BMJ Open Sport & Exercise Medicine

Head injury assessment in rugby union: clinical judgement guidelines

Éanna Falvey ![ORCID],[1,2] Ross Tucker ![ORCID],[1] Gordan Fuller,[3] Martin Raftery[1]

ABSTRACT

Background/aim Clinical judgement is a recognised component of a complete off-field concussion assessment. This study identifies guidance criteria for team medical staff when using clinical judgement in their decision-making process during the World Rugby off-field concussion-assessment screen (HIA1).

Methods Retrospective study of examining doctor clinical judgement in 1149 HIA1 assessments after a meaningful head impact event completed on rugby union players participating in elite-level international and national competitions between September 2015 and June 2018. We assessed (1) an abnormal subtest result as worse performance compared with preseason baseline values; (2) the proportion of cases where clinicians overruled abnormal HIA1 assessment subtest results and (3) made recommendations on how clinical judgement decisions may be made more safely based on the accuracy of clinical judgement decisions assessed against the final concussion diagnosis.

Results One or more subtests were abnormal compared with baseline values in 857 of 1149 HIA1 assessments. Clinical judgement was used to return players to the game despite abnormal subtest results on 424 out of 857 occasions (49%). In a significant majority of cases 356/424 (84%), clinical judgement decisions were correct, with players later cleared of a concussion. An application of guided clinical judgement potentially decreased false negative assessments by 33% (21/63).

Conclusions Clinical judgement should be applied in the diagnosis of concussion but done so cautiously. We propose doctors should only use clinical judgement to overrule either one of; or a combination of (1) an abnormal tandem gait and (2) one abnormal cognitive test.

New findings

► Experienced clinical judgement is an essential component of a thorough off-field assessment for suspected concussion.
► In most cases, clinical judgement improves the diagnostic accuracy of the World Rugby Head Injury Assessment.
► The guidance for doctors aims to limit false positive tests without increasing the risk of false negative tests. Strict application of these guidelines would potentially reduce false negative results by 33%.
► Doctors should exercise extreme care when considering use of clinical judgement if a player has symptoms, a clinical sign any failed Maddock's question or more than one failed cognitive subtest.

[1]Medical Department, World Rugby Limited, Dublin, Ireland
[2]College of Medcine & Health, University College Cork, Cork, Ireland
[3]Centre for Urgent and Emergency Care Research, School of Health and Related Research, University of Sheffield, Sheffield, UK

Correspondence to
Dr Éanna Falvey;
Eanna.Falvey@worldrugby.org

INTRODUCTION

Concussion is a common and high-profile injury in contact and collision sports. Elite sports have introduced protocols to identify and manage head impact events with the potential for concussion during matches (based on quadrennial concussion in sport consensus statements).[1 2] These typically involve pitch-side or off-field testing with a multimodal screening instrument to detect possible concussion following suspected head impact events.[3] The Berlin statement emphasises that the Sports Concussion Assessment Tool (SCAT5)[4] represents the most

well-established screening tool that baseline testing may be useful to interpret SCAT5 subtest results, and that clinical judgement is required to determine the final return-to-play decision.[1]

World Rugby mandates that any suspected concussion should leave the field of play immediately and not return. At a community level, this is regulated by the 'Recognise and Remove' programme.[5 6] In premier elite competitions, where all staff are experienced and appropriately trained, pitchside video is available, there is an untoward event review system and player welfare standards are adhered to, an off-field assessment of meaningful head-impact events is available. The World Rugby head injury assessment protocol (HIA)[7] has been in use since 2014—suspected concussion is still permanently removed—facilitating a 12 min off-field assessment of a player who has had a meaningful head-impact event.

A recent investigation evaluated the diagnostic accuracy of off-field concussion screening assessments using the HIA1 off-field assessment (HIA1)—an abridged version of the SCAT5, comprising seven subtests.[8] This investigation revealed that if baseline values were to be strictly applied, theoretical screen sensitivity would be 89.6%, specificity 33.9% and area under the receiver operating curve

(AUROC) 0.62. In comparison, real-life removal from play decisions by team doctors using clinical judgement at the time of screening to overrule abnormal subtests, resulted in improved specificity (86.6%) with a slight decrease in sensitivity (76.8%) and an overall improvement in diagnostic performance, with an AUROC of 0.82.

Thoughtful application of clinical judgement is best practice in concussion management, and although overall HIA1 assessment performance improved statistically when clinical judgement was applied, there was a trade-off between a favourable reduction in false positives (removal from play and no concussion) and the undesired small increase in false negatives (returned to play after screening and postgame diagnosed with a concussion). Conservative management of concussion necessitates that the application of clinical judgement is optimised, specifically to ensure that excessive numbers of false negatives are not created by clinician's judgements that overrule abnormal subtests. Clinical and pitchside experience are crucial to this process, but while less-experienced doctors gain this knowledge we must endeavour to make their work as safe as possible.

Information on which abnormal HIA1 assessment subtests (compared with baseline) are most often correctly overruled by clinicians, and which abnormal subtest results are most predictive for concussion, may guide the application of clinical judgement and improve real-life HIA1 assessment performance.

This study extends the previously published World Rugby HIA process study by Fuller[5] and aims to investigate in detail how doctors perform (clinical judgement) when assessing specific subtest abnormalities. Specific objectives were to analyse the impact of clinical judgement when it was used to 'over-rule' various subtest abnormalities, and to identify scenarios where clinical judgement should be made only with extreme caution, since it may either reduce diagnostic accuracy, or may fail to improve diagnostic accuracy enough to offset the risk of false negatives.

METHODS
Study design and sample
This was a retrospective cohort study using prospectively collected data from the World Rugby HIA database. The source population comprised rugby players participating in elite level men's international (five competitions) and national competitions[9] between September 2015 and June 2018. A total of 370 doctors completed at least one of the HIA1 assessments in this period, with a mean of 4.95 (1–53) assessments per doctor. This period was chosen as no significant operational changes were made to World Rugby HIA processes over this time. The subsequent study population included all players identified during play with a meaningful head impact event, but unclear consequences, undergoing off-field screening for possible concussion. In the event of missing data available case analyses were performed to provide the final study sample. We examined the impact of the examining

doctor's clinical judgement on the diagnostic accuracy of the HIA1 assessment for each of these real-life events.

Procedures
HIA1 assessment
The word rugby HIA is a three-stage process for managing sport-related concussion that has been described in detail previously, with full details provided in the web online supplemental appendix A. The HIA1 assessment, a modified SCAT5 test, comprises subtests of immediate and delayed memory, Maddock's questions, digits backwards, tandem gait, a modified symptom checklist and clinical signs subtests. An abnormal subtest result is defined as a worse performance than pre-season baseline values. If the team doctor performed the HIA1 assessment, the independent match-day doctor (MDD) observed the process, in many situations the MDD performed the HIA1 assessment (this is at the team doctor's request).

Clinical judgement
It is possible for players to demonstrate worse performance than baseline on HIA1 assessment subtests and still be returned to play if, in the assessing doctor's clinical judgement, the subtest failure is felt not to be due to concussion, for example, slower tandem gait time due to a suspected ankle sprain, or expected random variation in subtest performance.

Subsequent concussion diagnosis
A final diagnosis of concussion is determined according to the judgement of the team doctor based on their clinical findings over the 48-hour period posthead impact, that is, the diagnosis of concussion could be made during the second assessment done within three hours (HIA2) or the third assessment (HIA3) done within 24–36 hours or anytime thereafter.

Data collection
HIA process data are routinely recorded at the point of assessment by assessing physicians using the tablet-based, web-hosted, CSx data platform. Player-specific baseline values are available to doctors within the CSx platform at the time of the HIA1 assessment. Data are subsequently uploaded to the World Rugby HIA database. HIA assessment forms, from each of the three HIA process stages, are linked using unique player identifiers.

Analyses
The overall diagnostic accuracy of the HIA1 assessment, for both real-life performance and theoretical performance with strict application of baseline subtest scores, has been previously published by Fuller *et al.*[7] This study extends this previous analysis to further explore particular subtests and clinical judgement scenarios, including their effect on HIA1 assessment performance. The analysis proceeded in three stages. First, a description of the derivation of the study cohort. Second, an examination of the frequency of application of clinical judgement to HIA1 assessment subtests. Third, a calculation of the

effect of clinical judgement to individual and combinations of HIA1 assessment subtests on test accuracy.

For each HIA1 assessment, we noted whether the player returned to play following the assessment. If the doctor had used clinical judgement to over-rule a subtest, we noted whether this decision matched the final diagnosis of concussion (determined at the HIA2 and HIA3 phase of the HIA process). We calculated overall accuracy as true positive and negative results divided by all screens.

We evaluated the effect of clinical judgement on HIA1 assessment performance by calculating a ratio for how much the overall accuracy of the HIA1 assessment changed when clinical judgement was applied. A ratio greater than one implies greater diagnostic accuracy when clinical judgement was applied, whereas a ratio less than one indicates that diagnostic accuracy was impaired. We applied this analysis to various exploratory scenarios for specific combinations of HIA1 assessment subtests. We calculated the accuracy of the HIA1 assessment for subtests independently and did not account for the presence of other subtests that may copresent with each subtest.

Sample size, statistics, ethics and funding
A census sample of cases undergoing HIA1 assessments over the study period were included. Following recommendation by Hopkins, and to avoid over conservative interpretation of findings, 90% CI were used to interpret whether changes in diagnostic accuracy were meaningfully different. The width of CIs indicates the precision of results. Statistical analyses were carried out in Stata V.13.1 (StataCorp). This was a post hoc analysis of data from a study protocol that received ethical approval from the University of Sheffield. All players and medical staff provided informed consent for participation prior to the start of the season. All data were anonymised. The study was funded by World Rugby.

RESULTS
Study sample
A total of 1149 consecutive HIA1 assessments were performed in 980 individual players (recurrent events occurring in 191 players, ranging from 146 players with 2, to 1 player with 9 events) during competitive Rugby matches over the study period. A final diagnosis and/or a baseline reference screen was absent in 98 cases, leaving a study sample of 1051 complete screens with baseline and a final diagnosis. Sixty-seven of the 98 incomplete screens were, however, still available for analysis of specific subtests not requiring a baseline reference limit (symptoms, Maddocks and Clinical signs), resulting in n=1118 for those specific subtest analyses. Of the 1051 complete cases, 434 players were diagnosed with concussion, giving a prevalence of 41.3% (90% CI 38.0% to 44.6%).

After excluding all cases where a final diagnosis was absent (n = 63, 28 in the compliant group and 35 in the clinical judgement group), a total of 794 HIA1 assessments produced at least one abnormal subtest result compared with baseline values (76%, figure 1). Clinical judgement was applied on 389/794 occasions (49%). A team-affiliated doctor most often assessed the player 658/1051 (63%), compared with the independent MDD, 393/1051 (37%). Derivation of the study sample for evaluation of the performance of clinical judgement is shown in figure 1.

Overall effect of clinical judgement on HIA1 assessment accuracy
Table 1 details HIA1 assessment results compared with baseline values, as well as how often players were
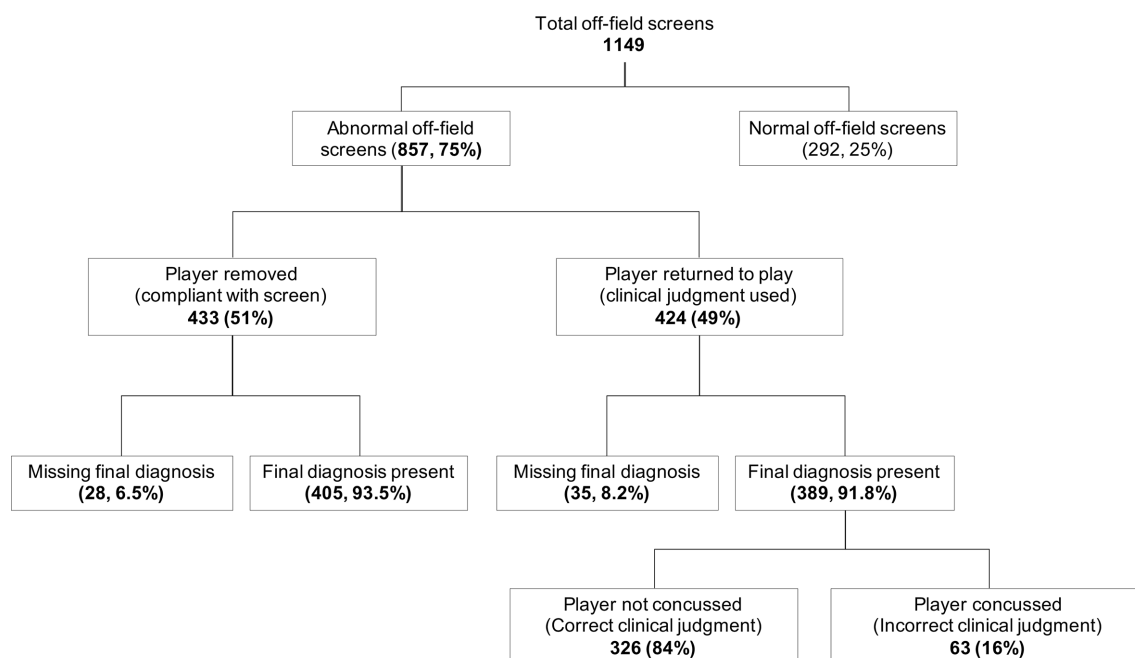

Figure 1 Derivation of study sample.

Table 1 Overall effect of clinical judgement on HIA1 assessment accuracy

Overall accuracy	Overall % (90% CI)
Theoretical accuracy (strict interpretation of subtests according to baseline value)	56.6 (54.1 to 59.1)
Real-life accuracy (taking clinical judgement into account)	82.5 (80.6 to 84.4)
Relative increase in overall accuracy when clinical judgement is applied	1.46 (1.33 to 1.59)
Theoretical true positive rate (sensitivity)	88.9 (86.5 to 91.4)
Real-life true positive rate (sensitivity)	77.0 (73.6 to 80.3)
Theoretical false positive rate (1—specificity)	66.1 (63.0 to 69.3)
Real-life false positive rate (1—specificity)	13.6 (11.3 to 15.9)

removed or returned to play. Overall accuracy of HIA1 assessment return to play decisions (ie, decision after HIA1 assessment results interpreted with clinical judgement) was 82.5% (90% CI 80.6% to 84.4%, n=1051). This was a 1.46-fold (90% CI 1.33 to 1.59) improvement in overall accuracy compared with theoretical diagnostic accuracy had no clinical judgement been applied, and subtest results had been strictly interpreted compared with baseline values (56.6%, 90% CI 54.1% to 59.1%). This improvement in accuracy was the result of a large reduction in the false positive rate (FPR) (from 66.1% for strict application of baseline comparisons, to 13.6% when clinical judgement is applied), partly offset by a smaller reduction in true positive rate (TPR) (from 88.9% to 77.0%) with clinical judgement (table 1). The practical outcome of clinical judgement was that 326 potential false positive cases (removed despite no concussion) were avoided, while 63 false negative cases (returned to play and later confirmed concussed) were created, compared with if baseline performance was strictly applied to all removal from play decisions.

Frequency of application of clinical judgement to HIA1 assessment subtests

There were 794 HIA1 assessments with one or more subtest results worse than baseline values and where a final diagnosis was present. Cases with specific abnormal subtest performances are shown in table 2. Tandem gait was the most common abnormal subtest (n=471/794) and Maddocks test the least frequently incorrect (n=56/794).

After excluding symptoms from the analysis of abnormal subtests, a single subtest performance was worse than baseline in 367 cases, with all subtests being abnormal in only two cases (table 3). Clinical judgement was applied in 71% (261) of cases where only a single subtest was abnormal. Clinical judgement was applied less frequently as the number of abnormal subtests increased, ranging from 39% (92/234) for two subtests to 0% when all subtests were abnormal.

The effects of clinical judgement to 'over-rule' abnormal HIA1 assessment subtest results differed for individual abnormal subtests and by the number of

Table 2 Description of clinical judgement applied to subtests within the HIA1 off-field screen

Specific abnormal subtests, irrespective of other subtests	Cases	Clinical judgement applied (% of cases)	Clinical judgement correct (% of clinical judgements)	Accuracy		Relative change
				Theoretical	Real life	
(A) Symptoms	261	13	64	74.0% (71.8%–76.2%)	74.9% (72.7%–77.1%)	1.01 (0.93–1.10)
(B) Failed Maddocks	56	16	89	62.1% (59.7%–64.5%)	62.7% (60.3%–65.1%)	1.01 (0.92–1.10)
(C) Failed Immediate memory	146	40	86	59.1% (56.6%–61.6%)	63.2% (60.7%–65.6%)	1.07 (0.98–1.17)
(D) Failed digits backwards	238	34	83	63.7% (61.2%–66.1%)	68.7% (66.3%–71.0%)	1.08 (0.99–1.18)
(E) Tandem Gait	471	57	84	52.5% (50.0%–55.1%)	69.8% (67.5%–72.2%)	1.33 (1.21–1.46)
(F) Delayed recall	273	32	83	62.8% (60.3%–65.2%)	68.3% (66.0%–70.7%)	1.09 (1.00–1.19)
(G) Clinical signs	232	7	59	73.0% (70.8%–75.2%)	73.3% (71.1%–75.4%)	1.00 (0.91–1.11)

Table 3 Description of clinical judgement applied to HIA1 off-field screens with abnormal subtests

No of abnormal submodes, excluding symptoms	Cases	Clinical judgement applied (% of cases)	Clinical judgement correct (% of clinical judgements)	Accuracy		Relative change
				Theoretical	Real life	
(A) All subtests abnormal	2	0		58.9% (56.4%–61.4%)	58.9% (56.4%–61.4%)	1.00 (0.91–1.10)
(B) Five subtests abnormal	11	0		59.2% (56.7%–61.7%)	59.2% (56.7%–61.7%)	1.00 (0.91–1.10)
(C) Four subtests abnormal	49	6	67	61.8% (59.4%–64.3%)	61.9% (59.5%–64.4%)	1.00 (0.91–1.10)
(D) Three subtests abnormal	100	22	82	62.7% (60.2%–65.2%)	64.0% (61.6%–66.5%)	1.02 (0.93–1.12)
(E) Two subtests abnormal	234	39	80	59.8% (57.4%–62.3%)	65.2% (62.8%–67.6%)	1.09 (0.99–1.19)
(F) Single subtests abnormal	367	71	86	47.2% (44.7%–49.7%)	65.0% (62.8%–67.6%)	1.38 (1.25–1.52)

abnormal subtests relative to baseline, as shown in tables 2 and 3. Clinical judgement was applied most frequently to overrule abnormal tandem-gait assessments (57% of 471 cases), and least often to overrule the presence of symptoms (13% of 261 cases as assessed against baseline symptom endorsement).

Table 4 shows the frequency of symptom endorsement during HIA1 assessments, the positive predictive value (PPV) of that symptom, the number of occasions where each symptom was overruled and the resultant accuracy of screens as a result of such clinical judgement. Symptoms with a higher PPV for a final concussion diagnosis ('feeling in a fog', 'nausea or vomiting', 'slowing down' and 'blurred vision') were less likely to be over-ruled than lower risk symptoms less predictive for concussion (table 3, over-ruled in 9.3% vs 27.4% of instances, respectively). The application of clinical judgement to

symptoms did not improve screen accuracy for any of the nine symptoms.

Accuracy of application of clinical judgement to individual HIA1 assessment subtests

For all individual subtests, application of clinical judgement increased the point estimate for overall accuracy of the HIA1 assessment compared with strict application of baseline value thresholds, as shown in table 2. Improvements ranged from very small and non-significant (1.01-fold for symptoms and for Maddocks 90% CI (0.92 to 1.10) to larger significant increases (1.33-fold for tandem gait, 90% CI 1.21 to 1.46).

Figure 2 indicates that clinical judgement applied to individual subtests reduced FPR and TPR, though to different degrees, with the greatest improvement in accuracy for tandem gait, 52.5%–69.8%, (see also table 2),

Table 4 Frequency of symptom endorsement during the HIA1, including individual positive predictive values

|	Cases	Concussion confirmed (PPV)	Clinical judgement applied (%)	Clinical judgement correct (%)	Theoretical accuracy, %	Real-life accuracy, %	Ratio
Feel in a fog	75	70 (93.3)	3 (4.0)	1 (33.2)	65.7	65.7	1.00
Nausea or vomiting	29	26 (89.7)	2 (6.9)	1 (50)	62.0	62.0	1.00
Slowing down	95	82 (86.3)	11 (11.6)	4 (36.4)	66.1	65.8	1.00
Blurred vision	96	82 (85.4)	5 (5.2)	5 (100)	66.0	66.5	1.01
Light or noise sensitivity	29	24 (82.8)	3 (10.3)	1 (33.3)	61.6	61.5	1.00
Headache	138	113 (81.9)	13 (9.4%)	9 (69.2)	67.8	68.2	1.01
Pressure in head	115	94 (81.7)	13 (11.3)	7 (53.8)	66.5	66.5	1.00
Dizziness	126	102 (81.0)	11 (8.7)	7 (63.6)	66.9	67.2	1.00
Unwell	47	38 (80.9)	4 (8.5)	3 (75)	62.5	62.7	1.00
Any symptom	320	249 (78)	45 (14.1)	30 (66.7)	75.8	77.2	1.02
High-risk symptom present	194	164 (84.5)	18 (9.3)	10 (55.6)	71.9	72.1	1.00
Low-risk symptom present but excluding high risk	126	85 (67.5)	27 (27.4)	20 (74.1)	63.9	65.0	1.02

PPV, positive predictive value.
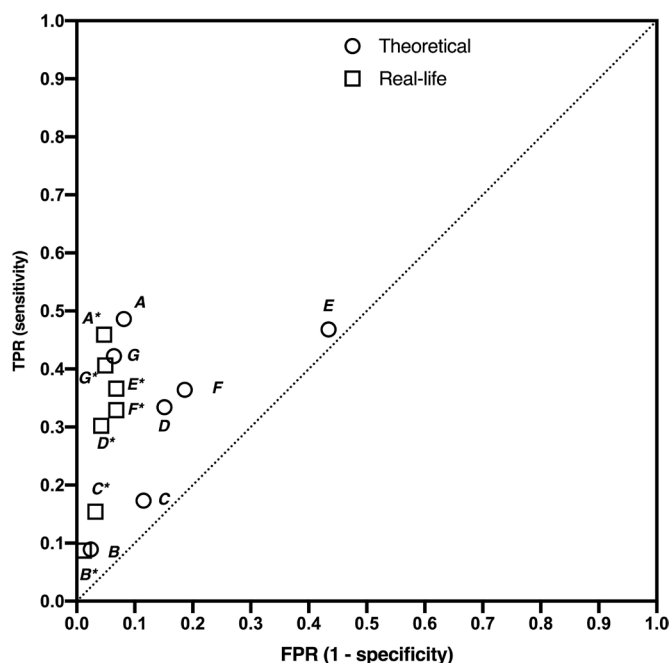
Figure 2 Sensitivity and specificity of individual HIA1 assessment subtests with and without clinical judgement applied. This analysis does not account for the presence of other subtests that may copresent with each subtest. The theoretical TPR and FPR are indicated by circles, while real-life TPR and FPR are shown as square symbols. Key: (A) symptoms (B) failed Maddocks (C) failed immediate memory (D) failed digits backwards (E) tandem gait (F) delayed recall (G) clinical signs. FPR, false positive rate; TPR, true positive rate.

Figure 3 Sensitivity and specificity of combinations of HIA1 assessment subtests with and without clinical judgement applied. The theoretical TPR and FPR are indicated by circles, while real-life TPR and FPR are shown as square symbols. key: (A) all modes abnormal, (B) five modes abnormal, (C) four modes abnormal, (D) three modes abnormal, (E) two modes abnormal, (F) single mode abnormal. FPR, false positive rate; TPR, true positive rate.

the result of a reduction in FPR from 43.4% to 6.8% (Figure 2, Symbol E and E*). In real terms, clinical judgement applied to abnormal tandem gait results reduced false positives by 226, with an increase of 44 false negatives. For Immediate Memory (C) and Digits Backwards (D), the improvement in diagnostic accuracy approached statistical significance (0.98–1.17 and 0.99–1.18, respectively, see table 2), while tandem gait (E) and delayed recall (F) improved significantly with clinical judgement.

Accuracy of application of clinical judgement to combinations of HIA1 assessment subtests

Figure 3 and table 3 show how clinical judgement affects accuracy when applied to a range of numbers of abnormal subtests. A single abnormal subtest was correctly overruled 86% of the time. This resulted in a 1.38-fold (90% CI 1.25 to 1.52) improvement in screen accuracy (47.2%–65.0%, table 3), by virtue of a large reduction in the FPR with only a slight decrease in TPR (see symbols F and F* in figure 3). Two abnormal subtests had a higher theoretical accuracy than one abnormal subtest (59.8% vs 47.2%, table 3) and clinical judgement was less likely to be correct in this situation (80% correct vs 86% for one abnormal subtest). The overall result was a 1.09-fold increase (90% CI 0.99 to 1.19, table 3) in HIA1 assessment accuracy, owing to improved FPR that was greater than the reduction in TPR when two abnormal
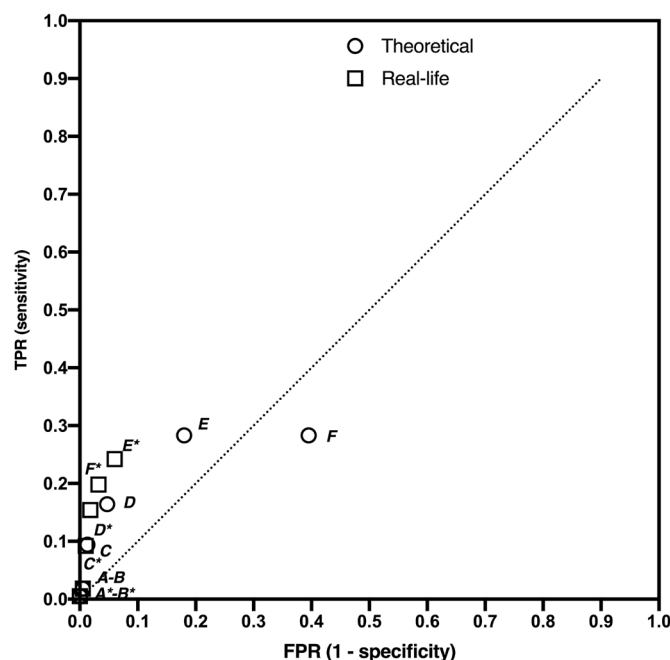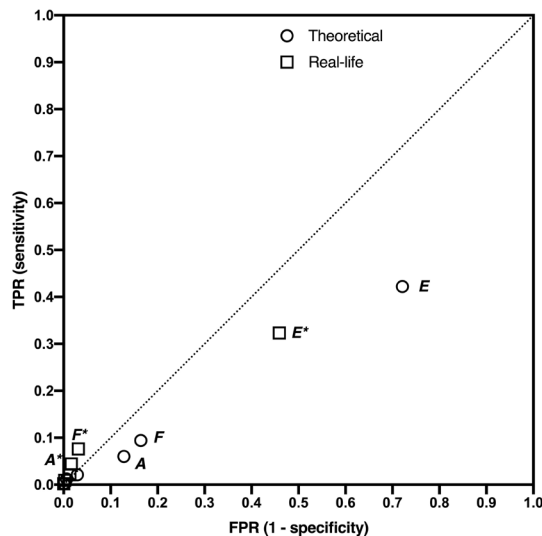
subtests were over-ruled by the clinician (symbol E and E* in figure 3).

For more than two abnormal subtests, there was no statistical benefit of clinical judgement on overall accuracy, as indicated by the relative change ranging between 1.00 and 1.02, and the large overlap in the 90% CI ranges for theoretical and real-life accuracy (table 3). At high numbers of abnormal subtests (4–6), FPR approached zero, indicating high specificity for these scenarios.

Exploration of performance clinical judgement with specific subtest combinations

An assessment of various combinations of the four cognitive subtests (Maddocks, Immediate Memory, Digits backwards and Delayed Recall) is shown in figure 4, with an associated table representing the data. For a single abnormal cognitive subtest, clinical judgement improved accuracy 1.11-fold (1.01–1.22), with FPR decreasing from 12.8% to 1.6% (symbol A and A*, figure 2). As the number of abnormal cognitive subtests increased, the effect of clinical judgement on accuracy diminished, though only nine instances where three or more cognitive subtests were abnormal were present in the cohort (figure 4).

When any cognitive subtest was abnormal, and no other subtest category (symptoms, balance or clinical signs) was abnormal, the accuracy of testing improved by 1.13-fold (1.03-fold to 1.25-fold) when clinical judgement was applied (symbol F and F*, figure 4).
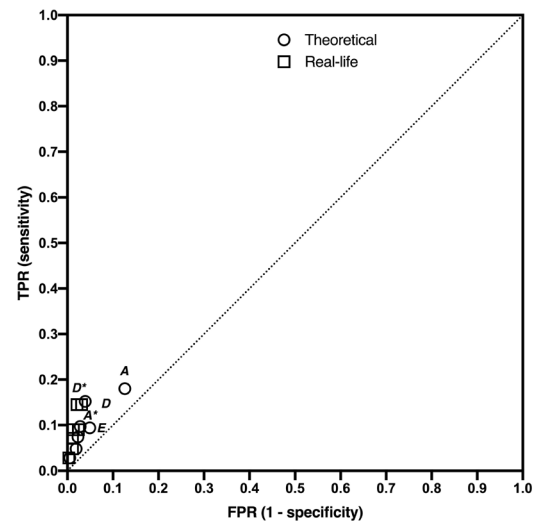
Figure 4 True positive rate (TPR) versus false positive rate (FPR) for cognitive subtests. The theoretical TPR and FPR are indicated by circles, while real-life TPR and FPR are shown as square symbols, with labelling identifying selected scenarios presented in the table below the figure, with* designating the real-life TPR and FPR combination.

Cognitive sub-test failures	Cases	Clinical judgment applied (% of cases)	Clinical judgment correct (% of clinical judgments)	Accuracy		Relative change
				Theoretical	Real-life	
A) One cognitive fail only	105	72%	91%	53.7% (51.1% - 56.2%)	59.6% (57.1% - 62.1%)	1.11 (1.01 - 1.22)
B) Two cognitive	27	37%	100%	57.8% (55.3% - 60.4%)	58.8% (56.3% - 61.3%)	1.02 (0.93 - 1.12)
C) Three cognitive	9	44%	75%	58.8% (56.3% - 61.3%)	59.0% (56.5% - 61.5%)	1.00 (0.91 - 1.10)
D) Four cognitive	1	0%	NA	58.8% (56.3% - 61.3%)	58.8% (56.3% - 61.3%)	1.00 (0.91 - 1.10)
E) Zero failed cognitive tests	628	31%	84%	33.8% (31.4% - 36.2%)	45.1% (42.6% - 47.6%)	1.34 (1.19 - 1.50)
F) Any cognitive fail and no other sub-test fails	142	63%	91%	53.0% (50.5% - 55.5%)	60.0% (57.6% - 62.5%)	1.13 (1.03 - 1.25)

Figure 5 True positive rate (TPR) versus false positive rate (FPR) for selected subtest combinations. The theoretical TPR and FPR are indicated by circles, while real-life TPR and FPR are shown as square symbols, with labelling identifying selected scenarios presented in the table below the figure, with* designating the real-life TPR and FPR combination. DB, digits backwards; DR, delayed recall; IM, immediate memory; TG, tandem gait.

Sub-test combinations	Cases	Clinical judgment applied (% of cases)	Clinical judgment correct (% of clinical judgments)	Accuracy		Relative change
				Theoretical	Real-life	
A) TG + only one cognitive mode abnormal	156	47%	80%	58.7% (56.2% - 61.2%)	62.9% (60.4% - 65.3%)	1.07 (0.98 - 1.17)
B) DB + Maddocks	15	7%	100%	59.6% (57.1% - 62.1%)	59.7% (57.2% - 62.1%)	1.00 (0.91 - 1.10)
C) DB + IM	59	20%	75%	61.1% (58.6% - 63.6%)	61.7% (59.2% - 64.1%)	1.01 (0.92 - 1.11)
D) DB + DR	90	16%	79%	62.7% (60.2% - 65.2%)	63.5% (61.0% - 65.9%)	1.01 (0.92 - 1.11)
E) Memory (IM + DR)	71	24%	88%	59.8% (57.3% - 62.2%)	61.0% (58.5% - 63.5%)	1.02 (0.93 - 1.12)
F) TG, DB + DR	46	17%	88%	60.4% (57.9% - 62.9%)	61.0% (58.5% - 63.5%)	1.02 (0.93 - 1.12)
G) TG + Memory	33	15%	100%	59.6% (57.1% - 62.1%)	60.0% (57.6% - 62.5%)	1.01 (0.92 - 1.11)

The combination or copresentation of different abnormal subtests is assessed and shown in figure 5. The diagnostic accuracy of multiple cognitive test abnormalities was not improved substantially when clinical judgement was applied. An abnormal tandem gait assessment plus one abnormal cognitive subtest resulted in a 1.07-fold (0.98–1.17) increase, from 58.7% to 62.9%, symbol A and A*, figure 5) though this did not reach statistical significance. For all combinations of digits backwards, immediate memory and delayed recall, clinical judgement achieved an increase in accuracy ranging between 1.01-fold and 1.02-fold.

Application of 'guided clinical judgement'

We next explored how a limited or guided version of clinical judgement might influence the creation of false negative cases arising from clinical judgement. To do this, we investigated specific scenarios described above where the effect of clinical judgement on screen accuracy was clearly insignificant. We found that this would prevent clinical judgement in the presence of any endorsed symptoms (table 2), clinical suspicion (table 2), abnormal Maddocks questions (table 2) or more than one failed cognitive subtest (figure 4). We found that this eliminated 21 false negative cases. Specifically, 1 instance of 3 cognitive subtest fails, 7 instances of clinical signs present, 12 cases with a symptom present and 1 case with a failed

Maddock's test present had been incorrectly overruled by the physician, creating false negative cases.

DISCUSSION

Summary of results

This study is the first to examine in detail how clinical judgement is applied during off-field concussion screening assessments, and the impact this has on the accuracy of these screens. We confirm that clinical judgement is commonly applied, significantly improving the overall diagnostic accuracy of the HIA off-field assessment, specifically by reducing false positive cases, but with the creation of a smaller number of false negative cases (when abnormal subtests are incorrectly over-ruled). Since such false negative cases are undesirable. Clinical judgement should be guided or cautioned against in certain situations, such that it is unrestricted only when the net effect is a large increase in overall accuracy . Our analysis indicates that such guided clinical judgement would reduce the number of false negative events (33% reduction), our primary target.

Certain subtests were less likely to be correctly over-ruled, and we thus recommend that doctors exercise significant caution when considering employing clinical judgement to over-rule endorsed symptoms, clinical

signs, failed Maddock's questions or more than one cognitive subtest fail.

Interpretation

Fuller *et al* previously described improved overall accuracy of the off-field concussion screening assessment when clinical judgement was applied.[7] This improvement is created by a large reduction in the number of false positives, though at a cost of a small increase in 'missed' concussions. This highlighted the important trade-off between sensitivity and specificity inherent in diagnostic tests. There is a tension between the medical priority of ensuring all concussions are detected, and a test which lacks specificity, unnecessarily removing many non-concussed players from play, which might create barriers to acceptance and adherence with the policies in the future.

The present study examined these clinical judgement decisions in greater detail, to advise clinicians when to exercise caution in applying clinical judgement. At all times when there is any doubt at the very least a player should be removed for assessment; if following assessment doubt remains, the player should not return to play. However, this cautionary advice must also not excessively constrain clinicians from making potentially beneficial judgement decisions and requires balance and clinical discretion to optimise overall screen accuracy.

We have found that improvements in overall accuracy are substantially greater for certain subtests compared with others when clinical judgement is applied (figures 1–4). Because clinical judgement may reduce both FPR and TPR, it should only be applied where overall accuracy is substantially improved. We identify this limit as any subtest scenario where clinical judgement does not improve overall accuracy by a factor greater than 1.0 when applying a 90% confidence limit to the ratio of real life to theoretical accuracy.

We explored the impact of 95% confidence levels but found that this limit would preclude clinical judgement in most instances, resulting in the persistence of large numbers of false positive cases. We; thus, believe that 90% confidence levels are more appropriate, when factoring in the quantitative chances of benefit, triviality and harm.

The relative value placed on false positives and negatives will ultimately determine where the optimum trade-off lies. This may be determined explicitly through preference-based research methods such as discrete choice experiments. More commonly values are integrated implicitly by team doctors through application of clinical judgement or can be imposed externally through recommendations or clinical guidelines by regulatory bodies.

Applying this pragmatic and conservative principle, we recommend that team doctors exercise significant caution before applying clinical judgement and over-ruling the following:

- ▶ Any symptom endorsed by the player as new or changed from baseline intensity and related to head trauma (figure 2 and table 2).
- ▶ Any clinical sign noted by the doctor, including altered emotional status (nervous, anxious, sad, irritable), drowsiness, difficulty concentrating or remembering, clinical suspicion of concussion despite normal testing.
- ▶ Any failed Maddock's questions (figure 2).
- ▶ More than one failed cognitive subtest (Immediate memory, delayed recall, digits backwards) (figure 3).

Our analysis revealed that clinical judgement was incorrect on 63 of the 389 occasions on which it was applied. Guided clinical judgement preventing over-ruling in the above abnormal subtest scenarios, would prevent 21 of the 63 false negative cases that arose from incorrect clinical judgement. This represents one-third of all false-negatives created by clinical judgement.

Symptoms

A player reporting any symptom during the off-field assessment had a 78% chance of having a concussion. Doctors rarely over-rule the presence of symptoms, either with respects to their presence (table 2) or when assessed against baseline performance (figure 2). We find that when they did, clinical judgement added little to accuracy (1.02-fold increase). Symptoms have previously been found to be the most predictive and sensitive indicator of concussion in the off-field screen.[7 10] Despite the non-specific nature of symptoms, which can lead to false positive cases,[9 10] we conclude that clinical judgement, although infrequently applied to symptom endorsement, does not substantially reduce false positive cases, and therefore, we do not recommended that team doctors overrule the presence of symptoms.

Specific subtest scenarios

We found the diagnostic accuracy of any single subtest abnormality to be low (figure 1, symbol F and F*). Clinical judgement used to overrule a single abnormal subtest improved test accuracy from 47.2% to 65.0%, the result of a large reduction in FPR with only a small drop in TPR (figure 1). The beneficial effect of clinical judgement was smaller for two abnormal subtests (figure 1, symbol E and E*), but remained significant, whereas three or more abnormal subtests were not significantly improved by clinical judgement. Our initial high-level analysis thus cautions against clinical judgement when more than two abnormal subtests are present.

We identified certain single subtests (figure 2) and paired cognitive subtests (figures 3 and 4) where abnormal results should not be over-ruled. This ultimately means that the only combination of two subtest abnormalities subject to clinical judgement is one abnormal cognitive test and an abnormal Tandem Gait assessment (figure 4, symbols A and A*).

Baseline test utility is potentially affected by some of the following issues: learning effects(contributing to

concussed players passing the off-field assessment), poor collection process, poor player effort and the relatively poor diagnostic accuracy of individual modes tested during the screen, all of which may increase the likelihood that strict compliance with baselines may decrease diagnostic accuracy.[9–11]

Comparison to literature

To date, there have been no studies examining the influence of clinical judgement on side-line concussion screening. However, there is consistent evidence from other disciplines (particularly emergency medicine) that clinical acumen is frequently applied when interpreting clinical decision rules or tests,[12–15] where clinical judgement may positively affect the utility of clinical decision rules. This tends to result in superior overall performance, with improved specificity offset by a small reduction in sensitivity, as we show here. For example, paramedics have been shown to 'over-rule' prehospital major trauma triage tools to make their own transport decisions on which patients require major trauma centre care.[13]

Limitations

Limitations of the present study include potential external and internal validity issues. The cohort is made up of elite male rugby union players—and similarly experienced and well-trained doctors—and as such may not be generalisable to other populations and sports. The subtests used in the present cohort include Immediate Memory and Delayed Recall using the 5-word lists, now largely replaced by 10-word lists, which is expected to enhance their sensitivity significantly. This may have implications for clinical judgement accuracy, since we presently suggest that these subtests are not statistically improved by clinical judgement. The introduction of the 10-word lists is expected to make the test more difficult to pass, possibly creating more false positive cases. Were this to occur, the frequency with which clinical judgement is applied may increase, and this clinical judgement stands to improve test accuracy more than our current findings suggest.[16]

We acknowledge a risk of bias in the final diagnosis of concussion, since most clinical screens (HIA2 and HIA3) are performed by the same doctor who performed the off-field screen. This may affect subsequent clinical screens when concussion is diagnosed, with concussion more likely to be rejected in cases where a player had been returned to play despite an abnormal off-field screen.

Another limitation of the present study is that the impact of clinical judgement is best evaluated when many cases are assessed. For some submode scenarios, we do not have enough cases with which to properly evaluate the implications. Similarly, we have too few cases to evaluate how unique subtest abnormalities are affected by clinical judgement. Our approach to this limitation is to caution against clinical judgement for these situations.

In some situations, clinicians over-ruled more than one subtest was abnormal compared with baseline. Because this is a retrospective study, we cannot identify which individual subtest was over-ruled.

This analysis (figures 2–4) assesses the outcomes when each subtest is abnormal and does not consider whether the subtest is abnormal in conjunction with other subtests. This accounts for the relatively high TPR and low FPR of each abnormal subtest, compared with what would be observed if each subtest was assessed as the only abnormal subtest in an HIA screen. An analysis of single subtest abnormalities (data not shown) revealed a similar pattern, but a significant reduction in cases available limits the interpretation of this unique abnormal subtest evaluation.

CONCLUSIONS

This study is the first to describe in detail how clinical judgement is used to interpret off-field concussion screening assessments.

Team doctors commonly use clinical judgement in rugby, this practice is supported by consensus opinion, but improved accuracy (primarily by reducing false positive cases) may cause a small increase in false negative cases.

Rugby will be implementing this clinical judgement guidance to manage this trade-off, and to ensure that false negative cases do not increase to the detriment of player safety. To achieve this, we advise against clinical judgement for the following combinations:
- Any symptom.
- Any clinical sign noted by the doctor.
- Any failed Maddock's questions (figure 2).
- More than one failed cognitive subtest (immediate memory, delayed recall, digits backwards).

had provided written informed consent for all data gathered as paart of the World Rugby Concussion management programme, to be used for research in a deidentified manner.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request. Original participant data belong to the players and the clubs/unions that generate such data. This may be provided on request to third parties. World rugby (the corresponding author) may facilitate the provision of that data, in terms of permissions and contacts, though there is not a single point of contact, since the data are generated globally from multiple teams and unions.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

ORCID iDs
Éanna Falvey http://orcid.org/0000-0003-3961-1805
Ross Tucker http://orcid.org/0000-0002-4294-4822

REFERENCES

1 McCrory P, Meeuwisse W, Dvořák J, *et al*. Consensus statement on concussion in sport-the 5th international conference on concussion in sport held in Berlin, October 2016. *Br J Sports Med* 2017;51:838–47.
2 Patricios JS, Ardern CL, Hislop MD, *et al*. Implementation of the 2017 Berlin concussion in sport group consensus statement in contact and collision sports: a joint position statement from 11 national and international sports organisations. *Br J Sports Med* 2018;52:635–41.
3 Raftery M, Kemp S, Patricios J, *et al*. It is time to give concussion an operational definition: a 3-step process to diagnose (or rule out) concussion within 48 h of injury: World Rugby guideline. *Br J Sports Med* 2016;50:642–3.
4 McCrory P, Meeuwisse W, Gioia GA. Sport concussion assessment tool - 5th edition. *Br J Sports Med* 2017;51:851–8.
5 Rugby W. World Rugby Concussion Resources: Recognise & Remove [Internet]. Available: https://playerwelfare.worldrugby.org/concussion
6 Hodgson L, Patricios J. Clarifying concussion in youth rugby: recognise and remove. *Br J Sports Med* 2015;49:966–7.
7 Fuller GW, Tucker R, Starling L, *et al*. The performance of the world rugby head injury assessment screening tool: a diagnostic accuracy study. *Sports Med Open* 2020;6:2.
8 World Rugby. World Rugby HIA Protocol [Internet]. Available: https://playerwelfare.worldrugby.org/content/getfile.php?h=b47de0127c9f35e604d51e2eac1815cc&p=downloads/concussion/HIA_Protocol_EN.pdf
9 Garcia G-GP, Broglio SP, Lavieri MS, *et al*. Quantifying the value of multidimensional assessment models for acute concussion: an analysis of data from the NCAA-DoD care consortium. *Sports Med* 2018;48:1739–49.
10 Harmon KG, Clugston JR, Dec K, *et al*. American medical Society for sports medicine position statement on concussion in sport. *Br J Sports Med* 2019;53:213–25.
11 Chin EY, Nelson LD, Barr WB, *et al*. Reliability and validity of the sport concussion assessment Tool–3 (SCAT3) in high school and collegiate athletes. *Am J Sports Med* 2016;44:2276–85.
12 Schriger DL, Elder JW, Cooper RJ. Structured clinical decision AIDS are seldom compared with subjective physician judgment, and are seldom superior. *Ann Emerg Med* 2017;70:338–44.
13 van Rein EAJ, van der Sluijs R, Raaijmaakers AMR, *et al*. Compliance to prehospital trauma triage protocols worldwide: a systematic review. *Injury* 2018;49:1373–80.
14 Sanders S, Doust J, Glasziou P. A systematic review of studies comparing diagnostic clinical prediction rules with clinical judgment. *PLoS One* 2015;10:e0128233.
15 Brehaut JC, Stiell IG, Visentin L, *et al*. Clinical decision rules "in the real world": how a widely disseminated rule is used in everyday practice. *Acad Emerg Med* 2005;12:948–56.
16 Echemendia RJ, Meeuwisse W, McCrory P. The sport concussion assessment tool 5th edition (SCAT5): background and rationale. *Br J Sports Med* 2017;51:848–50.