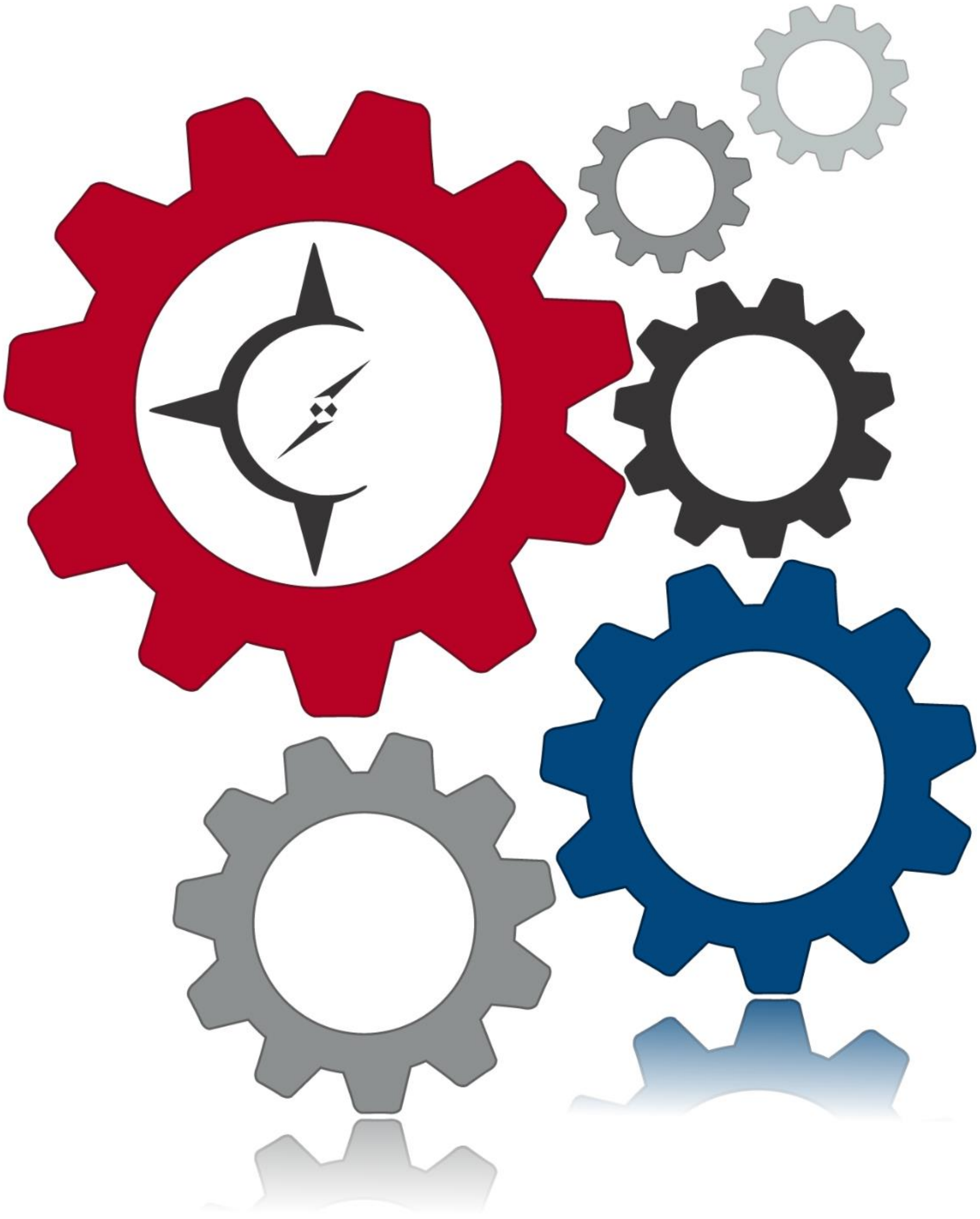
Linkage Review and Evaluation of Rugby Players and Referees added to the Integrated Data Infrastructure



Stephanie D’Souza
The University of Auckland

Chao Li
The University of Auckland

Kenneth L Quarrie
New Zealand Rugby

Barry J Milne
The University of Auckland

COMPASS
RESEARCH CENTRE

FACULTY OF ARTS
UNIVERSITY OF AUCKLAND
Waipapa Taumata Rau

Disclaimer

These results are not official statistics. They have been created for research purposes from the Integrated Data Infrastructure (IDI), which is carefully managed by Stats NZ.

For more information about the IDI, please visit https://www.stats.govt.nz/integrated-data.

Technical information

Counts described in this report have been random rounded to a base of 3, as per the confidentiality rules of Stats NZ.

Acknowledgements

This work was funded by World Rugby Limited and the New Zealand Rugby Foundation as part of Kumanu Tāngata: The Aftermatch Project. We would like to acknowledge Stats NZ and their staff for loading the New Zealand Rugby Register data to the IDI, for access to the IDI, and their timely review of output data. We would also like to thank the Public Policy Institute at the University of Auckland for access to their Statistics New Zealand data lab.

The Rugby Dataset and the Integrated Data Infrastructure

Original sample description

The Rugby Dataset contains anonymised data relating to high-level rugby players in New Zealand, that are contained within the New Zealand Rugby Register (Akers, 2016). The New Zealand Rugby Register, compiled by Clive Akers, contains information on players and referees involved in New Zealand rugby at a provincial union level or higher ('first-class rugby') between the years of 1870 to 2015. While dates of birth for players are included for a large majority of players in the rugby register, they are included for only a few of the referees.

Integrated Data Infrastructure

To allow for assessment of associations between rugby playing and a broad range of characteristics and areas of functioning, the Rugby Dataset was added to the Integrated Data Infrastructure (IDI). The IDI is a collection of New Zealand administrative data sources which have been linked at the individual level, de-identified, and made available for 'public-good' research under strict access controls maintained by the curator of the IDI, Stats NZ (Milne et al., 2019; Stats NZ 2021). Data added to the IDI are linked to a central concordance table (the 'spine'), which forms the base population of the IDI and can be thought of as an 'ever resident' New Zealand population (this is currently around 10 million individuals).

Table 1 indicates the breakdown in rugby players, referees and other individuals in the sample (N = 17,433) added to the IDI.

Table 1. Count of rugby players and referees in the original rugby cohort

Rugby players	Referees	Other[1]
16,101	972	360

Note: [1] Other includes other individuals involved in the sport, such as coaches.

Review of Fast Match Load

Fast Match Load

The Rugby Dataset was added to the IDI using what is known as a Fast Match Load (FML), to allow for a quicker addition to the database outside of one of the regular IDI refreshes (quarterly updates of IDI data). A list of variables in the loaded Rugby Dataset is provided in Table A1 in the Appendix. According to Stats NZ (n.d.), the FML replicates the standard IDI linking methodology (probabilistic linking) but without a review of the links made. Key information used for FML linking incudes names, date of birth and sex. More detail on this can be found in Annex A of the Fast Match Loader – User Manual (Stats NZ, n.d.). As an official review of linking quality is not conducted with FML data, the researcher is required to review the linking process.

The FML produces an snz_spine_uid for those in the Rugby Dataset that are able to be linked to the IDI spine. For those who were able to be linked to the spine, other relevant information in our linkage review includes the weight and the near exact indicator. The weight is a numeric value assigned to a record in the Rugby Dataset based on its similarity to a record in the spine, using the linking variables (i.e., names, date of birth, and sex). The near exact indicator indicates if the link is near exact or non-exact. Near exact links are links where all the linking variables are in agreement (with some small tolerance for errors), whereas non-exact links are all other links (e.g., where some of the linking variables are in agreement but others are not).

As FML linkage reviews are not conducted, two scenarios can occur that we may need to make decisions on:
Scenario 1: The same spine id is linked to more than one identity in the FML Rugby Dataset;
Scenario 2: Two spine ids are linked to the same identity in the FML Rugby Dataset.

Review process and results

Our first step was to determine how many records in the Rugby Dataset were able to be linked to the spine id. Our results showed that 16,830 (96.5%) of the full dataset had corresponding spine ids. Specifically, 15,555 (96.6%) of rugby players and 939 (96.3%) of referees were linked to the spine.

We then examined the weights of each linked record, stratified by near exact indicator status. Our results showed that while there was a lower mean weight in non-exact links, some of these links still had relatively high weights whereas some of the near-exact links had relatively low weights (see mean, minimum and maximum in Table 2). Given the overlapping range in weight values for near and non-exact links, we decided that focusing on near exact indicator status may be more informative in our decision-making than the weights.

Table 2. Weight descriptives by near exact indicator status

Indicator	N	Mean	SD	Min.	Max.
Near exact	15,231	28.33	6.92	5.21	49.20
Non-exact	1,599	17.36	7.39	0.01	42.58

Next, in those who were linked to the spine, we examined the occurrence of Scenarios 1 and 2 above. We found that no two spine ids were linked to the same record in the Rugby Dataset (Scenario 2). However, we did find that there were 192 spine ids linked to more than one identity in the FML Rugby Dataset. Of these 192 spine ids, 36 were linked to two identities in the Rugby Dataset that were discrepant on their near exact indicator. Specifically, one rugby identity showed a near-exact link whereas the other showed a non-exact link.

Based on the above information, we grouped the rugby records into four separate categories for determining which records to use in our analyses within the IDI. This was based firstly on whether or not the links were unique or a duplicate; a unique link is where one spine id was uniquely linked to an identity in the Rugby Dataset, whereas a duplicate link is on where one spine id was linked to more than one identity in the Rugby Dataset (i.e., Scenario 2 above). We also considered the near exact indicator status in our groupings. These groupings are defined as follows:

1. Definitely use – these are unique links and near-exact matches.
2. Probably use – these are unique links but non-exact matches.
3. Maybe use – this is part of a duplicate link, but this link is a near-exact match (and the other link is a non-exact match and therefore excluded).
4. Excluded – These are records that have either not been linked to the spine or duplicate links that were not grouped as 'maybe'. These duplicate links could have had both links flagged as near exact or non-exact matches, or is the non-exact link in a pair where the other duplicate is a near-exact link.

The numbers for these groupings and their weight descriptives are shown in Table 3.

Table 3. Weight descriptives by usability groupings

Usability grouping	N (%)[1]	Mean	SD	Min.	Max.
Definitely use	14,910 (85.5%)	28.44	6.88	5.21	49.20
Probably use	1,533 (8.8%)	17.33	7.37	0.01	42.58
Maybe use	36 (0.2%)	29.06	7.11	12.40	42.56
Excluded[2]	954 (5.5%)	21.23	6.98	1.43	40.63

Note: [1] Percentages indicate the proportion of the original rugby cohort (N = 17,433)
[2] Weight descriptives for the Excluded group have only been provided for the records that were able to be linked to the spine (n = 348).

The breakdown in rugby players and referees or coaches for those in the 'definitely use' group is presented in Table 4.

Table 4. Count of rugby players and referees in the 'definitely use' group

Rugby players	Referees	Other[1]
13,800 (85.7%)	819 (84.3%)	291 (80.8%)

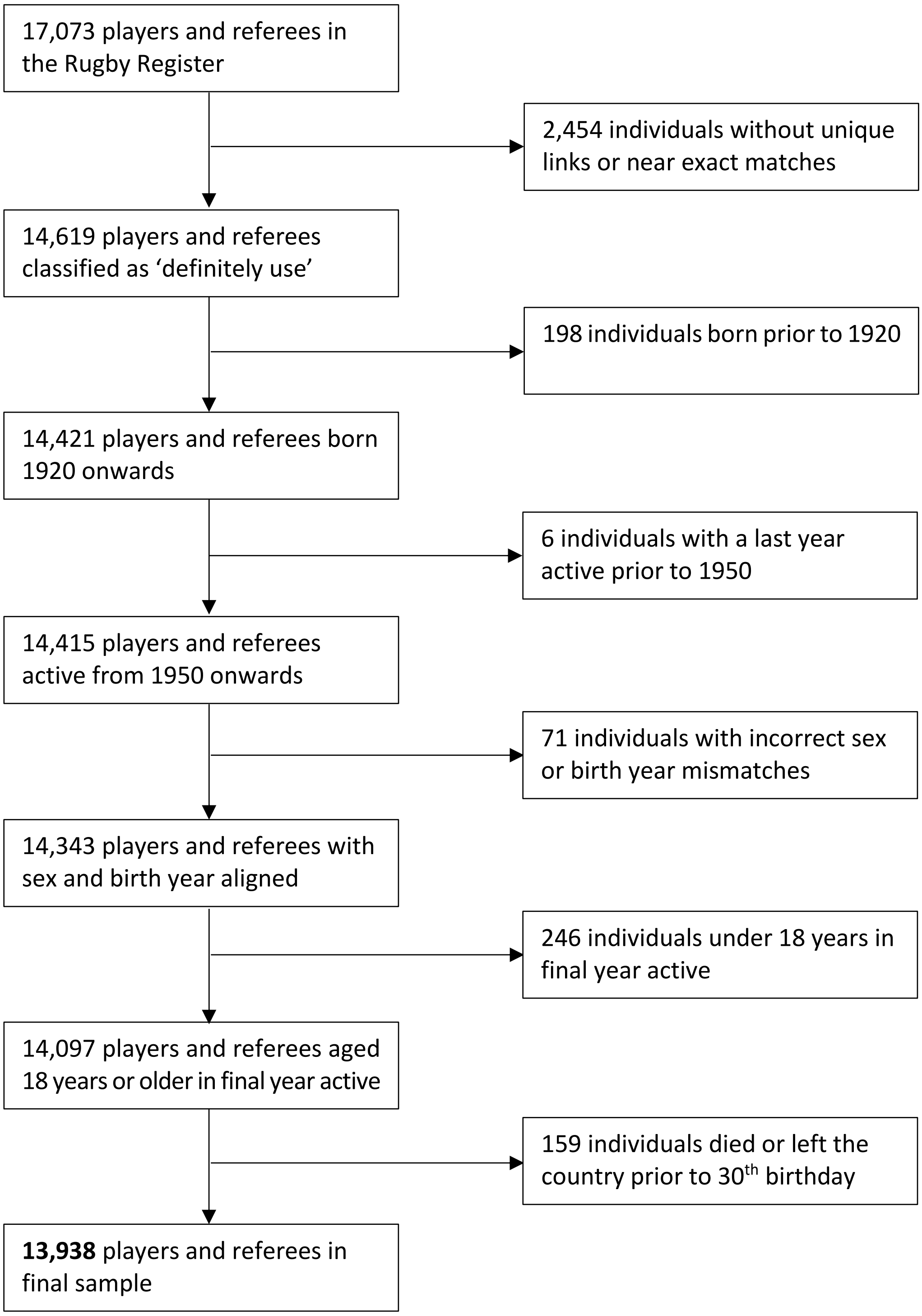
Additional inclusion criteria and data checking

We applied additional inclusion criteria and data checking to create our final cohort, detailed below.

1. Individuals must have been born from 1920 onwards. Reliable and complete date of birth information is available within the IDI from approximately 1920. As date of birth is used for linking, anyone born before 1920 may not be reliably linked to the IDI. Therefore, any individual with a birth year prior to 1920 in the Rugby Dataset was excluded.
2. Individuals must have been active (participation in first-class rugby) from 1950 onwards. This exclusion criteria is consistent with the funding proposal presented to World Rugby for work with this data. This allows for most players to be tracked from an age prior to when most chronic illnesses begin. Last year active information was obtained from the Rugby Dataset.
3. Sex and birth year (up to one year difference) must align across the IDI personal details and Rugby Datasets. Data linkage in the IDI is undertaken using the Fellegi-Sunter method of probabilistic linkage (Fellegi & Sunter, 1969), which can result in linkage errors (though false positive rates are generally relatively low; Milne et al., 2019). As these data were added to the IDI using a FML, we decided to evaluate potential linkage errors by comparing birth year and sex in the linked IDI records from the Personal Details table to birth year and known sex (all players and referees are male) in the Rugby Dataset. We allowed for a mismatch in birth year of up to one year.
4. Individuals must have been at least 18 in their final year active. Age in the final year active was calculated using birth year from the Personal Details table and last year active from the Rugby Dataset. Those who were under 18 in their final year active were excluded. This is consistent with our funding proposal and ethics approval for this project.
5. Individuals must have been alive and in the country at their 30th birthday. The average age of players in their final year active was 26 years (M = 26.19, SD = 4.58) amongst those in the 'definitely use' group. As such, the decision was made to track players from their 30th birthday, after most cohort members had exposure to the sport, to evaluate time to outcomes of interest. Therefore, any individuals who died or moved overseas prior to their 30th birthday, identified through the personal details and border movements tables, respectively, were excluded.

Figure 1 demonstrates the inclusion/exclusion criteria applied and the numbers excluded at each step. To avoid suppression due to small counts for referees, the flow chart provides counts aggregated across players and referees. The final cohort consisted of 13,938 individuals: 13,227 rugby players (82.2% of players in the original cohort) and 711 referees (73.1% of referees in the original cohort).

Figure 1. Cohort flow chart


17,073 players and referees in the Rugby Register
2,454 individuals without unique links or near exact matches
14,619 players and referees classified as ‘definitely use’
198 individuals born prior to 1920
14,421 players and referees born 1920 onwards
6 individuals with a last year active prior to 1950
14,415 players and referees active from 1950 onwards
71 individuals with incorrect sex or birth year mismatches
14,343 players and referees with sex and birth year aligned
246 individuals under 18 years in final year active
14,097 players and referees aged 18 years or older in final year active
159 individuals died or left the country prior to 30th birthday
13,938 players and referees in final sample

Demographic descriptives

A breakdown of demographics for rugby players and referees, shown in Table 5, indicated that referees were more likely to be European than players, with minimal non-European representation. Referees also had a greater mortality prevalence, though this is likely because these individuals appeared to be older relative to rugby players, as indicated by the birth year breakdown.

As noted previously, most players had date of birth information already available in the New Zealand Rugby Registry, but this was only available for a small proportion of the referees. Therefore, records from the New Zealand Births, Deaths and Marriages (BDM) register were used to obtain dates of birth for referees within the Rugby Register to help increase linkage rates. Data are made publicly available for anyone who was born at least 100 years ago, or who died at least 50 years ago and were born at least 80 years ago. Because dates of birth for referees were primarily sourced via the BDM register, it is unsurprising that referees who are older and/or have died are over-represented within the data. If the project is working with outcomes that are pertinent to older age (e.g., mortality, neurodegenerative diseases), these individuals will likely bias analyses and it may be best to exclude them.

Table 5. Demographic information for rugby players and referees in the final cohort

	Rugby players n (%)	Referees n (%)
Total response ethnicity		
European	10,425 (78.8%)	627 (88.2%)
Māori	2,424 (18.3%)	45 (6.3%)
Pasifika	486 (3.7%)	S
Asian	39 (0.3%)	S
MELAA	105 (0.8%)	S
Other	267 (2.0%)	21 (3.0%)
NZ vs. overseas born		
NZ born	12,564 (95.0%)	681 (95.8%)
Overseas born	663 (5.0%)	30 (4.2%)
Mortality age		
30–39 years	102 (0.8%)	S
40–49 years	222 (1.7%)	12 (1.7%)
50–59 years	414 (3.1%)	21 (3.0%)
60–69 years	741 (5.6%)	60 (8.4%)
70–79 years	1,077 (8.1%)	96 (13.5%)
80–89 years	921 (7.0%)	99 (13.9%)
90–99 years	129 (1.0%)	15 (2.1%)
Overall mortality prevalence	3,606 (27.3%)	303 (42.6%)

	Rugby players n (%)	Referees n (%)
Birth year		
1920–1929	1,473 (11.1%)	195 (27.4%)
1930–1939	2,562 (19.4%)	162 (22.8%)
1949–1949	2,550 (19.3%)	153 (21.5%)
1950–1959	2,301 (17.4%)	135 (19.0%)
1960–1969	2,361 (17.8%)	51 (7.2%)
1970–1984[1]	1,977 (14.9%)	12 (1.7%)

Note: Percentages given indicate the proportion as part of the final cohort (13,227 rugby players and 711 referees.

S = suppressed due to low counts (≤6).

[1] Due to low counts, the final birth year band aggregates those born 1970–1984 (1984 is the final year during which any cohort member was born).

Conclusion

Based on the review process, primary analyses should work with the 'definitely use' grouping and apply the inclusion/exclusion criteria specified under 'Additional inclusion criteria and data checking'. It may be best to work just with the rugby player data, as the referee group appears to be biased towards those who are of older age and have passed. Further, as most outcome data are available from 1988 at the earliest, additional exclusion criteria could also be applied to the Rugby Dataset by excluding those who died prior to 1988.

Sensitivity analyses could extend the cohort to include the 'probably use' grouping and/or the 'maybe use' grouping. This decision will depend on whether the researchers want to exclude non-exact matches and any links that were part of a duplicate link.

References

Akers, C. (2016). *New Zealand Rugby Register 1870–2015: The Players, Referees and Administrators in First-class Rugby*. New Zealand Rugby Museum.

Fellegi, I. P., & Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, 64(328), 1183–1210, https://doi.org/10.1080/01621459.1969.10501049.

Milne, B. J., Atkinson, J., Blakely, T., Day, H., Douwes, J., Gibb, S., Nicholson, M., Shackleton, N., Sporle, A., & Teng, A. (2019). Data Resource Profile: The New Zealand Integrated Data Infrastructure (IDI). *International Journal of Epidemiology*, https://doi.org/10.1093/ije/dyz014.

Stats NZ (n.d.). *Fast match load—External user manual*. Stats NZ.

Stats NZ (2021). Integrated Data Infrastructure; [accessed 2021 October 29]. https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure.

Appendix

Table A1. Variable details for Rugby Dataset loaded into the Integrated Data Infrastructure

Variable	Description
Rugby_Administrator	Indicator for provincial union/national union officers
age_first_played_nbr	Age in first year active for players
age_last_played_nbr	Age in last year active for players
birth_month_nbr	Birth month
birth_year_nbr	Birth year
career_games_count	Total number of career games
club_exposure_est_nbr	Exposure estimate for club play based on matches and estimated training
club_exposure_est_pos_nbr	Estimated number of club training sessions based on number of club matches played
coach_first_year	First year active for coaches
coach_ind	Indicator for coaches
coach_last_year	Last year active for coaches
coach_total_year	Number of years coaching first class rugby teams
first_played_year	First year active for players
last_played_year	Last year active for players
matches_per_year_nbr	Average number of first-class matches per year (career_games_count divided by years played count)
personality_ind	Indicator for rugby personalities in the public domain (e.g., rugby authors, historians, journalists, television presenters, etc.)
played_provincial_code	Yes/no for representing a provincial rugby team
position1_code	Primary position played for players
position2_code	Secondary position played for players who played two or more positions
position3_code	Third position played for players who played three positions
position_factor_nbr	Estimated intensity of contacts by position
pro_code	Indicator for All Blacks and/or Super Rugby players
ref_first_year	First year active for referees
ref_last_year	Last year active for referees
ref_matches_nbr	Count of first-class matches refereed
referee_ind	Indicator code for referees
rep_exposure_est_nbr	Exposure estimate for representative play based on matches and estimated training

Variable	Description
rep_exposure_est_pos_nbr	Representative exposure estimate from above multiplied by position factor
rep_training_est_nbr	Estimated number of representative training sessions based on number of representative matches played
selector_ind	Team selector indicator
snz_fml_3_uid	Unique identifier specific to the fast match loaded rugby data
snz_sex_code	Sex code (male or female)
total_exposure_est_nbr	Sum of club and representative exposure (includes position factor)
years_played_count	Total number of years active for players