 months or one season</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Non-response rate</td>
<td>Non-response rate reported and below 50%</td>
<td>*</td>
<td>-</td>
<td>-</td>
<td>*</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Total score</td>
<td>8 7 6 9 6 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Tarnutzer et al. 2016: Persistent effects of playing football and associated (subconcussive) head trauma on brain structure and function – a systematic review of the literature

Newcastle-Ottawa quality assessment Scale (NOS) – coding manual (modified after the scale initially published by Wells and co-authors)¹

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

CASE-CONTROL STUDIES

Selection

1) Is the case definition adequate?
   a) Independent validation based on confirmed (former) membership of an official (high school / college / university / amateur / professional) football team and regular participation in football play. *
   b) Self report with no reference to official source.
   c) No description.

2) Representativeness of the cases
   a) Consecutive or obviously representative series of cases and controls with recruitment from at least one entire football team or training group. *
   b) Not satisfying requirements in part a) or not stated.

3. Selection of controls
   a) Athletes regularly participating in non-contact sports that do not have a high risk for head injuries. *
   b) Athletes performing contact sports or /and having a high risk for head injury.
   c) Controls not performing any sports or people drawn randomly.
   d) No description.

4) Definition of controls
   a) No history of exposure to regular football play. *
   b) No description of potential (former) football play.

Comparability

1) Comparability of cases and controls on the basis of the design or analysis
   a) Study controls for age and sex *
   b) Study controls for history of non-football associated head injuries (i.e., either present or absent in both cases and controls) *

Exposure
Tarnutzer et al. 2016: Persistent effects of playing football and associated (subconcussive) head trauma on brain structure and function – a systematic review of the literature

1) Ascertainment of exposure
   a) Structured interview where blind to case/control status. *
   b) Standardized questionnaire. *
   c) Interview not blinded to case/control status.
   d) Written self-report only.
   e) No description.

2) Same method of ascertainment for cases and controls
   a) Yes. *
   b) No.

3) Non-Response rate
   a) Same rate for both cases and controls (i.e., absolute difference in response rate ≤10%) (e.g., 56% vs. 65%). *
   b) Non-respondents described.
   c) Rate different and no designation.
Tarnutzer et al. 2016: Persistent effects of playing football and associated (subconcussive) head trauma on brain structure and function – a systematic review of the literature

COHORT STUDIES

Selection

1) Representativeness of the exposed cohort

- Truly representative of the average football player in the community (e.g., recruitment from an entire football league). *
- Somewhat representative of the average football player in the community (e.g., recruitment from at least one entire football team or training group). *
- Selected group of players, e.g. with recent head injury, cognitive deficits.
- No description of the derivation of the cohort.

2) Selection of the non-exposed cohort

- Drawn from the same community as the exposed cohort *
- Drawn from a different source
- No description of the derivation of the non-exposed cohort

3) Ascertainment of exposure

- Secure record. *
- Structured interview or standardized questionnaire. *
- Non-structured interview or written self-report.
- No description.

4) Demonstration that outcome of interest was not present at start of study

- Yes. *
- No.

Comparability

1) Comparability of cohorts on the basis of the design or analysis

- Study controls for age and sex. *
- Study controls for history of non-football associated head injuries (i.e., either present or absent in both exposed and non-exposed subjects) *

Outcome

1) Assessment of outcome?

- Independent blind assessment stated in the paper (i.e., examiner such as neuropsychologist was unaware of head injuries or other critical events during the observation period). *
- Assessment (e.g., neuropsychological testing) was automatized (i.e., by use of IMPACT, CogState)*
- Self-report or non-blinded assessment by specialist.
- No description.
Tarnutzer et al. 2016: Persistent effects of playing football and associated (subconcussive) head trauma on brain structure and function – a systematic review of the literature

2) Was follow-up long enough for outcomes to occur?

a) Yes, observation period or duration since exposure was at least 6 months or one season. *

b) No, observation period or duration since exposure was less than 6 months.

3) Adequacy of follow up of cohorts

a) Complete follow up - all subjects accounted for. *

b) Follow-up rate ≥ 50%. *

c) Follow up rate < 50% and no description of those lost.

d) No statement.
Tarnutzer et al. 2016: Persistent effects of playing football and associated (subconcussive) head trauma on brain structure and function – a systematic review of the literature

CROSS-SECTIONAL STUDIES

Selection

1) Representativeness of the exposed cases

   a) Truly representative of the average football player in the community (e.g., recruitment from an entire football league). *
   b) Somewhat representative of the average football player in the community (e.g., recruitment from at least one entire football team or training group). *
   c) Selected group of players, e.g. with recent head injury, cognitive deficits.
   d) No description of the derivation of the cohort.

2) Selection of the non-exposed cases

   a) Drawn from the same community as the exposed participants. *
   b) Drawn from a different source.
   c) No description of the derivation of the non-exposed study participants.

3) Ascertainment of exposure

   a) Secure record. *
   b) Structured interview or standardized questionnaire. *
   c) Non-structured interview or written self-report.
   d) No description.

4) Selection of outcome parameters clearly specified in methods

   a) Yes. *
   b) No.

Comparability

1) Comparability of exposed and non-exposed participants on the basis of the design or analysis

   a) Study controls for age and sex. *
   b) Study controls for history of non-football associated head injuries (i.e., either present or absent in both exposed and non-exposed subjects) *

Outcome

1) Assessment of outcome?

   a) Independent blind assessment stated in the paper (i.e., examiner such as neuropsychologist was unaware of head injuries or other critical events during the observation period). *
   b) Assessment (e.g., neuropsychological testing) was automatized (i.e., by use of IMPACT, CogState)*
Tarnutzer et al. 2016: Persistent effects of playing football and associated (subconcussive) head trauma on brain structure and function – a systematic review of the literature

c) Self-report or non-blinded assessment by specialist.
d) No description.

2) Was observation period long enough for outcomes to occur?

a) Yes, observation period or duration since exposure was at least 6 months or one season. *
b) No, observation period or duration since exposure was less than 6 months.

3) Non-response rate

a) Non-response rate reported and below 50%, i.e. more than 50% of contacted subjects eventually participated (some of them may have afterwards been excluded for various reasons...). *
b) Non-response rate reported, but above 50%.
c) Non-response rate not reported.
Tarnutzer et al. 2016: Persistent effects of playing football and associated (subconcussive) head trauma on brain structure and function – a systematic review of the literature

References


Tarnutzer et al. 2016: Persistent effects of playing football and associated (subconcussive) head trauma on brain structure and function – a systematic review of the literature.


Tarnutzer et al. 2016: Persistent effects of playing football and associated (subconcussive) head trauma on brain structure and function – a systematic review of the literature

**Supplementary file 3: Table 1: key characteristics of included studies**

<table>
<thead>
<tr>
<th>Citation, year</th>
<th>Study design</th>
<th>Study population</th>
<th>Sample size (test / control group) (n)</th>
<th>Subjects with history of TBI excluded / Time since last concussion</th>
<th>Appropriate controls * / appropriate control of type-1 errors † / response rate (%)</th>
<th>Assessment of heading-frequency (rating)</th>
<th>Gender / age (mean±1SD, range) test / control group</th>
<th>Aspects evaluated</th>
<th>Neuro-cognitive tests applied</th>
<th>design issues/notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abreu et al. 1990¹</td>
<td>Retrospective case-control study</td>
<td>College football players / college tennis players</td>
<td>31 / 31</td>
<td>No / NA</td>
<td>yes / no / 62/220 (28%)</td>
<td>NA</td>
<td>Males /19.1±1.4 / 19.4±1.1</td>
<td>NCP</td>
<td>RPM, SDMT, perceptual speed test, PASAT</td>
<td>Contra: No form of type-1 error control reported</td>
</tr>
<tr>
<td>Adams et al. 2007²</td>
<td>Retrospective case-control study</td>
<td>College football players / non-playing controls</td>
<td>10 / 10</td>
<td>no / NA</td>
<td>No / yes / NR</td>
<td>NA</td>
<td>Males / 21±2 / 26±8y</td>
<td>MRI (VBM)</td>
<td>NA</td>
<td>Contra: No correlation analysis performed between atrophy, cognition and heading-frequency</td>
</tr>
<tr>
<td>Downs et al. 2002³</td>
<td>Retrospective case-control study</td>
<td>College (n=26) and professional (n=6, current or former) football players / swimmers</td>
<td>32 / 29</td>
<td>no / NA</td>
<td>no / no / NR</td>
<td>Headling-exposure-index (low-quality rating)</td>
<td>Mixed / 19.8±1.5y (college) and 41.5±9.8y, professional) / 19.5±1.2y (college) and 42.9±15.1y (elderly swimmers)</td>
<td>NCP</td>
<td>CPT, PASAT, FTT, WCST,</td>
<td>Contra: Dichotomy of age-ranges studied. No form of type-1 error control reported.</td>
</tr>
<tr>
<td>Ellemberger et al. 2007⁴</td>
<td>Retrospective case-control study</td>
<td>University football players with or without concussion with 6-8 months / NA</td>
<td>22 / NA</td>
<td>no (head trauma was inclusion criterion) / NA</td>
<td>NA / yes / NR</td>
<td>NA</td>
<td>Females / 22.7 (SD NR) / 22.3 (SD NR)</td>
<td>NCP</td>
<td>CVLT, Ruff 2 &amp; 7 selective attention test, BTA, SDMT, Stroop, Tower of London DX, Letter Fluency Test, forward and backward digit span, SRT, CRT</td>
<td>Contra: Unclear how players with concussion were selected. Pro: Bonferroni-correction for the number of comparisons applied.</td>
</tr>
<tr>
<td>Forbes et al. 2014⁵</td>
<td>Prospective cohort study</td>
<td>Interscholastic players with or without hx of concussion</td>
<td>210 / NA</td>
<td>no, used to segregate participants / 27.1±21.1m</td>
<td>NA / unclear / NR</td>
<td>Recorded prospectively by observer (high-quality rating)</td>
<td>Females / 16.0±1.3 / 15.9±1.3</td>
<td>NCP</td>
<td>ANAM (SRT, CPT, MTH, MSP, STN)</td>
<td>Contra: Low number of headers per game (1.3-1.4). No form of type-1 error control reported.</td>
</tr>
<tr>
<td>Guskiewicz et al. 2002⁶</td>
<td>Retrospective case-control study</td>
<td>Collegiate players / non-football contact sports athletes (n=96) and controls (n=53)</td>
<td>91 / 149</td>
<td>no / NA</td>
<td>no / yes / NR</td>
<td>NA</td>
<td>Mixed / 19.1±1.3y / 19.3±1.4y (non-soccer athletes) and 20.2±1.2y (controls)</td>
<td>NCP</td>
<td>TMT B, COWAT, Stroop, HVLT, SDMT, WDST,</td>
<td>Contra: Delay after last concussion is not reported. Pro: Tukey post hoc analyses performed for specific group comparisons.</td>
</tr>
</tbody>
</table>
Tarnutzer et al. 2016: Persistent effects of playing football and associated (subconcussive) head trauma on brain structure and function – a systematic review of the literature

<table>
<thead>
<tr>
<th>Study</th>
<th>Design</th>
<th>Participants</th>
<th>Sample Size</th>
<th>Sex</th>
<th>Age (range)</th>
<th>Imaging Modalities</th>
<th>Contra</th>
</tr>
</thead>
<tbody>
<tr>
<td>Janda et al. 2002⁷</td>
<td>Prospective cohort study</td>
<td>Youth football players / NA</td>
<td>57 / NA</td>
<td>NA</td>
<td>NA / unclear / NR</td>
<td>Recorded prospectively by observer (high-quality rating) / NA</td>
<td>No form of type-1 error control reported.</td>
</tr>
<tr>
<td>Jordan et al. 1996⁸</td>
<td>Retrospective case-control study</td>
<td>Professional football players / elite track athletes</td>
<td>20 / 20</td>
<td>NA</td>
<td>yes / no / 20/25 (80%)</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Kemp et al. 2016⁹</td>
<td>Prospective case-control study</td>
<td>Professional football players / university non-contact sports athletes</td>
<td>32 / 33</td>
<td>NA</td>
<td>yes / yes / NA</td>
<td>NA</td>
<td>Males / 18.8±0.8y / 20.8±1.3y</td>
</tr>
<tr>
<td>Koerte et al. 2012¹⁰</td>
<td>Retrospective case-control study</td>
<td>Professional football players / swimmers from competitive clubs</td>
<td>12 / 11</td>
<td>yes / NA</td>
<td>yes / yes / 12/40 (30%)</td>
<td>NA</td>
<td>Males / 19.7±1.6y / 21.4±2.8y</td>
</tr>
<tr>
<td>Koerte et al. 2015¹¹</td>
<td>Retrospective case-control study</td>
<td>Former professional players still active at recreational level / former professional athletes from non-contact sports still active at recreational level</td>
<td>11 / 14</td>
<td>yes / NA</td>
<td>yes / no / NR</td>
<td>Self-reported by player (low-quality rating)</td>
<td>Males / 52.0±6.8y / 46.9±9.7y</td>
</tr>
<tr>
<td>Koerte et al. 2015¹²</td>
<td>Retrospective case-control study</td>
<td>Former professional players still active at recreational level / former professional athletes from non-contact sports still actively participating in these sports</td>
<td>15 / 15</td>
<td>yes (only those with moderate / severe TBI) / NA</td>
<td>yes / no (NCP), yes (imaging) / NR</td>
<td>Self-reported by player (low-quality rating)</td>
<td>Males / 49.3±5.1y / 49.6±4.6y</td>
</tr>
</tbody>
</table>
Tarnutzer et al. 2016: Persistent effects of playing football and associated (subconcussive) head trauma on brain structure and function – a systematic review of the literature

<table>
<thead>
<tr>
<th>Study</th>
<th>Design</th>
<th>Participants</th>
<th>Cases / Controls</th>
<th>Recorded if concussion occurred within 3 months or if hx of TBI / NA</th>
<th>Recorded prospectively by observer (high quality rating)</th>
<th>Mixed / 15.9±1.2y / NA</th>
<th>NCP</th>
<th>Imaging (if applicable)</th>
<th>Contra:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kontos et al. 2011</td>
<td>Prospective cross-sectional study</td>
<td>Youth football players / NA</td>
<td>63 / NA</td>
<td>yes / no / NR</td>
<td>Mixed / 15.9±1.2y / NA</td>
<td>NCP</td>
<td>ImPACT 2.0</td>
<td></td>
<td>NA</td>
</tr>
<tr>
<td>Lipton et al. 2013</td>
<td>Prospective cross-sectional study</td>
<td>Youth football players / NA</td>
<td>37 / NA</td>
<td>yes / no / NR</td>
<td>Self-reported by player (low-quality rating)</td>
<td>Mixed / 30.9y (range:21-44y)</td>
<td>Imaging (DTI)</td>
<td>CogState (tests of psychomotor speed, attention, executive function and memory)</td>
<td>Contra:</td>
</tr>
<tr>
<td>Matser et al. 1998</td>
<td>Retrospective case-control study</td>
<td>Professional football players / elite noncontact sports athletes</td>
<td>53 / 27</td>
<td>yes / no / NR</td>
<td>Self-reported by player (low-quality rating)</td>
<td>Males / 25.4±4.1 / 24.5±4.5y</td>
<td>NCP</td>
<td>Same tests as in Matser et al. 1998</td>
<td>Contra:</td>
</tr>
<tr>
<td>Matser et al. 1999</td>
<td>Retrospective case-control study</td>
<td>Amateur football players / amateur athletes involved in swimming, track</td>
<td>33 / 27</td>
<td>yes / no / NR</td>
<td>Males / 24.9±4.2y / 24.5±4.5y</td>
<td>NCP</td>
<td>Same tests as in Matser et al. 1998</td>
<td>Contra:</td>
<td></td>
</tr>
<tr>
<td>Matser et al. 2001</td>
<td>Retrospective cross-sectional study</td>
<td>Professional soccer players / NA</td>
<td>84 / NA</td>
<td>no / NA</td>
<td>Self-reported by player (low-quality rating)</td>
<td>Males / 24 (median), 19-32 (10-90 percentile) / NA</td>
<td>NCP</td>
<td>Same tests as in Matser et al. 1998 except digit symbol forward, digit symbol backward</td>
<td>Contra:</td>
</tr>
<tr>
<td>Rutherford</td>
<td>Retrospective</td>
<td>Amateur</td>
<td>22 / 41 (17)</td>
<td>yes if head / yes / yes / Combination</td>
<td>Males / 20.6±1.2y /</td>
<td>NCP</td>
<td>NART, ROCF,</td>
<td></td>
<td>NA</td>
</tr>
</tbody>
</table>

Contra: Significantly larger intake of alcohol in control group. Insufficient control of type-1 errors (Bonferroni correction for 27 tests, but 276 tests applied, see also 16)
Contra: Significantly larger intake of alcohol in test group. Insufficient control of type-1 errors (Bonferroni correction for 8 tests, but >175 tests applied, see also 16)
Contra: No form of type-1 error control reported.
<table>
<thead>
<tr>
<th>Study</th>
<th>Design</th>
<th>Participants</th>
<th>Injury Occurrence</th>
<th>Follow-up</th>
<th>Outcome Measures</th>
<th>Control for Type-1 Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>et al. 2005</td>
<td>Case-control study</td>
<td>University football players / amateur rugby players and amateur non-contact sport players</td>
<td>Injury occurred within 3 months / NA</td>
<td>109/156 (70%)</td>
<td>20.6±1.2y (rugby) and 20.7±1.7y (non-contact sports players)</td>
<td>WAIS-R digit symbol/digit span, TMT A/B, Stroop, WMS Logical Memory, WMS Verbal Paired Associate Immediate and Delayed, TAP, WCST, alternate uses, COWAT</td>
</tr>
<tr>
<td>Rutherford et al. 2009</td>
<td>Retrospective case-control study</td>
<td>Amateur university football players / amateur rugby players and amateur non-contact sport players</td>
<td>Yes if head injury occurred within 3 months / NA</td>
<td>165/254 (65%)</td>
<td>Combination of prospective assessment by observer and self-reporting by player (high-quality rating)</td>
<td>NART, ROCF, WAIS-R digit symbol/digit span, TMT A/B, Stroop, WMS Logical Memory, WMS Verbal Paired Associate Immediate and Delayed, TAP, WCST, alternate uses, COWAT</td>
</tr>
<tr>
<td>Salinas et al. 2009</td>
<td>Prospective cross-sectional</td>
<td>Youth football players / NA</td>
<td>No / NA</td>
<td>NA / unclear / NR</td>
<td>Recorded prospectively by observer (high-quality rating)</td>
<td>TMT A/B, RAVLT, Tower of London DX, WISC-IV vocabulary and block design subtests, CRI-C</td>
</tr>
<tr>
<td>Stephens et al. 2010</td>
<td>Retrospective case-control study</td>
<td>School team football players / rugby players and non-contact sport players</td>
<td>Yes if head injury occurred within 3 months / NA</td>
<td>193/337 (57%)</td>
<td>Recorded prospectively by observer (high-quality rating)</td>
<td>WISC-R vocabulary Subtest, CDI-S, EPQI, STAI, ROCF, WAIS-R digit symbol/digit span, TMT, Stroop, WMS-R Logical Memory Immediate and Delayed, TAP, WCST, alternate uses.</td>
</tr>
<tr>
<td>Straume et al. 2005</td>
<td>Retrospective cross-sectional</td>
<td>Active professional football players / NA</td>
<td>No / NA</td>
<td>271/300 (90%)</td>
<td>Self-reported by player (low-quality)</td>
<td>CogSport (tasks: SRT, CRT, congruent reaction)</td>
</tr>
</tbody>
</table>

Pro: thorough control for type-1 errors by presenting a data analysis paradigm and applying Rom’s procedure.

Contra: no control for type-1 errors. Nonetheless no trend towards relation in the comparison group.
Tarnutzer et al. 2016: Persistent effects of playing football and associated (subconcussive) head trauma on brain structure and function – a systematic review of the literature

<table>
<thead>
<tr>
<th>Study</th>
<th>Type</th>
<th>Sample Description</th>
<th>Sample Size</th>
<th>Concussion History</th>
<th>Control Group Details</th>
<th>Control Strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Straume et al. 2009^24</td>
<td>Prospective cohort study</td>
<td>Active professional football players with or without concussion / NA</td>
<td>144 / NA</td>
<td>No / NA</td>
<td>NA / yes / NR NCP Males / 25.2 (SD NA, range: 18-34y)</td>
<td>Pro: Bonferroni correction for multiple statistical tests was applied in order to control for type-1 errors.</td>
</tr>
<tr>
<td>Svaldi et al. 2016^25</td>
<td>Prospective case-control study</td>
<td>High school football players / high school non-contact sports athletes</td>
<td>14 / 12</td>
<td>No</td>
<td>Yes / yes / NR Head acceleration event monitoring (high-quality rating) Females / 15.9 (range: 15-17) / 15.9 (range 15-18)</td>
<td>Pro: head accelerations were monitored and used to calculate a cumulative head acceleration exposure.</td>
</tr>
<tr>
<td>Tysvaer and Storli 1989^26</td>
<td>Retrospective case-control study</td>
<td>Active professional football players / men with various occupation and w/o hx of concussion</td>
<td>69 / 69</td>
<td>Yes (if unrelated to football) / NA</td>
<td>No / unclear / NR NA Males / 15 -34y EEG None</td>
<td>Contra: all abnormal EEGs were obtained in players labeled “not headers”. No mentioning of control for type-1 errors.</td>
</tr>
<tr>
<td>Tysvaer et al. 1989^27</td>
<td>Retrospective case-control study</td>
<td>Former professional football players / controls from „different occupational groups“</td>
<td>37 / 37</td>
<td>Yes (if unrelated to football) / NA</td>
<td>No / unclear / 43/50 (86%) NA Males / 49y (range: 35-64y) / no numbers („same age range“) EEG None</td>
<td>Contra: No EEG differences between „headers“ and „non-headers“. No mentioning of control for type-1 errors.</td>
</tr>
<tr>
<td>Tysvaer and Lochen 1991^28</td>
<td>Retrospective case-control study</td>
<td>Former professional football players / patients hospitalized for various disorders</td>
<td>37 / 20</td>
<td>Yes (if unrelated to football) / NA</td>
<td>No / no / NR NR Males / 49y (range: 35-64y) / no numbers („same age range“) NCP WAIS, TMT A/B, HWRAST</td>
<td>Contra: same patient group as in 27. Control group is poorly characterized (reason for hospitalization unclear). No form of type-1 error control reported</td>
</tr>
<tr>
<td>Vann Jones et al. 2014^29</td>
<td>Retrospective case-control study</td>
<td>Former professional football players / UK-based MCI prevalence study</td>
<td>92</td>
<td>no / NA</td>
<td>NA / no / 138 / 300 (46%) Heading-exposure-index (low-quality rating) Males / 67.5±7.0y (95%CI) / NA</td>
<td>Contra: Comparison with epidemiologic data. No control for physical activity.</td>
</tr>
<tr>
<td>Witol et al. 2002^30</td>
<td>Retrospective case-control study</td>
<td>Soccer players (high-school,</td>
<td>60 / 12</td>
<td>yes (if head trauma) no / no / NR Self-reported by player Males /21.3±4.9/ 22.9±3.6 NCP Shipley Institute of Living Scale, TMT</td>
<td>Contra: No mentioning of control for type-1 errors.</td>
<td></td>
</tr>
</tbody>
</table>
Tarnutzer et al. 2016: Persistent effects of playing football and associated (subconcussive) head trauma on brain structure and function – a systematic review of the literature

<table>
<thead>
<tr>
<th>Study</th>
<th>Type</th>
<th>Participants</th>
<th>Methodology</th>
<th>Controls</th>
<th>Outcome Measures</th>
<th>NCP</th>
<th>Pre/post task</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhang et al. 2013</td>
<td>Retrospective case-control study</td>
<td>High school football players / non-football players</td>
<td>12 / 12</td>
<td>unclear / unclear / NR</td>
<td>Self-reported by player (low-quality rating)</td>
<td>Females / 16.5y (range: 15-18y) for both groups.</td>
<td>NCP</td>
<td>Pre-point task and anti-point task</td>
</tr>
</tbody>
</table>

Abbreviations: AMIPB=adult memory and information processing battery; ANAM=automated neuropsychological assessment metrics; BESS=balance error scoring system; BIS=Barrett impulsivity scale; BTA=Brief Test of Attention; CDI-S=childhood depression inventory short form; Cho=choline; COWAT=controlled oral word association test; CPT=continuous performance test; Cr=creatine; CRT=choice reaction time; CRI-C=children’s concussion resolution index; CVLT=California verbal learning task; DTI=diffusion tensor imaging; ECF=eyes closed firm surface; ECU=eyes closed unstable foam surface; EHQ=Einstein heading questionnaire; EOF=eyes open firm surface; EOU=eyes open unstable foam surface; EPQJ=Eysenck personality questionnaire junior; FA=fractional anisotropy; fMRI=functional magnetic resonance imaging; FTT=finger tapping test; HVLT=Hopkins verbal learning test; HWRAST=Halstead Wepman Reitan aphasia screening test; hx=history; LOC=loss of consciousness; ml=myo-inositol; MRI=magnetic resonance imaging; MRS=magnetic resonance spectroscopy; MSP=matching to sample; MTH=math processing; NA=not available; NART=national adult reading test; NCP=neurocognitive performance; NR=not reported; PASAT=paced auditory serial addition task; RAVLT=Rey auditory verbal learning test; ROCF=Rey-Osterrieth complex figure; RPM=Raven progressive matrices test; SCOLP=speed and capacity of language processing test; SDMT=symbol digit modality test; SRT=simple reaction time; STAI=state-trait anxiety inventory; STN=Sternberg memory; TAP=test of attention performance; TBI=traumatic brain injury; TEA=test of everyday attention; TMT=trail-making test; TYM=test your memory questionnaire; WAIS=Wechsler adult intelligence scale; WCSS=Wisconsin card sorting test; WDST=Wechsler digit span test; WISC-IV=Wechsler intelligence scale for children-IV; WMS=Wechsler memory scale; WRAML=wide range of assessment memory and learning; |

* Only control groups that were age-matched and gender-matched and that participated in non-contact sports with a comparable physical-activity profile (e.g., swimming, track, tennis) were considered “appropriate” or low risk for bias.

† Only studies that reported sufficient controlling for multiple testing (e.g., by applying Bonferroni correction) were considered “appropriate” or low risk for type-1 errors, while studies falling short of these criteria were considered “inappropriate” or high risk.

References


https://mc.manuscriptcentral.com/bjsm
Tarnutzer et al. 2016: Persistent effects of playing football and associated (subconcussive) head trauma on brain structure and function – a systematic review of the literature


Supplementary file 4: Frequency of use of individual neurocognitive tests

Overall, more than 60 distinct neurocognitive tests were identified in the 23 studies reporting neurocognitive test performance. Most tests were used in few studies only and were abnormal in a minority of these studies only, as shown in detail in supplementary figure 1.

While we did not identify any studies that compared the diagnostic accuracy of individual neurocognitive tests, we analyzed which tests were most widely used in those neurocognitive domains considered most important in the context of (sub)concussive brain injury or mild traumatic brain injury (mTBI). We therefore restricted this analysis to the following domains: attention, memory, executive functions and visuospatial skills. The three most frequently applied tests for each domain and their resulting frequency of abnormal test results are shown in Table 1. On average, the fraction of neurocognitive tests with significant differences between the groups compared (football players vs. non-football playing controls) or linked to the frequency of heading or head-injuries was 0.29±0.18.

| NC domain          | Top-ranked |  | Second-ranked |  | Third-ranked |  | mean±SD |
|--------------------|------------|  |               |  |              |  |        |
|                    | Test       | Fraction [%] | Test       | Fraction [%] | Test       | Fraction [%] |        |
| Attention          | TMT A      | 3/10 [30]    | PASAT      | 0/6 [0]      | Digit span forward | 2/5 [40]    | 0.23±0.21 |
| Executive functions| TMT B      | 1/12 [8]     | Stroop     | 1/8 [13]     | WCST       | 2/7 [29]     | 0.16±0.11 |
| Memory and learning| ROCF delayed recall | 4/9 [44]  | ROCF immediate recall | 3/9 [33]  | WMS        | 2/6 [33]     | 0.37±0.06 |
| Visuospatial functions| ROCF copy raw | 2/8 [25]  | Benton facial recognition | 2/3 [67] | N/A        | N/A          | 0.46±0.29 |
| All pooled         | 0.27±0.15  | 0.28±0.29    | 0.34±0.06  | 0.29±0.18    |            |              |        |

Abbreviations: N/A=not available; NC=neurocognitive; NCT=neurocognitive test; PASAT=paced auditory serial addition task; ROCF=Rey-Osterrieth complex figure; SD=standard deviation; TMT=trail-making test; WCST=Wisconsin card sorting test; WMS=Wechsler memory scale;
Tarnutzer et al. 2016: Persistent effects of playing football and associated (subconcussive) head trauma on brain structure and function – a systematic review of the literature
Tarnutzer et al. 2016: Persistent effects of playing football and associated (subconcussive) head trauma on brain structure and function – a systematic review of the literature

**Supplementary figure 1**

Frequencies of all neurocognitive tests applied in a given subgroup are depicted in ascending order. Yellow bars refer to the total number of studies that applied this test and blue bars refer to the fraction of studies in which an abnormal test result was retrieved. Separate plots are provided for case-control studies comparing neurocognitive tests between football players and controls, for studies investigating the impact of heading-frequency on neurocognition and for studies focusing on the impact of the number of head-injuries on neurocognitive performance.

_Abbreviations:_ AMIPB=adult memory and information processing battery; BIS=Barrett impulsivity score; BTA=Brief Test of Attention; CPT=continuous performance test; Cr=creatine; CRT=choice reaction time; CRI-C=children’s concussion resolution index; CVLT=California verbal learning task; FTT=finger tapping test; HWRAST=Halstead Wepman Reitan aphasia screening test; HVLT=Hopkins verbal learning test; MSP=matching to sample; MTH=math processing; N/A=not available; NART=national adult reading test; NCP=neurocognitive performance; NR=not reported; PASAT=paced auditory serial addition task; RAVLT=Rey auditory verbal learning test; ROCF=Rey-Osterrieth complex figure; RPM=Raven progressive matrices test; SCOLP=speed and capacity of language processing test; SDMT=symbol digits modality test; SRT=simple reaction time; STN=Sternberg memory; TAP=test of attention performance; TEA=test of everyday attention; TMT=trail-making test; TYM=test your memory questionnaire; WAIS=Wechsler adult intelligence scale; WCST=Wisconsin card sorting test; WDST=Wechsler digit span test; WISC-IV=Wechsler intelligence scale for children-IV; WMS=Wechsler memory scale; WRAML=wide range of assessment memory and learning.
PERSISTENT EFFECTS OF PLAYING FOOTBALL AND ASSOCIATED (SUBCONCUSSIVE) HEAD TRAUMA ON BRAIN STRUCTURE AND FUNCTION – A SYSTEMATIC REVIEW OF THE LITERATURE

Tarnutzer AA M.D. (1,2), Straumann D M.D. (1,2), Brugger P PhD. (1), Feddermann-Demont N M.D. (1,2)

1) Department of Neurology, University Hospital Zurich and University of Zurich, Zurich, Switzerland
2) Swiss Concussion Center, Schulthess Clinic, Zurich, Switzerland

Short title: persistent effects of football on the brain

Corresponding author:
Alexander A. Tarnutzer, MD
Department of Neurology, University Hospital Zurich
Frauenklinikstr. 26, 8091 Zurich, Switzerland
Phone: 0041 44 255 11 11
Fax: 0041 44 255 43 80
Email: alexander.tarnutzer@access.uzh.ch

Statistics:
word count for the text (excluding legends, references & abstract): 4908
word count for the abstract: 250
character count for the title (including spaces): 153
number of figures: 2
number of tables: 5
supplementary files: 4
ABSTRACT

Aim/objective: There is ongoing controversy about persistent neurological deficits in active and former football (soccer) players. We reviewed the literature for associations between football activities (including heading/head-injuries) and decline in brain structure/function.

Design: Systematic literature review


Eligibility criteria for selecting studies: Original studies reporting on football-related persistent effects on brain structure/function. Results from neurocognitive testing, neuroimaging and electroencephalography were compared to controls and/or correlated with heading-frequency and/or head-injuries. Methodological quality was rated for risk-of-bias including appropriateness of controls, correction for multiple statistical testing, and assessment of heading-frequency and head-injuries.

Results: Thirty studies with 1691 players were included. Those 57% (8/14) of case-control studies reporting persistent neurocognitive impairment had higher odds for inappropriate control of type-1 errors (OR=17.35 [95%-CI=10.61-28.36]) and for inappropriate selection of controls (OR=1.72 [1.22-2.43]) than studies observing no impairment. Studies reporting a correlation between heading-frequency and neurocognitive deficits (6/17) had lower quality of heading-assessment (OR=14.20 [9.01-22.39]) than studies reporting no such correlation. In 7/13 studies (54%) the number of head-injuries correlated with the degree of neurocognitive impairment. Abnormalities on neuroimaging (6/8 studies) were associated with subclinical neurocognitive deficits in 3/4 studies.

Summary/conclusions: Various methodological shortcomings limit the evidence for persistent effects of football play on brain structure/function. Sources of bias include low-quality assessment of heading-frequency, inappropriate control for type-1 errors and inappropriate selection of controls. Combining neuroimaging-techniques with neurocognitive...
testing in prospective studies seems most promising to further clarify on the impact of football on the brain.

**Bullet statements**

*What is already known?*

- Repetitive head injuries and heading the ball were suggested to be linked to persistent neurocognitive impairments and structural brain abnormalities in advanced neuroimaging in both professional and amateur football players.

- However, there is ongoing controversy to what extent these findings are real or rather result from methodological limitations.

- A systematic assessment of existing studies using prospectively defined criteria is therefore needed to improve our understanding of persistent effects of football on the brain and to better estimate the role of methodological limitations in previous studies.

*What are the new findings?*

- While the majority of studies addressing the effect of football play and football-related injuries on neurocognitive functions reported significant impairment in at least one domain, methodological shortcomings were found to be more frequent in studies with reportedly significant findings.

- Evidence for a correlation between heading frequency and neurocognitive deficits was weak and likely biased by inaccurate heading-frequency estimates.

- Although the rate of football-related head injuries was reportedly higher in women than men, females were under-represented in studies that reported neurocognitive impairment compared to those studies that did not identify deficits, which may be
related to the fact that none of these studies included (retired) female professional football players.

- Combining neuroimaging and neurocognitive testing in prospective longitudinal and cross-sectional studies in male and female players to link structural and functional deficits seems most promising to further clarify associations between football play and brain abnormalities.

Key words: football (soccer), heading, head injury, neurocognitive testing, systematic review
INTRODUCTION

Concussions (i.e., a subtype of mild traumatic brain injury (mTBI) without structural abnormalities on conventional CT or MRI)\(^1\) represent 1-5% of all football-(soccer-) related injuries.\(^2\) While most players return to play within 7-10 days, head-trauma related symptoms may last for weeks to months in 10-15% and even persist in selected cases.\(^6\) Neurodegenerative disorders (such as Alzheimer disease) have been reported in retired professional football players and in athletes from other contact sports as rugby and American football.\(^7\) A postulated association between football play and chronic traumatic encephalopathy, however, remains controversial,\(^9\) and the effect of football-related concussions is not well understood.

Likewise, the impact of purposeful heading the ball to play and guide its direction – unique to football – on the brain has been debated.\(^10\) On average, players head the ball 1-16 times during a competitive football match,\(^12\) accumulating over a season to several hundred headings\(^17\) and to many thousand headings during a professional football career. This has raised concerns that heading may – similar to boxers receiving punches to the head – pose players at increased risk for “subconcussive” trauma,\(^20\) potentially resulting in neuronal damage similar to that in repetitive concussions but not accompanied by overt symptoms.\(^20\)

These considerations have led to uncertainty in football players and their (medical) attendants,\(^1\) albeit such a link is far from being established and the impact of parameters such as heading technique, player’s age and playing position remain unclear. Nonetheless, with raising concerns and facing a concussion litigation, the football federation of the USA issued in November 2015 a ban for heading in children aged 10 years or less and limited heading in children aged 11-13 years.\(^26\) Concerns that the maturing brain could be especially vulnerable to subconcussive head injury may have supported this decision.

Research interest in associations between concussion, heading and persistent changes of the human brain has grown substantially. At the end of the last century, a series of case-
control studies indicated persistent neurocognitive impairments in Dutch professional\textsuperscript{16,19} and amateur\textsuperscript{27} football players. These studies were the basis for further investigations addressing functional, structural and metabolic brain changes in football players. While some studies confirmed neurocognitive abnormalities compared to controls,\textsuperscript{28,29} others found no such evidence.\textsuperscript{14,15,30,31} Likewise, associations between neurocognitive deficits and heading-frequency were reported by some,\textsuperscript{16,17,28,29} but not by others.\textsuperscript{14,15,30-34}

In accordance with neuroimaging for TBI,\textsuperscript{35} different protocols were applied to study structural (diffusion-tensor imaging [DTI],\textsuperscript{36} voxel-based MR-morphometry [VBM]) and metabolic (functional MRI [fMRI], magnetic-resonance spectroscopy [MRS]) brain changes in football players and to correlate with neurocognitive function. In two small case-control studies, memory impairment was linked to cortical thinning in former professionals\textsuperscript{37} and to diffuse white-matter abnormalities in amateur players,\textsuperscript{17} while recently in a prospective case-control study over five years in professional players no such link could be drawn.\textsuperscript{38}

In summary, whether or not football play is linked to persistent changes in brain function/structure remains controversial. Against this background, we aimed to systematically review the literature on associations between football play and persistent changes in brain function/structure and the impact of heading-frequency and concussive head-injuries. Assessing study quality and identifying methodological limitations using standardized tools for reporting risk-of-bias was a special focus.\textsuperscript{10}
MATERIAL AND METHODS

Data sources and searches

A literature search (MEDLINE, Embase, PsycINFO, CINAHL, Cochrane CRCT, SportDiscus, Cochrane DSR=4) was performed (02/08/2016) to identify articles reporting on associations between football play and especially heading and football-related head-injuries and persistent structural/functional changes of the brain. The MEDLINE (OVID) search strategy was translated for each database, and is reported in supplement 1.

“Persistent” changes were defined as changes that were still recognizable >6 months after a potential impact or linked to exposure to football play for >6 months or a full season. We adhered to the time-frame usually applied for the persistent post-concussive syndrome, albeit no consensus-based definition of this term exists.

We also performed a manual search of reference lists from eligible articles. We did not seek to identify research abstracts from meeting proceedings or unpublished studies, nor non-English language studies. Retrospective or prospective studies with five or more participants were eligible. This review complies with PRISMA guidelines.

Study selection

We identified 2191 citations for screening and included 30 studies for quantitative synthesis (Figure 1) based on abstract and for selected studies full-text review by two experienced neurologists (NFD, AAT). Articles were selected using pre-determined criteria (supplement 1).

/* Figure 1 about here */
Data extraction and quality assessment

Reports on neurocognitive testing, neuroimaging, postural control and electroencephalography (EEG) were considered. Data extraction was performed by AAT and confirmed by NFD. When extracting data from selected studies, we assessed the type of study, the type of diagnostic tests performed, the frequency of heading and head-injuries, and the level of play, distinguishing between youth, high-school/college (including inter-scholar), university, amateur, active professional and former professional players. In studies reporting on neurocognitive testing, all tests applied were retrieved and assigned to the category that best described the domain of neurocognitive function evaluated. Categories were: abstract reasoning, attention, (verbal) creativity and divergent thinking, decision-making, executive functions, intelligence, language and language-associated functions, memory/learning, mood, motor skills, and visuospatial skills.

A standardized risk-of-bias assessment was performed using the Newcastle-Ottawa Scale (NOS). Its use for non-randomized case-control studies and observational studies has been promoted by the Cochrane Collaboration. The NOS requires rating the selection, comparability and exposure/outcome for a total of nine items. Study quality was rated as “good”, “fair” or “poor” (see and supplement 2). Studies rated as “poor” were excluded. The NOS included an assessment of the response rate when asked to participate. Studies with a low (<50%) participation rate, with participation rates differing >10% between football players and controls or studies that did not report response rates were rated as high-risk for selection bias. Whenever non-football playing controls were available (n=21 studies), their suitability was rated. Only control groups that were age- and gender-matched and that participated in non-contact sports with a comparable physical-activity profile (e.g., swimming, track, tennis) were considered “appropriate” or low-risk for bias. A distinct (i.e., lower/higher) physical-activity profile may introduce a bias and observed differences may be
attributed falsely to effects of football play. Controls falling short of these criteria were considered “inappropriate” or high-risk of bias.

Based on previously described methodological limitations, we further assessed the quality of included studies regarding assessment of heading-frequency and history of head-injuries and control for type-1 errors. Rutherford identified insufficient control for type-1 errors as a potential source for false-positive statistical differences. To follow-up on this limitation, we assessed methods for avoiding type-1 errors. Only studies that reported sufficient controlling for multiple testing (e.g., by applying Bonferroni correction) were considered “appropriate” or low-risk for type-1 errors, while studies falling short of these criteria were considered “inappropriate” or high-risk. Both heading-frequency and head-injuries may be overstated or understated by players, posing them at risk for recall-bias. Therefore, only studies that prospectively collected data on heading-frequency (e.g. by an independent observer) were considered low-risk of bias, while studies falling short of these requirements (e.g., relied on self-reported numbers, a heading-exposure-index) were considered high-risk. We did not require loss of consciousness for making the diagnosis of concussion and relied on the original study authors’ assessment.

Data synthesis and analysis

Excel 2011 (Microsoft Coorp., Redmont, USA) and Matlab 7.0 (The MathWorks, Nantuck, USA) were used for data analyses. Statistical analyses were performed using two-sample t-tests (with Bonferroni correction) and odds ratios (ORs) including 95%-confidence-intervals (CIs).
RESULTS

From the 32 studies included for qualitative synthesis (Figure 1), one was excluded because of “poor” quality on NOS and one was removed because of duplicity of data. Among the 30 studies included for quantitative synthesis (n=1691, 22.4% females), only six were prospective. Twenty-three studies (76.7%, n=1518) reported on results of neurocognitive testing, while data on neuroimaging were provided in eight studies (26.7%, n=143). Information on EEG (6.7%, n=106) was available from two; postural stability (3.3%, n=15) was provided in one study (Tables 1 and 2). Four studies reported on more than one modality (supplement 3). NOS ratings were “good” and “fair”, respectively, in 15 studies each (supplement 2). Key domains for neurocognitive testing of (sub)concussive brain injury (attention, executive functions, memory) were assessed by 18 of 23 studies, while in the remaining five studies one (n=3) or two (n=2) key domains were missing.

Case-control studies reporting on neurocognitive testing

Fourteen studies compared neurocognitive test results in football players (n=581) with those from controls (n=348). On average, 8.7±5.8 tests covering 4.4±1.9 categories were administered. Eight studies (57.1%) reported significantly lower results for the football players than for the controls in at least one test (2.8±2.7 tests, average±1SD). Most frequently, deficits of attention, executive function and memory were noted (Table 3 and Figure 2).

/* Figure 2 about here */

Studies reporting neurocognitive deficits had a higher rate of inappropriate control of type-1 errors (OR=17.35 [95%-CI=10.61-28.36]) and higher odds for inappropriate controls (OR=1.72 [1.22-2.43]) than studies not reporting any differences. The fraction of female players was about the same for both groups (OR=0.82 [0.47-1.47]). The fraction of younger...
players (youth/high-school/college) compared to more elderly players was larger in studies with negative findings than in those with significant deficits (OR=1.92 [1.38-2.68]) (Table 2).3).

**Impact of heading on neurocognitive functions**

A potential link between heading-exposure and performance on neurocognitive testing was analyzed in 17 studies (n=1173, 26.4% females). On average 8.5±5.1 neurocognitive tests covering 4.4±1.8 categories were obtained (Table 3). Two thirds of the studies did not find any relation between heading-frequency and neurocognitive test-performance, while six studies (35%) reported a correlation in 2.0±1.1 tests. Deficits of attention, memory and intelligence were most frequent (Table 4). The quality of heading-frequency assessment was lower for those studies reporting a link than for those studies without such a link (OR=14.2 [9.0-22.4]). The rate of inappropriate control of type-1 errors was similar in studies confirming or discarding such a link (OR=0.74 [0.53-1.04]). The fraction of players with more extensive exposure (amateurs, university, (former) professionals) was significantly higher amongst those studies reporting neurocognitive deficits than for those studies that did not observe deficits (OR=2.15 [1.63-2.85]).

**Impact of head-injuries on neurocognitive functions**

A potential association between previous head-injuries and neurocognitive deficits was investigated in 13 studies (n=1103, 26.6% females) (Table 3). On average, 11.3±4.9 neurocognitive tests covering 5.2±1.4 categories were obtained. Seven studies (54%) reported a correlation, with abnormalities noted in 2.4±1.8 tests. Deficits of visuospatial functions, decision-making, attention and executive function were most frequent (Table 4). The rate of inappropriate control of type-1 errors was lower in studies with positive findings compared to studies with negative findings (OR=0.20 [0.15-0.26]).
Data on previous head-injuries were available in ten studies with average numbers of concussions ranging between 1.0 and 2.1, with the most recent event between 6-8 months and several years ago. The assessment of previous head-injuries was based on players’ reports in all but two studies. Only football-related concussions were considered in 5/7 studies with positive findings and in 3/6 studies with negative findings, while the remaining five studies included other, non-football-related concussions as well or did not further specify.

Neuroimaging studies

We identified eight studies (n=143, 15.4% females) using imaging modalities focusing on brain structure (conventional MRI, VBM, DTI) or brain metabolism (fMRI, MRS). DTI (2 studies, n=49), conventional MRI (2 studies, n=44) and VBM (2 studies, n=25) were most frequently applied (Table 5). All studies used a case-control design with selection of controls rated as “appropriate” in six (75%). Most players were professionals (active=56; former=26) or amateurs (n=37). Only two studies were prospective. In one prospective study no conventional-MRI changes could be depicted in professional players (observation period=5 years). Prospectively observing female high-school players over one season using fMRI, significant reductions in frontotemporal cerebrovascular reactivity persisting up to 4-5 months after the season had ended were reported, resembling the pattern described in mTBI. Retrospectively, in former professionals, VBM demonstrated cortical thinning in the right inferolateral-parietal, temporal and occipital cortex and MRS showed higher choline and myo-inositol levels in the posterior-cingulate gyrus. In professional players, DTI indicated widespread white-matter abnormalities (albeit no changes in fractional anisotropy), but conventional MRI did not demonstrate changes related to the years of football-participation. In college-football players, VBM showed decreased grey-matter density and volume within the anterior-temporal cortex.
Four studies linked neuroimaging with neurocognitive data.\textsuperscript{17, 37, 38, 51} Cortical thinning was associated with worse performance on $\frac{1}{6}$ tests (Rey-Osterrieth complex-figure long-delay recall),\textsuperscript{37} Glutathion-levels were linked to inferior results in $\frac{1}{4}$ tests (trail-making-test B),\textsuperscript{51} and lower levels of fractional-anisotropy in parieto-occipital areas were associated with $\frac{1}{6}$ tests (poorer memory).\textsuperscript{17} In the only prospective study, neither changes in neurocognitive performance nor in conventional MRI could be depicted over an observation period of five years.\textsuperscript{38}

A link between heading-exposure and structural/metabolic neuroimaging-changes was investigated in five studies.\textsuperscript{17, 37, 44, 51, 54} Lifetime estimates of heading-numbers were inversely correlated with cortical thickness in the right parietal/occipital lobes\textsuperscript{37} and with myo-inositol and glutathione levels.\textsuperscript{51} Fractional-anisotropy levels in temporo-parietal white-matter were inversely correlated with the annual number of headings.\textsuperscript{17} A high cumulative head-acceleration exposure was linked to more profound reductions in cerebrovascular reactivity, outlasting the end of the season by 4-5 months before returning to baseline by month eight.\textsuperscript{54} No correlation between career heading-exposure and abnormalities on conventional MRI were reported in another study.\textsuperscript{44} The potential impact of remote head-injuries on brain structure was examined in two studies, both demonstrating no association.\textsuperscript{17, 44}

**EEG-studies**

Two studies used EEG in active (n=69)\textsuperscript{55} and former (n=37)\textsuperscript{56} professional male players. In both studies standard EEG-recordings were examined by a clinical neurophysiologist and EEGs were classified as “normal”, “slightly abnormal” or “abnormal” based on background-activity and alpha-activity. EEG-ratings in the players were compared to those in age-matched men of “various occupations”. With information on matched physical activities lacking in the controls, their quality was rated “inappropriate”. The rate of EEGs considered normal was lower in active and former players compared to controls. Amongst the
active players, all abnormal EEGs were observed in players who considered themselves as non-headers. Among former players, there were no EEG-differences between headers and non-headers.

Postural stability

One study (n=15) reported on balance, using the balance-error-scoring-system (BESS). This study described no significant differences between players and controls.
DISCUSSION

With the recently issued ban for heading in child football players, \(^4\) in the US, \(^26\) the ongoing debate about potential persistent effects of football and football-related (subconcussive) trauma on brain function received increased attention and caused uncertainty amongst football players, medical staff and media. Given the worldwide popularity of football, \(^52\) football-related health-issues may have far-reaching implications that have to be balanced and compared to benefits due to regular activity. This emphasizes the need to intensify hypothesis-driven research and the study of associations between football-play and persistent structural/functional changes of the brain.

Under-representation of female players

For most aspects evaluated, female players were in a minority, consistent with reportedly lower numbers of active female football players. \(^62\) While no conclusions could be drawn on football-related changes in neuroimaging and EEG, women were under-represented in studies that reported neurocognitive impairment compared to those not observing such deficits. This observation was unexpected since the rate of football-related head injuries was reportedly higher in women. \(^63-66\) Of note, none of the studies reported on (former) professional female players. Also, for studies reporting on neurocognitive testing, female players were over-represented in lower levels of play (youth, high-school/college) compared to higher levels (university, amateur, (former) professional) (OR=28.57 [19.25-42.41]). These observations suggest that cumulative exposure to football play or cumulative intensity has been lower in female players, not reaching levels that may be necessary to result in brain abnormalities. Future studies should pay special attention on functional/structural brain changes in female players with more extensive football exposure.
Neurocognitive testing in (former) football players

Applied in 77% of studies, neurocognitive testing remains the most common approach to investigate potential associations between football play and changes in brain function. Most studies dealt with effects of heading (74%) and head-injuries (57%). Over 60 different neurocognitive tests were used, most of them only in few studies. With all three key domains (attention, executive function and memory) assessed by 78% of studies, risk for false-negative results due to inappropriate selection of neurocognitive test domains seems low. Even for the most frequently used tests, in those domains considered most important in patients with (sub)concussive brain injury (supplement 4), the fraction of abnormal test results was low (0.29±0.18). This suggests that reported neurocognitive impairments were rather subtle and their detection may have depended on study-specific parameters as age, gender, level of play and selection of controls. Also, among all tests administered in a given study, those with abnormal outcome were infrequent (fraction=0.21±0.27). In 19/23 studies more than one neurocognitive test was applied to evaluate a single category (e.g. TMT-B and Stroop for executive functions). Noteworthy, in 42% of these studies discrepant test results in a given category were noted. This affected 37% (17/46) of all categories in studies that received multiple testing. These results suggest that changes are subtle and may get identified only by some tests. Moreover, these inconsistencies underline the importance of standardized neurocognitive testing in football.

Persistent neurocognitive changes in (former) players compared to controls

Fifty-seven percent of case-control studies reported persistent associations between football and neurocognitive impairment focusing on attention, executive function and memory. These categories are primarily mediated by the frontal and temporal lobes and are typically involved in mTBI. Based on the quality-assessment performed, several confounders must be considered to put the significance of these observations in context.
Probably of the most far-reaching implication is the finding that the rate of appropriate control for type-1 errors was smaller among studies with abnormal test results (OR=\textit{17.35 [10.61-28.36]}). These studies thus bear an increased risk for false-positive test results. Choosing the right controls is essential. Including control subjects without matching the profile of physical activity, might point to global effects of physical—activity rather than to football-related changes.\textsuperscript{14} While in our review controls were judged as “appropriate” in 67\% of case-control studies, the odds for inappropriate controls were higher for studies reporting neurocognitive deficits (OR=\textit{1.72 [1.22-2.43]}). This indicates that inappropriate selection of control subjects may represent a serious source of bias. As a consequence, caution is warranted when interpreting impairment in neurocognitive testing based on the existing literature. Noteworthy, with the fraction of younger players being larger in studies with negative findings than in those with significant deficits, this might suggest that the exposure duration was simply not long or intense enough to cause a significant effect, further limiting conclusions.

\textit{No clear evidence for heading-related persistent impairment of neurocognitive function}

Six of 17 studies that correlated heading-frequency with neurocognitive deficits reported a link (mostly for attention, executive functions, and memory), but these studies also contained more methodological limitations than those reporting no link. Most importantly, the assessment-quality of heading-frequency was lower in studies with positive findings (OR=\textit{14.2 [9.0-22.4]}). Self-reported heading-frequencies tend to be higher than those obtained by more reliable approaches,\textsuperscript{68} indicating potential risk of reporting bias. This emphasizes the need for prospective observer-based assessments of heading-frequency in future studies. Furthermore, studies so far remained incomplete in providing an accurate estimate of heading-exposure, since heading during practice sessions was not considered and other variables such as heading-technique and ball-properties and -velocity were not available. Studies focusing on former players with heading-exposure decades ago\textsuperscript{45 56 57} may also be biased by differences in
the properties of the ball – heavy leather balls, in common use until the mid-1970s and even heavier on wet undergrounds, should not be compared with the more lightweight balls used thereafter.

The ratio of more senior to more junior players was higher in studies reporting significant impact of heading than in those with negative findings (OR=2.15 [1.63-2.85]). For more senior players the lifetime heading-number is higher and the duration of exposure is longer. Current studies therefore cannot exclude that in more senior players neurocognitive deficits may eventually arise due to a decreased cerebral reserve-capacity in view of accumulated (sub)clinical head-trauma. Of note, in retired professional UK- football players no signs for accelerated cognitive decline were found. Furthermore, female players were under-represented in studies that noted a link between heading-frequency and neurocognitive impairments. A lower heading-frequency in female players and the skewed distribution of female players towards lower levels of play may explain this seemingly “protective effect” of female gender on heading-related neurocognitive impairments.

Based on our review, no firm conclusions on an association between heading-frequency and accelerated neurocognitive decline can be drawn. Methodological limitations identified more often in studies with positive findings emphasize caution in linking heading to persistent neurocognitive deficits. This extends conclusions of a previous review on acute and persistent effects of concussions and heading in football. Whether or not significant differences in neurocognition arise with increasing heading-exposure, awaits further clarification, especially in former professional players with extensive exposure and better control for head-injury.

Repetitive concussions may be linked to persistent cognitive impairment

A negative impact of (repetitive) head-injuries to brain function has been reported for other contact sports as American football, rugby or boxing. For football (soccer)
Barnes and colleagues estimated a 50%-risk that a professional male player will suffer a concussion within a 10-year period, while the corresponding figure for female players was 22%.\textsuperscript{26} Neuropathological changes associated with mTBI are axonal injuries with a focus on the orbito-frontal and temporal-polar zones.\textsuperscript{67} Cognitive functions affected most are delayed memory, executive functions, language and attention.\textsuperscript{77,78} Based on our review, 54% of studies addressing the influence of head-injuries on neurocognitive test-performance found a link in one or more categories. Seventy-one percent of studies with positive findings restricted their analysis to football-related concussions, which indicates a low-risk that football-unrelated concussions biased the results. Furthermore, the rate of studies with inappropriate control of type-1 errors was even lower amongst studies with positive findings compared to those with negative findings.

Distinguishing between effects secondary to heading and head-trauma,\textsuperscript{11,32} however, may be difficult or even impossible for several reasons: First, up to 50% of concussions are not reported by players or team physicians.\textsuperscript{63,79,80} Second, 89% of studies controlled only for recent (3-6 months) head-trauma or did not take head-trauma into account at all. Most studies relied on self-reporting of head-injuries, introducing risk of recall bias. Even more importantly, lack of control for head-injuries bears the risk that effects of heading and head-trauma are mixed as players with higher heading-frequencies tend to experience head-injuries more frequently.\textsuperscript{14,19,32} Third, definitions for football-related head-injuries have been applied differently,\textsuperscript{52} potentially resulting in inconsistencies between studies. Assuming that effects of heading and head-injury may have been intermingled, one would expect overestimating the link between heading-frequency and neurocognitive deficits. In fact, the opposite was true; refuting the assumption that head-injury related effects have significantly biased results in the assessment of heading-related persistent effects. In summary, the link between head-injuries and persistent neurocognitive impairment was moderate only and its impact may have been over-estimated due to several methodological shortcomings.
Structural and metabolic changes on neuroimaging

Neuroimaging in football players was driven by the hypothesis that repetitive subconcussive head-trauma result in structural/metabolic changes similar to those known from mTBI. Along with the anterior-posterior gradient in brain vulnerability, anterior regions may also be linked to executive and neurocognitive deficits in mTBI.

With a limited number of study subjects (n=143) and studies (n=8), different neuroimaging–modalities and levels of play, all but two studies reported significant brain changes in football players. These changes were localized preferentially to frontal and anterior-temporal regions. Structural abnormalities located in parieto-occipital areas, i.e., in areas opposite the presumed point of heading impact, may be explained in analogy to the principle of contre-coup injury.

In three studies with 63 players, changes depicted on neuroimaging could be linked to subclinical neurocognitive deficits, suggesting that these changes may be functionally relevant. While these studies used advanced neuroimaging, no changes could be observed on conventional MRI in the only prospective study (five-year observation period). Whereas there is extensive experience in the interpretation of standard MR-sequences, changes in advanced imaging as DTI, VBM and MR-spectroscopy are much more difficult to put into clinical context due to their relatively recent use.

A correlation between heading-exposure and persistent changes on neuroimaging was observed in four out of five studies, albeit reversible within eight months after season end in the only prospective study. While these findings suggest a possible link between heading-frequency and neuro-degeneration, the heading-exposure assessment was low-quality in four out of five studies, posing them at increased risk for recall bias. Correction for multiple statistical testing was reported by Svaldi and in one study by Koerte, but not in another, indicating possible increased risk of false-positive correlations. In the study by
Lipton, a link was observed only for players with more than 885-1500 headings per year, suggesting that heading below this threshold may be safe.\textsuperscript{17} Overall, taking the methodological limitations and the small sample sizes into consideration, support for a link between heading-frequency and persistent structural brain changes seems weak. Our review did not find any evidence for an association between (repetitive) concussions and structural brain changes on neuroimaging, albeit small sample sizes limit conclusions.\textsuperscript{17,44}

**Low-quality of EEG-based studies**

Both studies included reported higher incidences of EEG-abnormalities in male professional players than in controls.\textsuperscript{55,56} With only 106 players analyzed, conclusions on persistent effects of football on EEG can be considered preliminary \textit{only}. Caution is warranted as significant limitations were identified: control groups were rated “inappropriate” and the authors did not control for remote head-injury. The clinical implications of the higher incidence of EEG-abnormalities remain unclear, especially since no information on the location of EEG-abnormalities was provided. Both studies did not provide any evidence for a link between heading and EEG-changes as in the group of active football players all abnormal EEGs were observed in players who considered themselves as non-Headers.\textsuperscript{55} \textit{In the group of former players no EEG-differences were detected comparing headers and non-Headers.}\textsuperscript{56}

Nonetheless, the authors concluded that the reason for the EEG-abnormalities was most likely repeated minor head-trauma. Considering the limitations listed above, evidence in support of this conclusion is unconvincing.
Limitations

Studies varied in the criteria for reporting previous concussions and we were lacking details to validate the ratings. Self-reporting was standard in most studies, posing them at high-risk for recall bias. Therefore the impact of previous concussions on brain structure/function may have been over- or underestimated in our review. The impact of heading depends on many parameters. However, only scarce information about player’s position, circumstances of heading (purposeful, passively hit) and heading-skills could be retrieved. This limits the assessment of a possible association between heading and structural/functional brain changes. Lack of systematically controlling for timing and time lags between exposure and testing potentially weakens reported associations. Likewise failure to assess and/or correct for potential confounders as concussions unrelated to football, different physical-activity profiles, medical conditions (hypertension, overweight, diabetes) and life-style (e.g., alcohol consumption, smoking) may have biased associations between (former) football play and structural/functional brain changes.

A pooled analysis of individual study data (i.e., a meta-analysis) was not possible due to the heterogeneity in study design, data collection and data analysis. This may have resulted in the incorrect detection or masking of high-quality studies due to a higher number of low-quality studies.

We did not identify any studies that compared the diagnostic accuracy of different neurocognitive-testing procedures that assessed the same neurocognitive domains. Moreover, single studies obtaining more than one neurocognitive test for a specific domain often demonstrated discrepancies. This limits any recommendation on specific neurocognitive tests and emphasizes the need for prospective, controlled studies comparing the diagnostic accuracy of neurocognitive tests.
CONCLUSIONS

There is weak to non-existent evidence from the medical literature for football-related persistent functional and structural brain deficits and a putative role of (repetitive) head trauma in the development of neurocognitive impairment. Comparison of included studies was limited by various methodological shortcomings. Case-control studies reporting neurocognitive deficits in football players significantly more often included inappropriate controls and control of type-I errors than studies reporting no deficits. Furthermore, no clear link between heading-frequency and neurocognitive deficits could be established and low-quality assessment of heading-frequency was identified as the major confounder. Of special interest were studies that combined different modalities: while in four out of five neuroimaging studies structural and metabolic deficits could be correlated with heading-exposure, the clinical and preventive implications of these findings remain inconclusive, as most studies used a low-quality assessment of heading-frequency. In three out of four small case-control studies a link between neuroimaging abnormalities and subclinical neurocognitive deficits could be established, suggesting that these morphological and metabolic changes might be functionally relevant.

Further studies combining functional and structural modalities in larger numbers of football players with long-lasting football-exposure appear most promising to shed more light on a potential link between football play and brain structure/function. Such studies should include neurocognitive testing of attention, executive functions and memory as well of objective tests of cervical, vestibular and ocular motor function. They should also be prospective to appropriately control for confounders as history of head-injuries, heading-frequency and medical conditions. Further validation and head-to-head comparison is required to provide the basis of more standardized testing batteries to improve the quality of studies and to allow for better comparability between studies. A longitudinal and cross-sectional study design will help to determine whether identified subclinical structural and functional
abnormalities eventually progress to clinically relevant, symptomatic deficits or rather resolve again, especially after finishing exposure to football play.
FIGURES

Figure 1

Records identified through database searching (n = 4105)

Records after duplicates removed (n = 2191)

Records screened (n = 2191)

Records excluded (n = 2079)

Full-text articles assessed for eligibility (n = 112)

Studies included in qualitative synthesis (n = 32) *

Studies included in quantitative synthesis (n = 30)

Full-text articles excluded, with reasons (n = 80)

Not football-related (n=34, 43%)
Not persistent (n=29, 36%)
Abstract only (n=10, 13%)
Not about heading/head-trauma (n=4, 5%)
<5 cases (n=2, 3%)
No data (n=1, 1%)

Searches conducted
August 2, 2016

Database results
Medline = 1092    Embase = 1450
PsycINFO = 285   CINAHL = 532
Cochrane CRCT = 24  SportDiscus = 714
Cochrane DSR = 8

Number of duplicate records
N = 1919
TABLES
<table>
<thead>
<tr>
<th>Citation, year</th>
<th>Study design, level of play</th>
<th>Control group</th>
<th>Sample size (test/control group) (n)</th>
<th>Performance level</th>
<th>Observation / exposure duration to competitive football play (mean±1SD)</th>
<th>Focus of study</th>
<th>NOS rating (&quot;good&quot;, &quot;fair&quot;, or &quot;poor&quot;)</th>
<th>Controls appropriate / appropriate control of type I errors / response rate (%)</th>
<th>High-quality assessment of heading frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ameau et al. 2002</td>
<td>Retrospective, case-control</td>
<td>College tennis players 31 / 31, m</td>
<td>≥2 seasons (college level)</td>
<td>NCP football players vs. controls</td>
<td>Fair</td>
<td>Yes / no / 62/220 (28%)</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adams et al. 2007</td>
<td>Retrospective, case-control</td>
<td>College non-football playing College students 10 / 10, m</td>
<td>≥8 years</td>
<td>Grey-matter density and volume (VBM)</td>
<td>Fair</td>
<td>No / yes / NR</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adams et al. 2002</td>
<td>Retrospective, case-control</td>
<td>College / pro</td>
<td>22 / 22, f/m</td>
<td>NCP football players vs. controls</td>
<td>Fair</td>
<td>No / no / NR</td>
<td>No (Heading-exposure-index)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emsberger et al. 2007</td>
<td>Retrospective, cross-sectional</td>
<td>University N/A</td>
<td>6-8 months since concussion, 15 years of play</td>
<td>Impact of concussion on NCP</td>
<td>Fair</td>
<td>N/A / yes / NR</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forbes et al. 2014</td>
<td>Prospective, cohort</td>
<td>N/A</td>
<td>19 / 21 months since (last) concussion, 9 years of play</td>
<td>NCP football players vs. controls, impact of concussion</td>
<td>Fair</td>
<td>No / yes / NR</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Guskiewicz et al. 2002</td>
<td>Retrospective, case-control</td>
<td>College non-football contact sports (n=96), no sports controls (n=53) 91 / 149, f/m</td>
<td>≤5 years</td>
<td>NCP football players vs. controls</td>
<td>Fair</td>
<td>No / yes / NR</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ida et al. 2002</td>
<td>Prospective, cohort</td>
<td>Youth N/A</td>
<td>9 months (prospective)</td>
<td>Impact of heading on NCP</td>
<td>Good</td>
<td>N/A / unclear / NR</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Taffe et al. 2006</td>
<td>Retrospective, case-control</td>
<td>Active pro</td>
<td>80 / 80, m</td>
<td>Structural abnormalities (MRI)</td>
<td>Good</td>
<td>Yes / yes / 20/25 (80%)</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kemp et al. 2016</td>
<td>Prospective, case-control</td>
<td>Active pro</td>
<td>Various non-contact sports athletes from local university 92 / 33, m</td>
<td>NCP football players vs. controls, structural abnormalities (MRI)</td>
<td>Fair</td>
<td>Yes / yes / NR</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kevette et al. 2015</td>
<td>Retrospective, case-control</td>
<td>Active pro</td>
<td>former non-contact pro athletes 11 / 14, m</td>
<td>Brain biochemistry (MRS), correlation with NCP</td>
<td>Good</td>
<td>Yes / no / NR</td>
<td>No (self-reported)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kevette et al. 2015</td>
<td>Retrospective, case-control</td>
<td>Active pro</td>
<td>former non-contact pro athletes 11 / 14, m</td>
<td>Brain biochemistry (MRS), correlation with NCP</td>
<td>Good</td>
<td>Yes / no / NR</td>
<td>No (self-reported)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kontos et al. 2011</td>
<td>Prospective, cross-sectional</td>
<td>Youth N/A</td>
<td>Bw. 6.7±3.0 and 8.5±4.6 years</td>
<td>Impact of heading on NCP</td>
<td>Good</td>
<td>N/A / yes / NR</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lamont et al. 2013</td>
<td>Retrospective, case-control</td>
<td>Amateur N/A</td>
<td>≥5 years</td>
<td>White matter microstructure (DTI), correlation with NCP</td>
<td>Good</td>
<td>N/A / yes / NR</td>
<td>No (self-reported)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Matser et al. 2010</td>
<td>Retrospective, case-control</td>
<td>Active pro</td>
<td>elite non-contact sports athletes 53 / 27, m</td>
<td>NCP, impact of heading/head injuries</td>
<td>Good</td>
<td>Yes / no / NR</td>
<td>No (self-reported)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Matser et al. 2009</td>
<td>Retrospective, case-control</td>
<td>Amateur</td>
<td>33 / 22, m</td>
<td>NCP football players vs. controls</td>
<td>Good</td>
<td>Yes / no / 33/33 (100%)</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Matser et al. 2003</td>
<td>Retrospective, case-control</td>
<td>Amateur</td>
<td>swimming/track 33 / 22, m</td>
<td>NCP football players vs. controls</td>
<td>Good</td>
<td>Yes / no / 109/156 (70%)</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rutherford et al. 2005</td>
<td>Retrospective, case-control</td>
<td>University rugby players (n=17), non-contact sports (n=24) 22 / 41, m</td>
<td>≥5 years</td>
<td>NCP football players vs. controls, impact of head injuries</td>
<td>Good</td>
<td>Yes / yes / 165/254 (65%)</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rutherford et al. 2009</td>
<td>Retrospective, case-control</td>
<td>University rugby players (n=30), non-contact sports (n=22) 25 / 52, m</td>
<td>≥5 years</td>
<td>NCP football players vs. controls, impact of head injuries</td>
<td>Good</td>
<td>Yes / yes / 193 / 337 (55%)</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stephens et al. 2010</td>
<td>Retrospective, case-control</td>
<td>High school rugby players (n=27), non-contact sports (n=16) 54 / 43, m</td>
<td>3.4±1.0 years</td>
<td>NCP football players vs. controls, impact of head injuries</td>
<td>Fair</td>
<td>Yes / yes / 193 / 337 (55%)</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>Design</td>
<td>Population</td>
<td>Gender</td>
<td>Age</td>
<td>Duration</td>
<td>Exposure/Outcome</td>
<td>Prognosis</td>
<td>N/A/no</td>
<td>NCP</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>---------------</td>
<td>-----------------------------------</td>
<td>--------</td>
<td>-----</td>
<td>----------</td>
<td>---------------------------------------------------------------------------------</td>
<td>------------</td>
<td>--------</td>
<td>------</td>
</tr>
<tr>
<td>Straume et al. 2005</td>
<td>Retrospective-cross-sectional</td>
<td>Active pro</td>
<td>N/A</td>
<td>271 / NA, m</td>
<td>Btw. 2.8±2.4 and 3.3±2.4 years at highest league</td>
<td>Impact of heading/head injuries on NCP</td>
<td>Good</td>
<td>N/A / no</td>
<td>271/300 (90%)</td>
</tr>
<tr>
<td>Straume et al. 2009</td>
<td>Prospective-cohort</td>
<td>Active pro</td>
<td>N/A</td>
<td>144 / N/A, m</td>
<td>1 year (i.e., 1 season)</td>
<td>Impact of concussion on NCP</td>
<td>Good</td>
<td>N/A / yes</td>
<td>NR</td>
</tr>
<tr>
<td>Svaldi et al. 2016</td>
<td>Prospective-case-control</td>
<td>High school athletes from various non-collision sports</td>
<td>14 / 12, f</td>
<td>1 season plus 3-4 months</td>
<td>Cerebrovascular reactivity related to head acceleration (heading)</td>
<td>Good</td>
<td>Yes / yes</td>
<td>NR</td>
<td>Yes (head acceleration event monitoring)</td>
</tr>
<tr>
<td>Straume et al. 2009</td>
<td>Retrospective-case-control</td>
<td>Active pro</td>
<td>N/A</td>
<td>69 / 69, m</td>
<td>128 games in pro league (avg)*</td>
<td>EEG disturbances</td>
<td>Fair</td>
<td>No / unclear</td>
<td>NR</td>
</tr>
<tr>
<td>Tysvaer and Stoggl 1989</td>
<td>Retrospective-case-control</td>
<td>Former pro</td>
<td>men with various occupation</td>
<td>37 / 37, m</td>
<td>359 games in pro league (avg)*</td>
<td>EEG disturbances</td>
<td>Fair</td>
<td>No / unclear</td>
<td>43/50 (66%)</td>
</tr>
<tr>
<td>Tysvaer et al. 1991</td>
<td>Retrospective-case-control</td>
<td>Former pro</td>
<td>Men from “different occupational groups”</td>
<td>37 / 20, m</td>
<td>359 games in pro league (avg)*</td>
<td>NCP football players vs. controls; impact of heading</td>
<td>Fair</td>
<td>No / no</td>
<td>NR</td>
</tr>
<tr>
<td>Tysvaer and Lachen 1993</td>
<td>Retrospective-case-control</td>
<td>Former pro</td>
<td>patients hospitalized for various disorders</td>
<td>37 / 20, m</td>
<td>359 games in pro league (avg)*</td>
<td>EEG disturbances</td>
<td>Fair</td>
<td>No / no</td>
<td>NR</td>
</tr>
<tr>
<td>Van Jones et al. 2014</td>
<td>Retrospective-case-control</td>
<td>Former pro</td>
<td>UK-based MCI prevalence study*</td>
<td>92 / N/A, m</td>
<td>14.5±3.2 years of pro play</td>
<td>NCP football players vs. controls</td>
<td>Fair</td>
<td>N/A / no</td>
<td>138/300 (46%)</td>
</tr>
<tr>
<td>Wood et al. 2002</td>
<td>Retrospective-case-control</td>
<td>Inter-scholastic, amateur/pro</td>
<td>60 / 12, m</td>
<td>Btw. 9.1±5.0 and 13.9±6.5 years</td>
<td>NCP football players vs. controls; impact of heading</td>
<td>Fair</td>
<td>No / no</td>
<td>NR</td>
<td>No (self-reported)</td>
</tr>
<tr>
<td>Zhang et al. 2013</td>
<td>Retrospective-case-control</td>
<td>High school</td>
<td>non-football players (high school level)</td>
<td>12 / 12, f</td>
<td>5 to 15 years (median: 8y)</td>
<td>NCP football players vs. controls; impact of heading</td>
<td>Fair</td>
<td>Unclear / unclear</td>
<td>NR</td>
</tr>
</tbody>
</table>

Abbreviations: avg=average; btw=between; CT=computed tomography; DTI=diffusion tensor imaging; EEG=electroencephalography; MRI=magnetic resonance imaging; MRS=magnetic resonance spectroscopy; N/A=not available; NCP=neurocognitive performance; NOS=Newcastle-Ottawa Scale, NR=not reported; pro=professional; VBM=voxel-based morphometry.

* These studies did not provide years of play, but games played. Assuming a total of 30-60 games per season, in all cases more than one season of professional play can be assumed.
Table 2: summary information about included studies (n=30)

<table>
<thead>
<tr>
<th>Control of type-1 errors</th>
<th>Studies (n, [%])</th>
<th>Female football players (n [%])</th>
<th>Male football players (n [%])</th>
<th>All football players (n [%])</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inappropriate / unclear</td>
<td>18 [60.0]</td>
<td>275 [72.6]</td>
<td>898 [68.4]</td>
<td>1173 [69.4]</td>
</tr>
<tr>
<td>Total</td>
<td>30 [100.0]</td>
<td>379 [100.0]</td>
<td>1312 [100.0]</td>
<td>1691 [100.0]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Selection of controls</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Appropriate</td>
<td>12 [40.0]</td>
<td>12 [11.1]</td>
<td>224 [54.2]</td>
<td>236 [43.3]</td>
</tr>
<tr>
<td>N/A</td>
<td>10 [33.3]</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Total</td>
<td>30 [100.0]</td>
<td>108 [100.0]</td>
<td>412 [100.0]</td>
<td>521 [100.0]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Response rate*</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>High (&gt;50%)</td>
<td>7 [23.3]</td>
<td>0 [0.0]</td>
<td>462 [35.2]</td>
<td>462 [27.3]</td>
</tr>
<tr>
<td>Low (≤50%)</td>
<td>3 [10.0]</td>
<td>0 [0.0]</td>
<td>135 [10.3]</td>
<td>135 [8.0]</td>
</tr>
<tr>
<td>Not reported</td>
<td>30 [100.0]</td>
<td>379 [100.0]</td>
<td>1312 [100.0]</td>
<td>1691 [100.0]</td>
</tr>
<tr>
<td>Total</td>
<td>30 [100.0]</td>
<td>379 [100.0]</td>
<td>1312 [100.0]</td>
<td>1691 [100.0]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gender</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Football players</td>
<td>N/A</td>
<td>379 [22.4]</td>
<td>1312 [77.6]</td>
<td>1691 [100.0]</td>
</tr>
<tr>
<td>Control subjects</td>
<td>N/A</td>
<td>108 [20.7]</td>
<td>413 [79.3]</td>
<td>521 [100.0]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Category†</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NCT</td>
<td>23</td>
<td>365</td>
<td>1153</td>
<td>1751</td>
</tr>
<tr>
<td>Case-control studies</td>
<td>14† [18.2]</td>
<td>29 [38.2]</td>
<td>52 [67.2]</td>
<td>104 [66.0]</td>
</tr>
<tr>
<td>Impact of heading</td>
<td>17 [27.2]</td>
<td>310</td>
<td>863</td>
<td>1459</td>
</tr>
<tr>
<td>Impact of head-injuries</td>
<td>13‡ [21.9]</td>
<td>188</td>
<td>810</td>
<td>980</td>
</tr>
<tr>
<td>Neuroimaging</td>
<td>8‡ [13.8]</td>
<td>22</td>
<td>121</td>
<td>138</td>
</tr>
<tr>
<td>EEG</td>
<td>2‡ [3.4]</td>
<td>0</td>
<td>106</td>
<td>106</td>
</tr>
<tr>
<td>Postural control</td>
<td>1‡</td>
<td>0</td>
<td>15</td>
<td>6‡</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level of play‡</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Youth</td>
<td>3 [10.0]</td>
<td>69 [18.2]</td>
<td>100 [7.6]</td>
<td>169 [10.0]</td>
</tr>
<tr>
<td>Professional</td>
<td>8 [26.7]</td>
<td>0 [0.0]</td>
<td>677 [51.6]</td>
<td>677 [40.0]</td>
</tr>
<tr>
<td>Total</td>
<td>30 [100.0]</td>
<td>379 [100.0]</td>
<td>1312 [100.0]</td>
<td>1691 [100.0]</td>
</tr>
</tbody>
</table>

Abbreviations: EEG=electroencephalography; N/A=not available; NCT=neurocognitive testing.

* Fraction of subjects that agreed to participate after being invited to participate in the study. This includes both football players and control subjects.

† Some studies provided testing for more than one modality (e.g. neurocognitive testing and neuroimaging or neuroimaging and balance testing), resulting in a total study number larger than 31. Within neurocognitive testing, also some studies provided case-control data as well as a correlation analysis for e.g. heading-frequency and neurocognitive deficits in the football players.

‡ Level of play as reported in the original studies. Age-range for youth football players was 10-13 and 13-18. The category “High-school / college” includes the “interscholastic” as well.
Table 3: Studies reporting on neurocognitive testing (NCT) in football players *

<table>
<thead>
<tr>
<th></th>
<th>Studies with significant differences in NCT†</th>
<th>Studies without significant differences in NCT</th>
<th>All studies</th>
<th>Stats</th>
<th>Studies with significant impact on NCT †</th>
<th>Studies without significant impact on NCT</th>
<th>All studies</th>
<th>Stats</th>
<th>Studies with significant impact on NCT †</th>
<th>Studies without significant impact on NCT</th>
<th>All studies</th>
<th>Stats</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control of type-1 errors (n studies / n subjects [%])</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Selection of controls (n studies / n control subjects [%])</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appropriate</td>
<td>5 / 155 [35.8]</td>
<td>10 / 380 [65.4]</td>
<td>28 [100]</td>
<td></td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Inappropriate/unclear</td>
<td>3 / 104 [3.0]</td>
<td>1 / 293 [0.3]</td>
<td>3 [100]</td>
<td></td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Response rate (n studies [%])</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High (&gt;50%)</td>
<td>3 / 23 [13.0]</td>
<td>1 / 109 [0.5]</td>
<td>4 [17.4]</td>
<td></td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Low (≤50%)</td>
<td>0 / 96 [0.0]</td>
<td>2 / 263 [0.8]</td>
<td>2 / 100</td>
<td></td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Not reported</td>
<td>5 / 200 [2.5]</td>
<td>7 / 295 [2.4]</td>
<td>7 / 100</td>
<td></td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Football players according to gender (n subjects [%])</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender controls (n subjects [%])</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Females</td>
<td>27 [9.3]</td>
<td>69 [25.9]</td>
<td>96 [34.4]</td>
<td></td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Males</td>
<td>218 [70.7]</td>
<td>171 [64.1]</td>
<td>389 [13.7]</td>
<td></td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>245 [100]</td>
<td>138 [100]</td>
<td>383 [100]</td>
<td></td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Sample size (mean±SD)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Football players</td>
<td>33±17</td>
<td>53±22</td>
<td>65±26</td>
<td></td>
<td>48±22</td>
<td>80±35</td>
<td>61±36</td>
<td>p=0.371</td>
<td>59±43</td>
<td>116±101</td>
<td>85±78</td>
<td>p=0.208</td>
</tr>
<tr>
<td>Controls</td>
<td>3±5±3</td>
<td>27±22</td>
<td>30±26</td>
<td></td>
<td>32±22</td>
<td>55±35</td>
<td>47±38</td>
<td>p=0.114</td>
<td>75±44</td>
<td>132±104</td>
<td>106±74</td>
<td>p=0.162</td>
</tr>
<tr>
<td>NCTs (mean±SD)</td>
<td>9.1±5.4</td>
<td>8.2±5.4</td>
<td>8.7±5.5</td>
<td></td>
<td>11±5.5</td>
<td>7.2±4.6</td>
<td>8.5±3.4</td>
<td>p=0.773</td>
<td>14±4.1</td>
<td>5.2±3.1</td>
<td>10±1.4</td>
<td>p=0.086</td>
</tr>
<tr>
<td>Level of play (n studies / n subjects [%])</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Youth</td>
<td>0 / 0 [0.0]</td>
<td>0 / 0 [0.0]</td>
<td>0 / 0 [0.0]</td>
<td></td>
<td>0 / 0 [0.0]</td>
<td>3 / 169 [19.1]</td>
<td>3 / 169 [19.1]</td>
<td></td>
<td>3 / 169 [19.1]</td>
<td>3 / 169 [19.1]</td>
<td>3 / 169 [19.1]</td>
<td></td>
</tr>
<tr>
<td>Amateur</td>
<td>1 / 33 [3.0]</td>
<td>0 / 0 [0.0]</td>
<td>1 / 33 [3.0]</td>
<td></td>
<td>1 / 37 [12.8]</td>
<td>0 / 0 [0.0]</td>
<td>1 / 37 [12.8]</td>
<td></td>
<td>1 / 37 [12.8]</td>
<td>0 / 0 [0.0]</td>
<td>1 / 37 [12.8]</td>
<td></td>
</tr>
<tr>
<td>Former professional</td>
<td>0 / 0 [0.0]</td>
<td>0 / 0 [0.0]</td>
<td>0 / 0 [0.0]</td>
<td></td>
<td>0 / 0 [0.0]</td>
<td>0 / 0 [0.0]</td>
<td>0 / 0 [0.0]</td>
<td></td>
<td>0 / 0 [0.0]</td>
<td>0 / 0 [0.0]</td>
<td>0 / 0 [0.0]</td>
<td></td>
</tr>
</tbody>
</table>

*OR = OR ratio, OR = odds ratio, N = number, † = p-value, ‡ = p-value, § = p-value, †† = p-value, †‡ = p-value, †‡† = p-value, †‡‡ = p-value, †‡§ = p-value.
Abbreviations: CI=confidence interval; NCT=neurocognitive testing; NR=not reported; OR=odds ratio including 95% confidence intervals.

* Studies may have provided data both for neurocognition in football in general and specifically related to heading-frequency and/or the number of head-injuries.

† In these studies statistically significant (p<0.05) differences (between cases and controls or related to heading-frequency or the number of head-injuries) in at least one neuropsychological test were reported.

‡ Odds ratio (OR; including 95% confidence intervals in brackets) comparing fractions of a given condition (e.g., appropriateness of type-1 errors or gender) between studies with or without significant differences in NCT.

§ Controls were considered appropriate if the following criteria were met: 1) age- and gender-matched, 2) physical activities of comparable intensity but without body contact, i.e., swimming, running or cycling, 3) no additional (recreational) exposure to football or other contact sports.

¶ This includes both football players and control subjects.

|| Two-sample t-tests using Bonferroni correction for multiple testing.

** Odds ratio for high-quality heading-assessment (prospective recorded or combined prospective and retrospective approach) vs. low-quality heading-assessment (self-reported heading rates or calculation of heading-exposure-risk based on player’s position).

†† Odds ratio between studies with or without significant differences in NCT comparing younger (youth, high school, college, interscholastic) and more elderly (university, amateur, professional, former professional) football players.
### Table 4: Distribution of NCT categories (in alphabetical order) and percentage of abnormal tests

<table>
<thead>
<tr>
<th>Category</th>
<th>Case-control studies (n=14)</th>
<th>Studies reporting on heading-frequency (n=17)</th>
<th>Studies reporting on the number of head-injuries (n=13)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Studies with sign. effect / all studies reporting abnormal NCT in at least one category (n=8) [%]</td>
<td>Studies with sign. effect / all studies reporting abnormal NCT in at least one category (n=6) [%]</td>
<td>All Studies with sign. effect / all studies reporting abnormal NCT in at least one category (n=7) [%]</td>
</tr>
<tr>
<td>Abstract reasoning</td>
<td>0/3 [0]</td>
<td>0/2 [0]</td>
<td>0/2 [0]</td>
</tr>
<tr>
<td>Creativity</td>
<td>0/2 [0]</td>
<td>0/0 [0]</td>
<td>0/2 [0]</td>
</tr>
<tr>
<td>Decision making</td>
<td>0/0 [0]</td>
<td>0/0 [0]</td>
<td>0/3 [0]</td>
</tr>
<tr>
<td>Language</td>
<td>0/4 [0]</td>
<td>0/3 [0]</td>
<td>0/6 [0]</td>
</tr>
<tr>
<td>Motor skills</td>
<td>1/3 [33]</td>
<td>1/3 [33]</td>
<td>1/3 [33]</td>
</tr>
<tr>
<td>Visuospatial functions</td>
<td>2/7 [29]</td>
<td>2/5 [40]</td>
<td>0/2 [0]</td>
</tr>
</tbody>
</table>

Abbreviations: NCT = neurocognitive testing; sign.=significant.

* Studies may have provided data both for neurocognition in football in general and specifically related to heading-frequency and / or number of head-injuries.

| Categories with ≥50% of studies with abnormal NCT results are bold. |
Table 5: Persistent effects of football on the brain: neuroimaging studies (n=58)

<table>
<thead>
<tr>
<th>Effect confirmed (n, [%])</th>
<th>No effect (n, [%])</th>
<th>All (n, [%])</th>
</tr>
</thead>
<tbody>
<tr>
<td>Studies (n)</td>
<td>6 [25.0]</td>
<td>2 [25.0]</td>
</tr>
<tr>
<td>Selection of controls (n studies / n subjects [%]) *</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appropriate</td>
<td>4 / 52 [83.9]</td>
<td>2 / 37 [100.0]</td>
</tr>
<tr>
<td>Inappropriate / unclear</td>
<td>1 / 10 [16.1]</td>
<td>0 / 0 [0.0]</td>
</tr>
<tr>
<td>N/A</td>
<td>1 / N/A</td>
<td>0 / N/A</td>
</tr>
<tr>
<td>Total</td>
<td>6 / 62 [100.0]</td>
<td>2 / 37 [100.0]</td>
</tr>
<tr>
<td>Gender football players (n subjects [%])</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Females</td>
<td>22 [22.2]</td>
<td>0 [0.0]</td>
</tr>
<tr>
<td>Males</td>
<td>37 [77.8]</td>
<td>44 [100.0]</td>
</tr>
<tr>
<td>Total</td>
<td>59 [100.0]</td>
<td>44 [100.0]</td>
</tr>
<tr>
<td>Gender controls (n subjects [%])</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Females</td>
<td>12 [19.4]</td>
<td>0 [0.0]</td>
</tr>
<tr>
<td>Males</td>
<td>50 [80.6]</td>
<td>37 [100.0]</td>
</tr>
<tr>
<td>Total</td>
<td>62 [100.0]</td>
<td>37 [100.0]</td>
</tr>
<tr>
<td>Sample size (mean±1SD)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Football players</td>
<td>16.5±10.2</td>
<td>22.0±2.8</td>
</tr>
<tr>
<td>Controls</td>
<td>12.4±2.1</td>
<td>18.5±2.1</td>
</tr>
<tr>
<td>Type of imaging (n studies / n football players [%])</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conventional MRI</td>
<td>0 / 0 [0.0]</td>
<td>2 / 44 [100.0]</td>
</tr>
<tr>
<td>Diffusion-tensor imaging (DTI)</td>
<td>2 / 49 [49.5]</td>
<td>0 / 0 [0.0]</td>
</tr>
<tr>
<td>Voxel-based MR-morphometry (VBM)</td>
<td>2 / 25 [25.3]</td>
<td>0 / 0 [0.0]</td>
</tr>
<tr>
<td>MR-spectroscopy (MRS)</td>
<td>1 / 11 [11.1]</td>
<td>0 / 0 [0.0]</td>
</tr>
<tr>
<td>Functional MRI (fMRI)</td>
<td>1 / 14 [14.1]</td>
<td>0 / 0 [0.0]</td>
</tr>
<tr>
<td>Total</td>
<td>6 / 99 [100.0]</td>
<td>2 / 44 [100.0]</td>
</tr>
<tr>
<td>Level of play (n studies / n football players [%])</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Youth</td>
<td>0 / 0 [0.0]</td>
<td>0 / 0 [0.0]</td>
</tr>
<tr>
<td>High school / college</td>
<td>2 / 24 [24.2]</td>
<td>0 / 0 [0.0]</td>
</tr>
<tr>
<td>University</td>
<td>0 / 0 [0.0]</td>
<td>0 / 0 [0.0]</td>
</tr>
<tr>
<td>Amateur</td>
<td>1 / 37 [27.4]</td>
<td>0 / 0 [0.0]</td>
</tr>
<tr>
<td>Professional</td>
<td>1 / 12 [12.1]</td>
<td>3 / 44 [100.0]</td>
</tr>
<tr>
<td>Former professional</td>
<td>2 / 26 [26.3]</td>
<td>0 / 0 [0.0]</td>
</tr>
<tr>
<td>Total</td>
<td>6 / 99 [100.0]</td>
<td>2 / 44 [100.0]</td>
</tr>
<tr>
<td>Correlating neuroimaging with NCT (n studies)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Structural and NCT changes correlate</td>
<td>3 [23.1]</td>
<td>1 [8]</td>
</tr>
<tr>
<td>No correlation</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>N/A</td>
<td>3 [23.1]</td>
<td>1 [8]</td>
</tr>
<tr>
<td>Correlating neuroimaging with heading-frequency (n studies)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Changes on neuroimaging and heading-frequency correlate</td>
<td>4 [26.4]</td>
<td>0</td>
</tr>
<tr>
<td>No correlation</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>N/A</td>
<td>5 [31.2]</td>
<td>1</td>
</tr>
<tr>
<td>Correlating neuroimaging with the number of head-injuries (n studies)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Changes on neuroimaging and number of head-injuries correlate</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>No correlation</td>
<td>1 [6.3]</td>
<td>1</td>
</tr>
<tr>
<td>N/A</td>
<td>3 [18.4]</td>
<td>0</td>
</tr>
</tbody>
</table>

Abbreviations: NCT=neurocognitive testing; N/A=not available

* Controls were considered appropriate if the following criteria were met: 1) age- and gender-matched, 2) physical activities of comparable intensity but without body contact, i.e., swimming, running or cycling, 3) no additional (recreational) exposure to football or other contact sports.
FIGURE LEGENDS

Figure 1
Flow chart depicting the selection process of identified articles. * One study was excluded because of duplicity of data, another study was excluded because of a “poor” risk-of-bias rating (Newcastle-Ottawa Scale).

Figure 2
Spider plots illustrating in how many studies the different neurocognitive categories were evaluated (in red) and how often an abnormal test result was retrieved in this category (in blue). The number of studies is provided along the intersection of the web. Separate plots are provided for all case-control studies (panel A), for all studies investigating the impact of heading on neurocognitive tests (panel B) and for all studies reporting on the impact of head injuries on cognition (panel C). Irrespective of the study categorization, attention, executive functions, memory, and visuospatial functions were the cognitive domains most frequently tested and also most frequently impaired.
REQUIRED STATEMENTS

Acknowledgements: We thank Dr. K. Alix Hayden (Libraries & Cultural Resources University of Calgary, Calgary, Canada) for optimizing the search strategy and performing the literature search.

Funding sources for this study: none.

Conflict of interest:
Dr. Tarnutzer reports no conflict of interest.
Prof. Brugger reports no conflict of interest.
Prof. Straumann reports no conflict of interest.
Dr. Feddermann-Demont reports no conflict of interest.
REFERENCES


Point-to-point reply to the reviewers’ comments

Reviewer: 1

Comments to the Author

Specific Review:
The manuscript was evaluated using the PRISMA guideline.

Introduction and abstract: There are no concerns with the abstract and introduction sections when using PRISMA. The “what is known” and “what does this study add” fits with the manuscript.

Methods: Overall the methods are described adequately. Including a detailed search strategy and figures illustrating the implantation of the exclusion criteria to find suitable original articles. Two points to consider are using PRISMA are:

• The database search was only limited to the MEDLINE database. This method risks missing original studies within other databases. Future studies could include several databases to strengthen the review.

Reply from the authors: For this revised version of the manuscript we have extended the number of databases searched. Besides MEDLINE we now take the following databases into account as well: Embase, PsycINFO, CINAHL, Cochrane CRCT, SportDiscus, Cochrane DSR = 4. This literature search was performed by Dr. A. Hayden, who is now acknowledged at the end of the manuscript. Note that Dr. Hayden also slightly modified and extended the search query to optimize functioning in the different databases and to cover the key interests of this review optimally. The refined literature search was performed for these additional six databases and was repeated for MEDLINE to cover the same period for all seven databases. Details about the search strategies for these different databases are now provided in the methods and in supplement 1.

Searching all seven databases resulted in 2192 unique citations (including 5 citations identified by screening the bibliography of included manuscripts). After title/abstract and full-text search, 30 manuscripts were included (after removal of one manuscript rated as “poor” quality in the Newcastle-Ottawa Scale and one manuscript due to duplicity of data). Furthermore, compared to the first version of the manuscript (reporting on 31 manuscripts), three studies (Kaminski et al. 2007; 2008; Colvin et al. 2010) were removed due to a too brief observation period (<6 months). By updating the database search (previous search performed in December 2015), 2 additional manuscripts were identified. By searching Embase, PsycINFO, CINAHL, Cochrane CRCT, SportDiscus, Cochrane DSR = 4 as well, one additional study that met the inclusion criteria was identified (Salinas et al. 2009).

Removal and addition of the manuscripts mentioned above did not significantly change the main findings in our review.
•The process of data extraction is stated however there is no mention of any process of confirming data between reviewers

Reply from the authors: The process of confirming data between reviewers is now stated: “Data extraction was performed by AAT and confirmed by a second author (NFD).

Results: Results seem clear with easy to read subtitles providing flow to the manuscript.

•Table 2: Is difficult to follow it is not clear that the Yes/No category in Table 2 refers to the results of the original studies i.e. do they show a deficit (yes) or no deficit (no). I would recommend adjusting this title to make this more explicit.

Reply from the authors: We thank reviewer 1 to pointing this out. We changed the title to “studies with significant differences in NCT” and “studies without sign. differences in NCT” for the case control-studies. For studies focusing on heading-frequency and head-injuries, we changed the categories to “Studies with significant impact on NCT” and “Studies without sign. impact on NCT”.

Also the first line of table 2 page 23, line 6. Under ‘studies’ the categories of Impact of heading-frequency on NCT and Impact of head-injuries on NCT do not have percentages.

Reply from the authors: Percentages have been added as suggested.

Discussion: The papers original aim area answered is discussed with relevance to other papers in the field. The manuscript addresses the limitations of the source papers well. Conclusion: clear with indications of potential further research in the area.

Recommendations: adjusting the method section to state whether there was a process of confirmation between reviewers during data extraction and revision of Table 2 to improve the clarity of the information.

Reply from the authors: We have implemented these changes as outlined in detail below in this revised version of the manuscript.
Reviewer: 2

Comments to the Author
This is a timely review on an important topic. The review has sought to bring some coherence to a wide body of literature on the long-term impact of heading and head injuries on subsequent brain structure and function. The review tackles a very comprehensive question however in doing so seems to lose direction at times.

There are a number of issues, which need clarification in order to show the data to its full potential.

Reply from the authors: We now provide detailed answers to all issues raised by the reviewer below.

1. A fundamental issue is the operational definition of "long-term". This is fundamental to the review and appears without justification or reference to any publication or other. This choice of cut-off value needs considerable justification, as it has the potential to confound the types of studies being reviewed.

Reply from the authors: Previously we have required minimal duration of observation (after a head injury) or exposure to football play (in either case-control studies or observational studies) of only 4 weeks. We agree that this may not be considered “long term” by many. We therefore have extended the minimal observational period to six months and at the same time have changed the wording. We now use the term “persistent” instead of “long-term” as for the latter one no consensus definition is available and some may require an exposure duration of more than 5 years to qualify for “long-term”.

There is no sharp definition of the term “persistent” either. However, in the context of the persistent postconcussive syndrome, most researches require a duration of 3-6 months to consider symptoms or findings “persistent” (see http://emedicine.medscape.com/article/828904-overview for details on commonly used time frames). For our review we required a minimal duration of either 6 months or a full season of play. In practice, this takes into account the rhythm of medical assessment in football, with annual pre-season and / or post-season assessments being used in part of the studies. An observation period of one full season therefore will be considered sufficiently long enough as well to assess persistent symptoms. For cross-sectional (i.e., retrospective) studies and retrospective case-control studies we required exposure to organized football play for at least six months or one full season as well. For studies included, average duration of exposure to organized football play was clearly above one season as reported in table 1 (new) in detail.

This is now specified in the methods section: “ “Persistent” changes were defined as changes that were still recognizable >6 months after a potential impact or linked to exposure to football play for >6 months or a full season. We adhered to the time-frame usually applied for the persistent post-concussive syndrome, albeit no consensus-based definition of this term exists.”

2. The introduction is overly long and could be reduced and include some mention of the methodological tools which are available to ensure the quality of the review. It is noted that issues study methodology etc have been raised but more specific attention needs to be paid to how they will be addressed.
Reply from the authors: We have edited, streamlined and slightly shortened the introduction and at the same time now put more attention to the applied methodological tools to obtain an estimate of risk-of-bias of included studies. We now use the Newcastle-Ottawa Scale (NOS; Wells et al.: The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp), which is a nine-item score designed for non-randomized trials including case-control studies and cohort studies. Its use is promoted by the Cochrane Collaboration (Cochrane handbook for systematic reviews of interventions version 5.1.0 [updated March 2011]. Edited by Higgins JPT, Green S. [http://www.cochrane-handbook.org]). We also evaluated the Downs and Black scale, however, opted for the NOS due to its more practical use and reasonable number of items to rate. The NOS was adjusted to reflect the key research questions asked in this review and therefore also takes the previously assessed items of history of head injury (i.e., study controls for presence / absence of previous non-football associated head injuries) and the response rate into account. Besides the NOS, we assess the quality of statistical analyses, how heading-frequency was determined and whether diagnosis of previous head-injuries was based on medical files or not.

3. Please comment on the reason why only one database was selected for searching.

Reply from the authors: For this revised version of the manuscript we have extended the number of databases searched. Besides MEDLINE we now take the following databases into account as well: Embase, PsycINFO, CINAHL, Cochrane CRCT, SportDiscus, Cochrane DSR = 4. This literature search was performed by Dr. A. Hayden, who is now acknowledged at the end of the manuscript. Note that Dr. Hayden also slightly modified and extended the search query to optimize functioning in the different databases and to cover the key interests of this review optimally. The refined literature search was performed for these additional six databases and was updated for MEDLINE to cover the same period for all seven databases. Details about the search strategies for these different databases are now provided in the methods and in supplement 1.

4. Key methodological detail is included as Supplementary material - all key information should be included in the Methods section of the manuscript.

Reply from the authors: We have extended the methods section as requested by the reviewer and we have also added a limitations section as requested by another reviewer. As a result, however, we are now about 400 words above the word limit of 4500 words.

5. In presenting the data it would be helpful to include an initial Table which identifies the individual studies, their characteristics and study quality etc. Some of this information is in the Suppl file but needs to be included in the core text as Table 1 so the reader is fully informed and can better understand the following data.

Reply from the authors: We now provide a core table as suggested by the reviewer (new table 1). This table summarizes the most important information about each study, including the focus of research, the level of play, the risk of bias and the duration of exposure to football or the observation period after an event (concussion).
6. Was the correction for multiple testing the only quality assurance measure used?
Reply from the authors: In order to assess the quality of the obtained data, several quality assurance measures were used, including the correction for multiple testing. Other measures were use of objective approaches to determine the number of headings (i.e., done prospectively by independent observer), reporting of the fraction of subjects that participated in relation to all subjects invited to participate and how data on history of head injuries was gathered (based on medical files or only on the participants’ self-report). Some of these aspects are now provided in a more structured and standardized fashion using the Newcastle-Ottawa Score.

7. Studies in the various Tables need to be identified. This could be facilitated by the establishment of a core Table 1 as mentioned above.
Reply from the authors: We now provide a core table (new table 1) as suggested, summarizing all key information as e.g. appropriateness of controls, type-1 error correction of assessment of heading frequency. We also provided more references to specific studies in the other tables, however, opted against extensive use to avoid overloading of tables (especially for table 3). We do feel confident that the newly added table 1 will help the reader obtain additional information as needed.

8. It would be useful to have more detail on the "appropriate/inappropriate" designations. The information provided is somewhat superficial considering that this is a key element of the manuscript.
Reply from the authors: We rated all studies with respect to how information about 1) heading frequency and 2) history of concussion was obtained and whether the statistical analysis took into account the presence of multiple statistical tests. Furthermore (4) the selection of controls was judged as well. In the revised version of this manuscript, items 1 and 4 are part of the NOS and have been prospectively defined before the ratings were performed by two reviewers. Criteria for low or high risk of bias for each item are provided in detail in the supplementary file 2. A summary of this approach is provided in the methods section when describing the NOS.

Regarding the correction for multiple statistical tests, we required studies to explicitly state that measures were taken to compensate for multiple statistical tests. If no such measures were reported or if it was stated that no correction was applied, we rated these studies as high risk for type 1 error. This is now explained with more detail in the methods section.

9. Table 2. Can you please clarify how the designation "Yes/No" was determined. Are the ORs presented for all data or just for studies considered to have "appropriate" level of control?
Reply from the authors: Any study that reported significant differences in neurocognitive test performance either when comparing cases and controls or in relation to heading frequency or history of head injuries was labelled as “yes”. All studies not demonstrating any differences in neurocognitive test performance between the different groups were labelled “no”. Note that
these labels (yes/no) have been modified as suggested by reviewer 1. Furthermore, we now provide a footnote explaining the requirements for meeting these categories:

“In these studies statistically significant (p<0.05) differences (between cases and controls or related to heading-frequency or the number of head-injuries) in at least one neuropsychological test were reported.“

The ORs compare specific aspects of studies with significant differences and studies with no significant differences, e.g. the frequency of appropriate control of type-1 errors. This is now specified in a footnote:

„ORs compared fractions of a given condition (e.g., appropriateness of type-1 errors or gender) between studies with or without significant differences in NCT.“
Reviewer: 3

Comments to the Author
Thank you for the opportunity to review this manuscript. The topic and rationale for the review are well thought out and would be of value to the readers of BJSM. The study aim was to review the literature on long-term effects of football on brain function and structure and the impact of heading and head injuries.

Overall the manuscript is well written and easy to follow. The introduction is clear and succinct but there are limitations in the methodology that could significantly affect the results and subsequent conclusions. Without these addressed I think the study does not deliver on the potential full aims of a study of this nature.

Firstly the reliance on only one database (Medline-PubMed) would be seen as a limitation especially for study with such extensive search topics/categories and the inclusion of both psychological and technical data alongside the medical literature.

Reply from the authors: Based on the reviewers’ concern that restricting the literature search to a single database, we have extended the search to these additional databases: Embase, PsycINFO, CINAHL, Cochrane CRCT, SportDiscus, and Cochrane DSR = 4. The search queries for these databases were adjusted by Dr. A. Hayden, librarian, University of Calgary. Key elements of the search strategies applied for these different databases can be found in the methods section and in full detail in the online-only supplement 1.

Secondly, the authors have based their risk of bias assessment on a previous authors finding of methodological inadequacies in the literature some 13 years previous (Rutherford et al. 2003). Limiting the risk of bias evaluation to the categories of selection controls, heading assessment and control of type-I errors while at face value is relevant, does not fully evaluate the methodological vigour of each paper reviewed from my perspective. I would have expected a risk of bias assessment would include an industry recognised standard tool that had been assessed for reliability and validity to be used as a benchmark of which there are many available. From there, modifications could have been made to include other methodological inclusions such as listed above to make it more specific to the review topic if necessary and justified. A risk of bias assessment using appropriate cut-off scores could then be used to grade studies for further inclusion in the analysis as opposed to using individual bias criterion’s to grade data after the results are known (i.e. positive/negative finding).

Reply from the authors: We now use the Newcastle-Ottawa Score (NOS) as standardized methodological tool to assess risk of bias in included studies. The NOS has been promoted by the Cochrane Collaboration as suitable tool to assess risk of bias in non-randomized studies including case-control studies and cohort studies. For our purposes we have slightly modified some of the nine items and also designed a modified version of the NOS for cross-sectional studies (not originally supported by the NOS). Such adjustments to the specific review topic are promoted for the NOS. Overall quality of studies was rated as good, fair or poor according to a previous publication (McPheeters et al. 2012). Studies rated as “poor” quality were excluded from the review. 15 studies were rated as “good”, 15 as “fair”, while one study rated as “poor” was excluded from the analysis.

Thirdly, (associated with the second issue above) and perhaps most importantly, I believe
there are significant limitations in the analysis of the literature presented in this review. Traditionally a meta-analysis should be used to pool the results of homogeneous studies. Instead the authors have chosen to group all studies of positive and negative findings and then analyse on individual bias criterions as counts (OR’s) of these groupings. This could lead to the incorrect detection or masking of high quality studies in the analysis due a higher number (count) of lower quality studies. Registering the review in an appropriate repository such as Prospero may have remedied some of the above methodological issues before the study was undertaken, however as it stands I don’t think I can truly interpret the study findings based on these methodological limitations.

Reply from the authors: Indeed pooling data to allow a meta-analysis of prespecified outcome parameters would have made this review more powerful. However, as pointed out by the reviewer, such data pooling is only possible for homogeneous studies. The studies available on football-related effects on the brain structure and function, however, did not meet criteria for such data pooling. Several factors have prevented from performing a meta-analysis: 1) study groups and controls were distinct, ranging from college students to former professional football players in others and to perform analyses in subgroups, to few studies could be identified. 2) The output parameters obtained were not uniform. In the case of neuropsychological testing, the set of tests applied was different in almost all studies, both regarding the specific tests used and the neuropsychological domains covered. In the case of neuroimaging, the analyses performed were distinct in all seven studies included, ranging from magnetic resonance spectroscopy, voxel-based morphometry to fiber tracking. 3) The level of granularity for reporting results was low for most studies, severely limiting data extraction for any meta-analysis. Obtaining sufficiently detailed numbers (i.e., on a single subject level) would have required requesting whole data sets from the authors of the studies included and was not feasible. Facing these limitations, we decided to perform a systematic review of the literature instead, using an assessment of quality to allow a rating on the impact of individual studies.

We now acknowledge this limitation at the end of the discussion section:
“A pooled analysis of individual study data (i.e., a meta-analysis) was not possible due to the heterogeneity in study design, data collection and data analysis in identified studies. This may have resulted in the incorrect detection or masking of high quality studies due to a higher number of low quality studies.”

In summary, without a more conventional risk of bias assessment and pooled analysis of individual results (where possible) I cannot fully trust the interpretation of the findings of this study in its current form. The study does have much merit and is likely to be well received if the authors can address the above methodological issues but I understand it will be a significant amount of work.

Reply by the authors: We thank the reviewer for this valuable input. As suggested, we have extended the basis of literature search (addressing point 1) and also now provide a more conventional risk of bias assessment by using the Newcastle-Ottawa Score (point 2). Regarding point 3, a careful analysis of the data available indicated that a pooled analysis of individual results was not feasible. This limitation is now discussed at the end of the manuscript.
Reviewer: 4

Comments to the Author

Thank you for the opportunity to review this important systematic review looking at the impact of playing football on brain structure and function. Not only is it important to understand injury risk associated with sport from a governance, injury prevention/safety perspective but there are also important legal and financial ramifications, noting the ongoing case (and payments already made) regarding the US national football league settlement case. The Jeff Astle Foundation have been very vocal in the media in England on the subject of this review and the aims of this study therefore have wide and important ramifications.

As the authors highlight, we lack valid and reliable methods for measuring exposure and 23/31 studies collect this retrospectively. Furthermore the clinimetric / validity of the methods for determining brain structure and function are also unclear. So when multiple different methods for measuring the unknown exposure and supposed/possibly related effects, it becomes evident that we cannot answer the aims of this study.

With so many different tests being used to assess brain structure and function is it possible to compare outcomes and draw conclusions? The main focus of this paper and the discussion should focus perhaps on the validity and reliability of the tests as these are the main determinants of any conclusions and will help improve study methods moving forwards.

Reply by the authors: Admittedly, the heterogeneity of identified studies puts restrictions to the conclusions that can be drawn and heavily influenced the recommendations for future studies proposed at the end of the manuscript. As suggested by the reviewer, we extended the discussion on the validity and reliability of the tests as applied in the studies identified. This was done both for the large number of neuropsychological tests and for neuroimaging studies.

The evidence for the different imaging approaches chosen is generally weak and the link that authors of these papers have tried to make between structural changes (e.g. by VBM, DTI, conventional MRI) and brain function (as assessed by neurocognitive test performance) is therefore not well established. We further emphasize these shortcomings.

The three most important domains in suspected mTBI are attention, executive functions and memory. These three domains were covered in 18 out of 23 studies reporting on neurocognitive test performance in football players. Most frequently used tests in these for domains were TMT-A (attention), TMT-B (executive functions) and Rey-Osterrieth complex figure recall (memory). More detailed information on this aspect is now provided in supplementary file 4 (Table 1 in this supplement). Based on these observations, future studies addressing functional brain changes in football players should at least include a single test from these domains. Based on our review, however, we cannot make recommendations for a specific test.

The title should not include the phrase 'long term' as the cut-off used to define this is described in the paper as 4 weeks. Many of the studies look at youth, high school and university subjects, (ages 10-18 years), so long-term inferences cannot be made and the title becomes very misleading and does not really reflect the paper content.
Reply by the authors: As pointed out by the reviewer, “long-term” may not be suitable to describe effects studied in some of the manuscripts included (e.g., with observation periods of only 6-12 months). We therefore replaced “long-term” with “persistent”. Symptoms or signs are usually considered “persistent” if present for more than three to six months. For our study, we required a duration of observation (if studies were prospective) or exposure (if studies were retrospective) for more than six months. In retrospective studies, this reflected the total duration of football play, either restricted to certain levels (e.g., first division or national team) or since onset of competitive football play (usually in youth). This is specified in the methods section.

Furthermore the title is quite vague as heading a ball and concussion are separate events (one is an injury and the other does not meet the definition of injury) and this review joins all these together in one review. If we are looking at the risks associated with playing football then this is fine, but to look at this and concussion together makes little sense. It should be one or the other as they are not the same. More importantly it may inadvertently mislead readers.

Reply by the authors: This review indeed focuses on the risks associated with playing football. We modified the title to better reflect this link: “Persistent effects of playing football and associated (subconcussive) head trauma on brain structure and function – a systematic review of the literature”.

There are numerous potential causes of structural and functional brain limitations in later life and we cannot differentiate between these potential factors regarding the findings within each study. Timing and time lags between heading/concussion and testing, along with controlling for all other factors that could potentially affect the testing methods need to be accounted for.

Reply by the authors: We agree with the reviewer that all these parameters should be controlled and accounted for in an optimally designed study. For the studies included in this systematic review, however, this was often not the case. In some studies such confounders were obvious, as for example for alcohol consumption (Matser et al. 1998; 1999). Such observations were stated in the notes section in supplement 3, summarizing all studies included in a table.

We now address this also more specifically in the limitations section: “Lack of systematically controlling for timing and time lags between exposure and testing potentially weakens reported associations. Likewise failure to assess and/or correct for potential confounders as concussions unrelated to football, different physical-activity profiles, medical conditions (hypertension, overweight, diabetes) and life-style (e.g., alcohol consumption, smoking) may have biased associations between (former) football play and structural/functional brain changes.”

Given what is known about head injury should men and women have been compared together or analysed separately.

Reply by the authors: In the results section and in table 3 we do address gender-specific aspects of head-injuries. We show that women are relatively under-represented in studies reporting significant associations between head injuries and neurocognitive impairment and
we also address this in the discussion section. While five studies included data from female football players, only two of those studies studied female players only. Furthermore, in these three studies with mixed-gender data, no separate analysis based on gender was provided. Therefore too few datasets were available to perform a meaningful analysis restricted to female players.

*If retrospective studies are even more problematic and lower quality then why include 23 of them?*

Reply by the authors: One important aim of this review was to provide an overview of the quality of studies available. As such, various limitations became evident when assessing these studies, as reflected in the scores obtained in the risk of bias assessment obtained (Newcastle-Ottawa Scale). This includes a retrospective study design in most studies as mentioned by the reviewer.

However, restricting our review to prospective studies was not deemed feasible to the heterogeneity of study populations, research questions asked and experimental paradigms applied (neuroimaging vs. neurocognitive testing vs. balance testing).

*This article is however an important step in trying to understand the long term risks of brain 'damage' through playing football. I would have like the authors to spend more focus on the specific tests used and to provide the readers/future researchers with guidance and solutions on how to realistically approach future research so that the type 1 errors and other major methodological flaws highlighted are not repeated ad infinitum*

Reply by the authors: We agree with the reviewer that guidance for future researchers and the readers regarding study design for studies addressing long-term effects of football play on brain structure and function are needed. A validation of specific neurocognitive tests (as e.g. the trail-making-test A and B or the Stroop test), however, is limited by their large number and varying combination in the different studies. While the use of computerized testing batteries (e.g., IMPACT 1.0 / 2.0, AMAN, CogTest) may help in standardizing testing procedures, currently few studies have implemented computer-based testing (6/30) and no single testing battery was used in more than 2 studies. We did not identify any studies that compared the diagnostic accuracy (i.e., combined sensitivity and specificity) of different neurocognitive testing procedures that assessed the same neurocognitive domains (i.e. TMB B vs. Stroop). Moreover, single studies obtaining more than one neurocognitive test for a specific domain often demonstrated discrepancies between such tests as mentioned in the discussion.

This is now acknowledged in the limitations section: “We did not identify any studies that compared the diagnostic accuracy of different neurocognitive-testing procedures that assessed the same neurocognitive domains. Moreover, single studies obtaining more than one neurocognitive test for a specific domain often demonstrated discrepancies. This limits any recommendation on specific neurocognitive tests and emphasizes the need for prospective, controlled studies comparing the diagnostic accuracy of neurocognitive tests.”

Therefore, we consider the evidence to give recommendations about the use of specific neurocognitive tests or computerized testing procedures as too weak. We do agree that validation and consensus for specific neurocognitive testing procedures are important goals for future research in this area. Currently used computerized testing procedures have not been
designed to pick up long-term neurocognitive changes, rather they were put together to detect acute deficits after concussion.

We do provide a recommendation, albeit less granular, stating that future studies should be prospective and longitudinal and should tightly control for confounders such as life-style factors (alcohol, smoking), medical conditions (especially when looking at retired football players) including hypertension, diabetes and overweight. We have added the recommendation that further validation and head-to-head comparison is required to provide the basis of more standardized, ideally computer-based testing batteries to improve the quality of studies and to allow for better comparability between different studies.

Abstract:

Aim/objective:

Line 7 - Not sure about word 'estimate' - this study is trying to determine if there is an association with structural/functional decline rather than estimation.

Reply by the authors: We have replaced “estimate” by “association” as suggested: “We reviewed the literature for associations between football activities (including heading and head-injuries) and decline in brain structure and function.”

Page 5:

Line 5 - football is not a contact sport. It is not supposed to be contact and the rules are written to try and ensure it remains non contact.

Reply by the authors: FIFA Rules in football distinguish between legal and illegal contact between players. While there are obviously different types of illegal contact including charges, jumps, kicks, pushes or strikes that will be sanctioned, there are also accidental and legal contacts. Legal contacts refer to shoulder-to-shoulder contact, while accidental contact may for example include head-to-head collision while heading the ball. Based on these rules, football is usually considered a contact sport (see e.g. the following essay: http://coachingamericansoccer.com/soccer-rules/soccer-is-a-contact-sport)

Line 10 - CTE and tau protein findings remain very controversial and should not be mentioned without explaining the reasons for it being controversial.

Reply by the authors: We have modified the sentence under discussion, now emphasizing more that CTE and tau protein are still controversial:

“Neurodegenerative disorders (such as Alzheimer disease) have been reported in retired professional football players and in athletes from other contact sports as rugby and American football. A postulated association between football play and chronic traumatic encephalopathy, however, remains controversial, and the effect of football-related concussions is not well understood.”

Lines 14-16 - It is worth explaining this statement about the banning of heading - what was the evidence base? If done through fear of litigation or with evidence based backup then very important to explain.
Reply by the authors: When U.S. Soccer issued a statement in November 2015 announcing an extensive player safety campaign, they also reported a resolution in a concussion litigation filed in 2014. Although it is stated that the player safety campaign was developed independently from the lawsuit, it is likely that it has been influenced by this lawsuit.

To take this possibility into account, we have modified the sentence under discussion:

“With raising concerns and facing a concussion litigation, the football federation of the USA issued in November 2015 a ban for heading in children aged 10 years or less and limited heading in children aged 11-13 years.”

Line 23 - need to define head injury? Are the ones in this study just those requiring treatment? What about the many that do not require treatment? This does not reflect the title of the paper.

Reply by the authors: We did not re-classify study subjects regarding the presence / absence of previous concussion but relied on the judgement of the study authors as details to allow a sound rating of single cases were missing. We did not require treatment as inclusion criteria (mostly because this information was often not available). Furthermore, criteria for head injury varied amongst studies. While some studies required loss of consciousness or amnesia, others did not or did not further specify. This is obviously a confounder. We now describe in more detail how the history of concussions was obtained (low risk vs. high risk of recall bias) and address this also in the limitations section.

Line 45 - frequency depends on position. Only have to look at match data at high level to determine. Amateur and professional players likely to differ greatly. Main issue with paper is that biomechanical/force studies of heading a ball ought to be mentioned as different headers produce different forces. It is not possible to say all headers in football are the same.

Reply by the authors: We agree with the reviewer that heading technique, player’s position and player’s experience/technical skills are relevant variables when assessing the impact of heading. Overall, players’ positions were reported in very few studies only and heading technique was not specified. Furthermore, none of the papers distinguished between purposeful heading and accidental heading. These aspects are now addressed in the limitations section at the end of the discussion:

“The impact of heading depends on the player’s position and heading skills, however, only scarce information about parameters as player’s position, circumstances of heading (purposeful, passively hit) and heading skills could be retrieved. This limits the assessment of a possible association between heading the ball and structural / functional brain changes.”

Note also that amateur players made a small fraction only (70 players out of 1691 players).

Line 52-54 - I am not sure this is relevant and could be excluded from the paper.

Reply by the authors: The reviewer is referring to the sentence: “Low-quality assessment of heading-frequency, inappropriate control for type-I errors and inappropriate selection of controls were identified as sources of bias.” We do think that referring to the key parameters
that were assessed imperfectly is an important conclusion of this review and tentatively suggest keeping this sentence.

The third paragraph is contentious as concerns were raised that heading a ball during the brain maturation years may cause more damage. However the frequency, forces and whole raison d'être of this logic is questionable, because most related factors are not understood from research. There are other sports such as boxing that do require punches to the head - worth drawing lessons from other sports to justify the reasons for this paper.

Reply by the authors: We have modified this paragraph, emphasizing more areas of uncertainty and also providing examples from other sports such as boxing. The paragraph now reads:

“Likewise, the impact of purposeful heading the ball to play and guide its direction – unique to football – on the brain has been debated. On average, players head the ball 1-16 times during a competitive football match, accumulating over a season to several hundred headings and to many thousand headings during a professional football career. This has raised concerns that heading may – similar to boxers receiving punches to the head – pose players at increased risk for “subconcussive” trauma potentially resulting in neuronal damage similar to that in repetitive concussions but not accompanied by overt symptoms. These considerations have led to uncertainty in football players and their (medical) attendants, albeit such a link is far from being established and the impact of parameters such as heading technique, player’s age and playing position remain unclear. Nonetheless, with raising concerns and facing a concussion litigation, the football federation of the USA issued in November 2015 a ban for heading in children aged 10 years or less and limited heading in children aged 11-13 years. Concerns that the maturing brain could be especially vulnerable to subconcussive head injury may have supported this decision.”

Paragraph page 6
Lines 17-34 - so there are many tests that have been used and no agreement on radiological function and structure. There are 40 definitions of concussion in the literature and no gold standard assessment, so what hope is there currently of determining abnormal (not reaching clinical disease definitions) brain function and structure from controls. What is the natural decline in both as we age?

Reply by the authors: We agree with the reviewer that there is currently no consensus on a set of neurocognitive tests or a specific, clinically relevant imaging protocol. Currently published imaging studies therefore can only be considered as pilot studies requiring further confirmation in larger, better characterized study groups. These limitations in imaging studies and associations made with neurocognitive test performance are now emphasized more in the discussion.

Regarding the natural decline in brain function and structure, normative values have not been established for the imaging protocols applied here (VBM, DTI, MRS). This question therefore cannot be answered based on small case-control studies.

Page 6 Line 50
This paper does not review the long term effects of football as we cannot measure exposure? It is looking at trends in football players at a set time in their lives. For this reason it does not
measure the impact of heading and head injuries. This paper looks at associations. To use the word 'effects' (line 50) implies causation.

Reply by the authors: the sentence under discussion has been modified; the word “effects” has been replaced by “association” as suggested:

“Against this background, we aimed to systematically review the literature on associations between football play and persistent changes in brain function and structure and the impact of heading-frequency and head-injuries in this context.”

I think the introduction is muddled as it tries to cover multiple complex subjects. The introduction needs to steer the reader to the content of the paper and look to other head injury specific sports for reasons rather than football. Areas of uncertainty need to be much clearer. I do not think the title of the paper should include heading and head injury without explanations of the two and definitions. Not all headers are the same. Are there as many headers in an under 13 game of football as a professional game? What about concussion and when a concussion occurred during a life?

Reply by the authors: We have extensively modified the introduction. As suggested by the reviewer, we now refer to other sports with increased risk for head injuries including boxing and American football. Furthermore, we put a focus on areas of uncertainty including effects of heading in relation to players age and the concept of subconcussive head injury secondary to repetitive purposeful heading of the ball. Indeed, heading frequency in youth players was lower than in professionals.

Regarding the title of the review with its referral to heading and head injuries, we now provide clear definitions of both terms. For heading, this refers to “Likewise, the impact of purposeful heading the ball to play and guide its direction – unique to football – on the brain has been debated.”\textsuperscript{10,11} For concussive head injury, we refer to the definition of the 4th concussion consensus conference statement: “Concussions (i.e., a subtype of mild traumatic brain injury (mTBI) without structural abnormalities on conventional CT or MRI)\textsuperscript{1} represent 1-5\% of all football-(soccer-) related injuries.\textsuperscript{2,3,4,5}.

Materials & methods:
Is a 4 week cut-off really 'long-term'?! This is very important as the title should not include 'long-term' based on a 4 week follow-up. This is misleading. Please change the title to reflect time period.

Reply by the authors: Admittedly, a cut-off of 4 weeks will be difficult to be called long-term. Noteworthy, duration of exposure and how it was defined depended on the study type. For retrospective studies (usually performed as cross-sectional studies) total exposure duration to football play (e.g. as a professional or amateur) can be considered (also when addressing effects of repetitive heading). For head injuries, time since the last concussion was reported by several studies and might be the most practical approach to quantify lag here. For prospective studies, the observation period between baseline testing and follow-up testing is essential as effects of football play during this period are investigated. To better reflect the main aim of this review – i.e. its focus on persistent effects of football play – we now require studies to provide an observation period lasting at least 6 months. As a result, two studies from Kaminski and co-workers (2007; 2008) with observation periods of only 2 to 3 months are now excluded from the systematic review.
The specific observation periods are provided in table 1 (new in the revised version of the manuscript). For most retrospective studies, total duration of exposure to football play was considered, being usually in the range of 5 or more years.

Is the inclusion of retrospective studies valid?

Reply by the authors: Ideally we would have considered only prospectively designed studies. However, due to the relatively few prospective studies anticipated, we did not require studies being prospective to be included. Obviously, by including retrospective studies as well, this may have posed these analyses at increased risk for bias. We tried to minimize this risk by providing standardized ratings for risk of bias by use of the Newcastle-Ottawa Score (NOS). When looking at the NOS for each study, generally prospective studies were rated as “good” twice as often as “fair”, while for retrospective studies, ratios (“good” vs. “fair” were about even with 11 vs. 13), confirming this observation.

Is the inclusion of studies on youth/children really long term?

Reply by the authors: We identified 2 studies reporting on youth football players (Kontos et al 2011 and Janda et al. 2002), while in all other studies participants were at least 15 years old on average (high-school, college or university students). We do agree with the reviewer, that these studies are likely not perceived as “long-term” observational studies. As mentioned above, we have replaced “long term” by “persistent” and now require a period of observation of at least six months. In these studies focusing on youth players, one study was prospective (Janda et al. 2002) with an observation period of 9 months. The other study (Kontos et al. 2011) was retrospective and players participated in competitive football for at least 6.7±3.0 years on average. These studies are therefore in a range of observation duration usually considered valied when addressing persistent symptoms or signs.

If more stringent study selection were used this paper might be more informative?

Reply by the authors: Regarding neuroimaging studies, all were retrospective but two, so requiring a prospective study design would have reduced to number of studies included strongly. Regarding studies focusing on neurocognitive testing procedures, only five were prospective in study design. We therefore see little space for using more stringent study selection criteria. Rather, we opted for less strict inclusion criteria but detailed reporting on potential bias and limitation. However, selection criteria are now somewhat tighter in the sense that we require an observation period of at least six months to be included. Whether more restrictive study selection criteria would have impacted or even changed our main findings, remains open.

How can we compare athletes aged 10-18 years and athletes in their 60's and imply long term follow-up? I am left wondering whether the eligibility criteria should be tighter to avoid confusion for readers.

Reply by the authors: These groups obviously represent the extremes regarding age of the spectrum of players reviewed here. While we have removed the term “long-term” as
suggested by the reviewer, we have also increased the minimal observation period required (now at least six months).

Supplementary file - page 35 line 53 - how was a decision made on papers by discussion?

Reply by the authors: Disagreements were resolved by discussed and adjunction of a third reviewer if needed. This is now stated: “Discrepancies in selection status and reasons for exclusion were settled between the two reviewers by discussion and adjunction of a third reviewer if needed.”

Why was a quality assessment/score not completed? Appreciate limitations of all quality assessment systems but best one available should be selected and used.

Reply by the authors: We now use the Newcastle-Ottawa Score, which has been recommended by the Cochrane Collaboration as suitable for risk of bias assessment of non-randomized studies. Detailed results are presented in supplementary file 2 and summary findings are reported in table 1 and in the results section of the manuscript. Overall quality of studies is then rated as good, fair or poor. Studies rated as “poor” quality were excluded from the review.

Why are studies looking at males and females not separated? They should be. Also age very relevant.

Reply by the authors: We would like to refer the reviewer to our answer provided further above when this issue was already raised by the reviewer. As mentioned, we felt that the data was too scarce to provide a gender-specific analysis (only two studies reporting on effects of head injuries in female football players only).

What about other sports - many studies from US - so how many play American football as well, which is a major risk factor? Concussion is much more frequent outside of sport than inside and head injuries in children/youth. How do the studies account for this?

Reply by the authors: Participation in other contact sports, such as American football or boxing was assessed by some studies and was considered an exclusion criterion (no participation in other contact sports) in some studies. However, information about possible exposure to other contact sports, for example in recreation, is lacking in many studies. While this indicates a possible risk of bias, such exposure may have rather been limited and well below the degree of exposure to football play while in an organized team. This limitation is now also addressed in the discussion section, emphasizing that undisclosed or unrecognized participation in other contact sports such as boxing or American football may have contaminated effects of football play on brain structure and function.

Discussion
Page 13, line 12
There are benefits to regular activity to balance and compare with risks.
Reply by the authors: We have rephrased the sentence under discussion (“Given the worldwide popularity of football, football-related health-issues may have far-reaching implications” in the sense that we also emphasize positive effects of football play.

“Given the worldwide popularity of football, football-related health-issues may have far-reaching implications that have to be balanced and compared to benefits due to regular activity.”

Page 16
This study does not in general look at long term follow-up.

Reply by the authors: We have replaced “long term” by the term “persistent” and do emphasize looking for associations rather than effects: “This emphasizes the need to intensify hypothesis-driven research and the study of associations between football play and persistent structural and functional changes of the brain.”

Page 16, lines 30-36
Or natural decline with age, CVD or smoking, or chronic disease or many many others factors not related to football. This is very speculative.

Reply by the authors: We agree with the reviewer that cardiovascular disease, smoking, chronic disease and natural decline with age are potential confounders as well. In general, we tried to emphasize these limitations more at the end of the discussion section.

Page 17, lines 32-38
Different in UEFA studies and clearly at different levels of game.

Reply by the authors: Here we refer to active professional football players only to give the reader an idea about the risk of football-related head injury in those exposed the most. We agree with the reviewer, that in players with less exposure to football play as amateur players this risk will lower. We emphasized in the discussion also, that duration of exposure may play an important role for developing persistent neurocognitive deficits.

Page 21 – Conclusion
First 2 sentences contradict each other. I would say the evidence is 'weak to non-existent' for a relationship. The methodological shortcomings are to a small extent determined by the broad inclusion criteria. It is not possible to compare results as the tests used are so varied and so unreliable - table 3.

Reply by the authors: We agree that evidence is weak to non-existent and have adjusted the sentence under discussion accordingly. We also emphasize now in the conclusions that validation of specific neurocognitive tests is needed in the future:
“Further validation and head-to-head comparison is required to provide the basis of more standardized testing batteries to improve the quality of studies and to allow for better comparability between studies.”
I would have liked to have seen a solutions based table with a summary of each test and the evidence base to support each test, because this study highlights that they cannot be compared.

Reply by the authors: We have identified over 60 different neuropsychological tests that covered a total of 9 domains in our systematic review. Most tests were applied in few studies (as shown in figure 1 in supplement 4). For specific domains, several tests were proposed, sometimes in the same study. Regarding persistent effects for (sub)concussive head injury, the following domains are most crucial: attention, memory and executive functions. We now provide more detail on the most widely used and possibly most suitable tests from these four domains, however, such recommendations are preliminary. There are no controlled studies comparing the diagnostic accuracy of different neurocognitive tests in football players.

Editor(s)' Comments to Author:

Associate Editor
Comments to the Author:
Thank you for submitting this manuscript to BJSM. The reviewers feel that it has potential but requires some substantial changes. Please review their comments especially with respect to your methodology including the searching of only 1 database and the lack of risk of bias assessment.

Reply by the authors: We have implemented an established risk of bias assessment tool (Newcastle-Ottawa Score) and have also increased the number of databases searched from one to seven. We feel confident that by applying these two main modifications the paper gained clarity and improved in reflecting strengths and weaknesses of existing studies in the field.