

# International Journal of Sports Science & Coaching

## A systematic review of sport related packages within the R CRAN repository

|                               |  |
|-------------------------------|--|
| Journal:                      | <i>International Journal of Sports Science &amp; Coaching</i>  |
| Manuscript ID                 | Draft  |
| Manuscript Type:              | Review   |
| Date Submitted by the Author: | n/a  |
| Complete List of Authors:     | Casals, Martí ; Research Centre Network for Epidemiology and Public Health, Sport and Physical Activity Studies Centre (CEEAF), Faculty of Medicine. ; University of Vic - Central University of Catalonia<br>Fernandez, Jose; Australian Catholic University<br>Martínez, Víctor; Universitat Politècnica de Catalunya<br>Lopez, Michael; NFL<br>Langohr, Klaus; Universitat Politècnica de Catalunya Departament d'Estadística i Investigació Operativa<br>Cortes, Jordi; Universitat Politècnica de Catalunya Departament d'Estadística i Investigació Operativa  |
| Keywords:                     | statistical software, R, sports statistics, sports analytics, sports science, statistical education  |
| Abstract:                     | <p>The sports statistics community is rapidly evolving. R is an open-source software in constant development that has gained a lot of popularity within statistical communities and sport analytics. This work presents a systematic review of the available R CRAN sport packages following PRISMA guidelines. We consider all packages active as of 18th February 2021.</p> <p>A total of 81 sport R packages created since 2006 were detected. Of these, 35.9% were authored by an American national, 69.1% include a dataset, 43.2% provide vignettes, and 65.4% have been updated at least once. The sport with the highest representation is basketball (n=14, 17.3%), followed by soccer (n=12, 14.8%) and packages related to physical activity (n=11, 13.6%). There are no sports packages directed solely for the female gender, while 59.3% packages are focused on professional athletes. Fifty packages (61.7%) are related to the Sports performance analysis category, and web scraping (n=43, 53.1%) was the main functionality in the collected packages.</p> <p>The creation of new R packages in the area of sports could help solve questions and tasks that still remain a challenge in this field, while continuing to help to improve the level of statistical education and computational thinking skills.</p> |
|                               |  |

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



1  
2  
3 *Title of the Article:*  
4

5 **A systematic review of sport related packages within the R CRAN repository**  
6

7  
8 Martí Casals<sup>1,2,3</sup>, José Fernández<sup>4</sup>, Victor Martínez<sup>5</sup>, Michael Lopez<sup>6</sup>, Klaus Langohr<sup>7</sup>,  
9  
10 Jordi Cortés<sup>7</sup>  
11  
12  
13  
14  
15

16 <sup>1</sup>Sport and Physical Activity Studies Centre, Faculty of Medicine, Universitat de Vic -  
17  
18 Universitat Central de Catalunya, Vic, Catalunya, Spain  
19

20 <sup>2</sup>Sport Performance Analysis Research Group (SPARG), University of Vic - Central  
21  
22 University of Catalonia, Vic, Barcelona, Spain.  
23

24  
25  
26 <sup>3</sup>National Institute of Physical Education of Catalonia (INEFC), University of Barcelona  
27  
28 (UB), Barcelona, Spain  
29

30  
31 <sup>4</sup>Faculty of Health Sciences, School of Behavioral and Health Sciences, Australian  
32  
33 Catholic University, Fitzroy, VIC, Australia  
34

35  
36 <sup>5</sup>Universitat Politècnica de Catalunya, Barcelona Spain  
37

38  
39 <sup>6</sup>National Football League, New York, NY  
40

41  
42 <sup>7</sup>Department of Statistics and Operations Research, Universitat Politècnica de  
43  
44 Catalunya, Jordi Girona, 31, 08034 Barcelona, Spain  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 **Correspondence to** Dr Martí Casals, Sport and Physical Activity Studies Centre  
4  
5 (CEEAF), University of Vic-Central University of Catalonia (UVic-UCC);  
6  
7 [marticasals@gmail.com](mailto:marticasals@gmail.com), [marti.casals1@umedicina.cat](mailto:marti.casals1@umedicina.cat)  
8  
9

10 Martí Casals

11  
12 Sport and Physical Activity Studies Centre (CEEAF), University of Vic – Central University of  
13 Catalonia (UVic-UCC), C. Dr. Antoni Vilà Cañellas, s/n, 08500 Vic, Spain.

14  
15 ORCID: <https://orcid.org/0000-0002-1775-8331>

16  
17 Twitter: @CasalsTMarti

18  
19 His research interests are Statistics of Education, Sports Statistics, Biostatistics and Statistical  
20 Thinking.  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

## 1 **Abstract**

2 The sports statistics community is rapidly evolving. R is an open-source software in  
3 constant development that has gained a lot of popularity within statistical communities  
4 and sport analytics. This work presents a systematic review of the available R CRAN  
5 sport packages following PRISMA guidelines. We consider all packages active as of  
6 18th February 2021.

7 A total of 81 sport R packages created since 2006 were detected. Of these, 35.9% were  
8 authored by an American national, 69.1% include a dataset, 43.2% provide vignettes,  
9 and 65.4% have been updated at least once. The sport with the highest representation is  
10 basketball (n=14, 17.3%), followed by soccer (n=12, 14.8%) and packages related to  
11 physical activity (n=11, 13.6%). There are no sports packages directed solely for the  
12 female gender, while 59.3% packages are focused on professional athletes. Fifty  
13 packages (61.7%) are related to the Sports performance analysis category, and web  
14 scraping (n=43, 53.1%) was the main functionality in the collected packages.

15 The creation of new R packages in the area of sports could help solve questions and  
16 tasks that still remain a challenge in this field, while continuing to help to improve the  
17 level of statistical education and computational thinking skills.

18  
19  
20 **Keywords:** statistical software; R; sports statistics; sports analytics; sports science;  
21 statistical education

22

## 23 **Introduction**

24 Statistics is a discipline that gathers interest not only amongst professionals and  
25 academics but also from sports participants, including managers, analysts, video analysts,  
26 journalists, trainers, physiotherapists, psychologists, nutritionists, coaches, fans, and even  
27 athletes themselves.

28 The sports statistics community has grown and evolved in recent years thanks in part to  
29 the promotion of the section on Statistics in Sports  
30 (<https://community.amstat.org/sis/home>) of the American Statistical Association (ASA),  
31 which was created in 1992. Scientific journals such as the Journal of Quantitative  
32 Analysis in Sports (JQAS), which was launched in 2005, was the first academic journal  
33 dedicated to the statistical analysis of sport and has become well-known worldwide. More  
34 recently, the Journal of Sports Analytics originated in 2015, and includes a multitude of  
35 sport executives on its advisory board. Other statistics journals, such as Chance, The  
36 American Statistician, and The Statistician, regularly publish articles promoting statistical  
37 analysis of sports data, along with books related to sports statistics <sup>1</sup>.

38 On college campuses, there are a multitude of opportunities for students and professionals  
39 to become involved. Research groups dedicated to sports statistics have been formed at  
40 different universities (e.g.: Harvard, Simon Fraser, Carnegie Mellon, California, and  
41 Berkeley). Additionally, formal sport statistics education (units / credits) as part of  
42 undergraduate and postgraduate courses at universities (countries like the US, UK and  
43 Australia have increased the offer of university courses related to sports analytics in recent  
44 years). In the US, Syracuse, among other universities, now offer sports analytics major.  
45 Several conferences about this topic are held regularly (e.g.: Joint Statistical Meetings  
46 (JSM), Carnegie Mellon Sports Analytics Conference, New England Symposium on

1  
2  
3 47 Statistics in Sports (NESSIS), Cascadia Symposium on Statistics in Sports (CASSIS),  
4  
5 48 UConn Sports Analytics Symposiums). The interest and evolution of applied statistics in  
6  
7 49 sports science in academia have also grown alongside other technological innovations  
8  
9 50 currently available within sport domains, new specializations of data science in sports, or  
10  
11 51 also in sports medicine, and scientific articles in computer vision, and sports science  
12  
13 52 journals <sup>2</sup>.

14  
15  
16  
17  
18 53 Nowadays, sports analytics departments have been formed or expanded at different  
19  
20 54 professional clubs depending on the country and sport, enabling better collaboration  
21  
22 55 between coaching staff members (analysts, physical trainers, physiotherapists, coaches)  
23  
24 56 and statisticians or data scientists <sup>3-5</sup>.

25  
26  
27  
28 57 The interest in sports statistics in both academia and the sports industry has also coincided  
29  
30 58 with the boom and recognition of the statistician and especially the data scientist as the  
31  
32 59 sexiest profession of the 21st century according to Harvard Business Review <sup>6</sup>. Data  
33  
34 60 science or sports statistician roles are already in demand in the sports industry at  
35  
36 61 professional clubs in different sports leagues in the United States (NBA, MLB, MLS,  
37  
38 62 NFL) and even in Europe (soccer leagues in particular). Statistical thinking skills  
39  
40 63 developed in doctoral or master's degrees in statistics or related fields as well as  
41  
42 64 computational thinking with the statistical R or Python frameworks are usually requested  
43  
44 65 as part of these job opportunities.

45  
46  
47  
48  
49  
50 66 The current work presents an in-depth review of sport-related packages within the R  
51  
52 67 ecosystem. R is a programming language for statistical computing and graphics supported  
53  
54 68 by the R Core Team and the R Foundation for Statistical Computing<sup>7</sup>. RStudio is an  
55  
56 69 integrated development environment (IDE) for R. It includes a console, syntax-

1  
2  
3 70 highlighting editor that supports direct code execution, as well as tools for plotting,  
4  
5 71 history, debugging and workspace management.  
6  
7  
8

9 72 As a free and open-source software with constant improvements that allow a more  
10  
11 73 comfortable and practical workspace, R has gained a lot of popularity within statistical  
12  
13 74 computing and also in the world of sports statistics. R capabilities are extended through  
14  
15 75 user-created packages <sup>8</sup>. The R ecosystem is constantly evolving, as the available  
16  
17 76 packages are updated over time, and new ones are dynamically created. Packages contain  
18  
19 77 code, data, and documentation in a standardized collection format that can be installed by  
20  
21 78 users of R, typically via a centralized software repository such as the Comprehensive R  
22  
23 79 Archive Network (CRAN), Bioconductor, or GitHub <sup>9</sup>. CRAN is the best known and most  
24  
25 80 popular R repository because, over and above other repositories, it is one of the few that  
26  
27 81 routinely performs checks and makes regular checks of contributed packages.  
28  
29  
30  
31

32  
33 82 The location where the packages are stored is called the library <sup>10</sup>. Functions are  
34  
35 83 commonly distributed within R packages. Functions are “self-contained” modules of code  
36  
37 84 that accomplish a specific task. Functions usually take in some sort of data structure  
38  
39 85 (value, vector, data frame etc.), process it, and return a result <sup>11</sup>.  
40  
41  
42

43 86 In recent years, there has been an increase in the creation of tutorials, courses, books, and  
44  
45 87 even package applications within the R Ecosystem tailored to different sporting  
46  
47 88 environments <sup>1,12–15</sup>. In other words, the key probably remains in statistical education to  
48  
49 89 advance sports statistics from future generations.  
50  
51

52  
53 90 Sports statistics are particularly well-suited for the use in education because of the wealth  
54  
55 91 and availability of data, the pre-existing knowledge of sports among students, and the  
56  
57 92 ease with which it can be used to promote data literacy. The numerous sources of publicly  
58  
59 93 available data make solving real-world problems an option that is not available in many  
60



1  
2  
3 94 other fields. Since many students are familiar with the rules and strategies in sports, they  
4  
5 95 have the subject-matter expertise necessary to more easily understand how to  
6  
7 96 appropriately solve these problems. This expertise also makes sports analytics examples  
8  
9  
10 97 great options for non-technical introductions to statistical reasoning and statistical  
11  
12 98 thinking. A recent book by Benjamin Baumer et al<sup>16</sup> introduces examples of sabermetrics  
13  
14 99 and data with some R packages to show the possibilities of statistical education and sports  
15  
16  
17 100 analytics. Sport is a convenient and familiar context to use in teaching or in demonstrating  
18  
19 101 a new statistical method. Just at the 2021 JSM conference one of the invited panels of the  
20  
21 102 Statistics in Sports section was entitled "Using Sports Analytics to Inspire Student Interest  
22  
23 103 in Statistics and Storytelling with Data".

24  
25  
26 104 Therefore, given the growing interest in statistics in the field of sport, and the evolution  
27  
28 105 of the popularity of the R Ecosystem, we determined it appropriate to make a systematic  
29  
30 106 review of the available R packages related to sports.  
31  
32  
33  
34  
35 107

1  
2  
3 108 **Methods**

4  
5  
6 109 **Systematic Review**

7  
8  
9 110 We consider all packages active as of 18th February 2021.

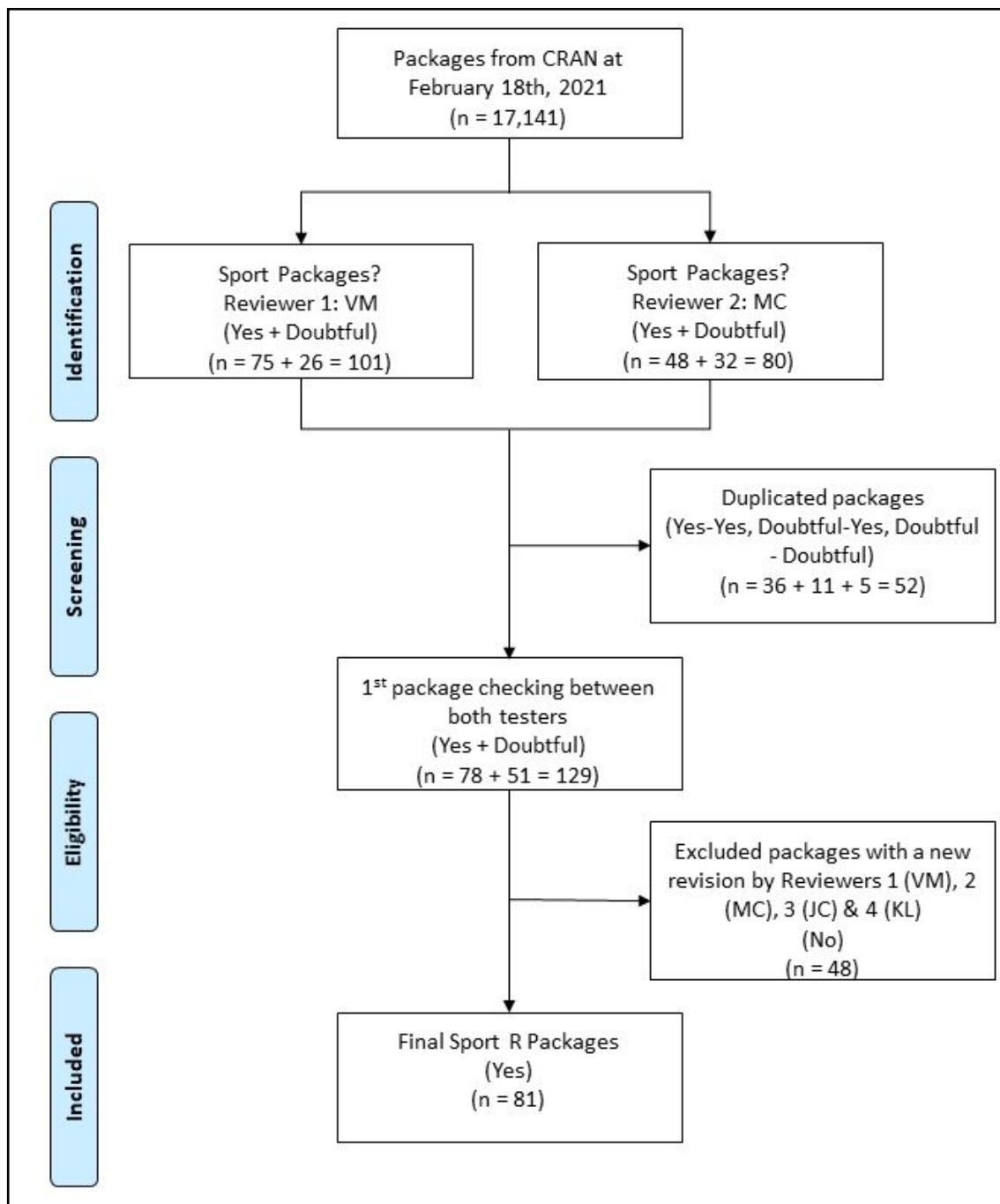
10  
11  
12 111 The approach implemented to conduct this systematic review follows the PRISMA  
13  
14 112 statement <sup>17</sup>, which includes a flow chart that describes the workflow of the whole review  
15  
16 113 process.

17  
18  
19 114 17,141 R packages were present in the CRAN repository on February 18th, 2021. VM  
20  
21 and MC conducted a first independent manual search of sports packages. Regarding the  
22  
23 115 definition of sport, we considered all those sports mentioned on the web *topend* sports  
24  
25 116 (<https://www.topendsports.com/>). In the first phase of this systematic review, both VM  
26  
27 117 and MC took into consideration keywords related to the sport (leagues, instruments, or  
28  
29 118 similar words). These terms could specifically be the name of sports (e.g, Football,  
30  
31 119 Basketball, Soccer), sports competitions (e.g, NBA, MLB, NFL), eSports (LoL, Dota2,  
32  
33 120 Fortnite), accessories or sports instruments (e.g, Accelerometer, Racket, Ball),  
34  
35 121 federations or information channels (e.g, ESPN, FIBA, FIFA), or other words also related  
36  
37 122 to sport (e.g, Sport, Cup, Game). VM and MC found a total of 101 and 80 packages,  
38  
39 123 respectively. The Matthews correlation coefficient <sup>18</sup> was 0.61 between both reviewers,  
40  
41 124 which represents a strong positive agreement despite the classification of some packages  
42  
43 125 within the sports settings was unclear. Among these, there were 52 repeated packages  
44  
45 126 leading to 129 packages after the first phase. Of those, 51 (39.53%) packages were  
46  
47 127 classified as questionable, and they were reviewed in more detail. JC and KL helped to  
48  
49 128 decide the suitability of these packages. Among the 51 questionable sports R packages,  
50  
51 129 only 3 packages were incorporated in the final list giving the 81 R sports packages from  
52  
53 130 the CRAN repository (see Figure 1).  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 132 To validate our manual search, an R script was created using the *findFn* function from the  
4  
5 133 *sos* package. The number of packages returned was reduced to 3,051. This number was  
6  
7 134 so high because generic words such as *Running*, *Ball*, *Game*, *Cup*, or *Score* were included  
8  
9 135 in the list, and these can be found as a part of other words, or they have a different meaning  
10  
11 136 unrelated to sport. Finally, the use of this function was discarded due to its low specificity.  
12  
13  
14

15 137  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

For Peer Review

138 **Figure 1:** PRISMA: Flowchart of the systematic review of R packages

139

140 **Collected information**

141 Once the final list was obtained, some variables related to the characteristics of the  
 142 package; the characteristics of the sport; and statistical topics were collected. Table 1  
 143 shows the variables collected within each of these categories.

144

145 **Table 1:** Variable characteristics by package

| Variable                                       | Description   |
|--|---|
| <b>General characteristics of the packages</b> |   |
| Tutorial                                       | <i>(Yes/No). If the package contains a vignette/tutorial</i>                              |
| Creation date                                  | Date of the first version   |
| Initial version                                | Number of the initial version   |
| Last version date                              | Date of the last version of the package   |
| Current version                                | Number of the current version   |
| R journal                                      | <i>(Yes/No) the package has a paper related in R Journal</i>                              |
| Citation                                       | Package citation  |
| Creation Year                                  | Year the package was launched   |
| Current Year                                   | Year of the last version  |
| Creator's Country                              | Country of the creator or the first contributor of the package                            |
| Creator's Gender                               | Male/Female   |
| <b>General characteristics of the sport</b>    |   |
| Sport  | The sport which the package is focused on   |
| Gender   | Gender which is the package is focused on ( <i>Male, female o Both</i> )                  |
| Category                                       | Professional category ( <i>Professional, Amateur o Both</i> )                             |
| JQAS   | <i>(Yes/No) if the package is mentioned in the Journal of Quantitative Analysis</i>       |
| Category                                       | Five categories: 1) <i>Sports Performance Analysis</i> ; 2) <i>Sports technology</i> ; 3) |
| <b>Data and statistical methodology</b>        |   |
| Data   | <i>(Yes/no) If the package contains any dataset</i>                                       |
| Data analysis                                  | <i>(Yes/no) If the purpose of the package is the data analysis</i>                        |
| Methodology                                    | The main type of statistical techniques used  |

146 The complete dataset can be found at

147 [https://github.com/jordicortes40/CRAN\\_R\\_Sports\\_Packages](https://github.com/jordicortes40/CRAN_R_Sports_Packages). Furthermore, a R Shiny

148 web app was developed to enable readers to dynamically query the list of R sport

149 packages included in this review and can be found on <https://shiny->

150 [eio.upc.edu/pubs/sport-R-packages/](https://shiny-eio.upc.edu/pubs/sport-R-packages/).

1  
2  
3 151 **Statistical analysis**  
4  
5

6 152 Categorical variables are described with absolute and relative frequencies. R, version  
7  
8 153 4.0.2., was used to perform the analyses.  
9

10  
11  
12 154 **Results**  
13

14  
15 155 **Descriptive analysis**  
16

17  
18 156 Of the total of 81 selected packages, 46 (56.8%) do not contain user tutorials, while 35  
19  
20 157 (43.2%) provide vignettes. Only 3 of the 81 packages (3.7%) are presented in R journal  
21  
22 158 articles. Twenty-nine packages (35.8%) were written by a developer or author from the  
23  
24 159 United States of America, 9 (11.1%) from the United Kingdom, and 6 (7.4%) from  
25  
26 160 Canada, respectively. Moreover, most of the packages have been developed by males  
27  
28 161 (n=73; 90.1%). To be more accurate, if we calculate a ratio per population of each of the  
29  
30 162 countries, the rates would be 0.08 per 10<sup>6</sup> inhabitants of the US, 0.13 per 10<sup>6</sup> inhabitants  
31  
32 163 of the UK, and 0.15 per 10<sup>6</sup> inhabitants of Canada, respectively.  
33  
34  
35

36  
37 164 Regarding the version and update of the revised packages, 53 packages (65.4%) have  
38  
39 165 been updated at least once (Table 2).  
40  
41

42 166 **Table 2:** Description of the collected variables  
43  
44

| Variable (N = 81)                             | Category | n (%)      |
|---|----------|------------|
| <b>General characteristics of the package</b> |          |            |
| Tutorial                                      | Yes      | 35 (43.2%) |
| Package update                                | Yes      | 53 (65.4%) |
| R Journal                                     | Yes      | 3 (3.7%)   |
| Creator's country                             | USA      | 29 (35.8%) |
|   | UK       | 9 (11.1%)  |
|   | Canada   | 6 (7.4%)   |

|   |                                    |            |
|---|------------------------------------|------------|
| Creator's gender                            | Male                               | 73 (90.1%) |
| <b>General characteristics of the sport</b> |                                    |            |
| Sports                                      | <i>American Football</i>           | 9 (11.1%)  |
|   | <i>Baseball</i>                    | 7 (8.6%)   |
|   | <i>Basketball</i>                  | 14 (17.3%) |
|   | <i>Cycling</i>                     | 9 (11.1%)  |
|   | <i>Football</i>                    | 12 (14.8%) |
|   | <i>Ice Hockey</i>                  | 5 (6.2%)   |
|   | <i>Physical Activity</i>           | 11 (13.6%) |
|   | <i>Running</i>                     | 5 (6.2%)   |
|   | <i>Other*</i>                      | 32 (39.5%) |
| Gender                                      | <i>Male</i>                        | 38 (46.9%) |
|   | <i>Female</i>                      | 0 (0%)     |
|   | <i>Both</i>                        | 43 (53.1%) |
| Category                                    | <i>Amateur</i>                     | 10 (12.3%) |
|   | <i>Professional</i>                | 48 (59.3%) |
|   | <i>NS*</i>                         | 23 (28.4%) |
| JQAS  | Yes                                | 2 (2.5%)   |
| Category classification                     | <i>Athlete Health</i>              | 4 (4.9%)   |
|   | <i>eSports</i>                     | 3 (3.7%)   |
|   | <i>Movement integration</i>        | 9 (11.1%)  |
|   | <i>Sports performance analysis</i> | 50 (61.7%) |
|   | <i>Sports technology</i>           | 15 (18.5%) |
| <b>Data and statistical methodology</b>     |                                    |            |
| Data  | Yes                                | 56 (69.1%) |
|   | No                                 | 22 (27.2%) |
|   | Yes (dependence on other packages) | 3 (3.7%)   |
| Statistical methodology <sup>§</sup>        | Yes                                | 40 (49.4%) |
| Methodology type                            | <i>Compute</i>                     | 17 (21%)   |

|  |                      |            |
|--|----------------------|------------|
|  | <i>Descriptive</i>   | 7 (8.6%)   |
|  | <i>Prediction</i>    | 8 (9.9%)   |
|  | <i>Processing</i>    | 14 (17.2%) |
|  | <i>Scraping</i>      | 43 (53.1%) |
|  | <i>Simulation</i>    | 7 (8.6%)   |
|  | <i>Visualization</i> | 8 (9.9%)   |
|  | <i>Other**</i>       | 17 (21%)   |

167 Other\*: Arbitrage, Athletics, Australian football, Billiard, Chess, Cricket, Darts, Dota2, Injury Sports Medicine, Mahjong, Swimming,

168 Scuba diving, Softball, Speedway, Squash, Tennis, Volleyball, Walking.

169 Other\*\*: Categorization, Classification, Decoding, Extraction, Modeling, Preprocessing, Probabilities.

170 NS \*: Not specified.

171 <sup>s</sup>The classification of the packages for this variable is based on the description in the CRAN reference manual. If the description  
 172 explicitly indicates that the library contained statistical methodology, it was marked with a Yes. Even so, some packages could contain  
 173 some methodology not mentioned in the package description.

174

175 Figure 2A shows the number of packages created by year in the period from 2006 to 2021.

176 The first active package was created in 2006, and it was not until 2010 that a 2nd package  
 177 was created. Since 2010, there is a growing trend indicating that, generally each year, the  
 178 creation of sports-related R packages increases. However, it is worth mentioning that R  
 179 packages created many years ago may not be maintained and therefore we may have  
 180 missed them in our review. Figure 2B shows the number of packages according to the  
 181 year of their most recent update. Again, a growing trend is observed: for example, there  
 182 are 48 sports-related packages that have an update in 2020 or 2021, pointing out the  
 183 growth of sports topics within the R community.

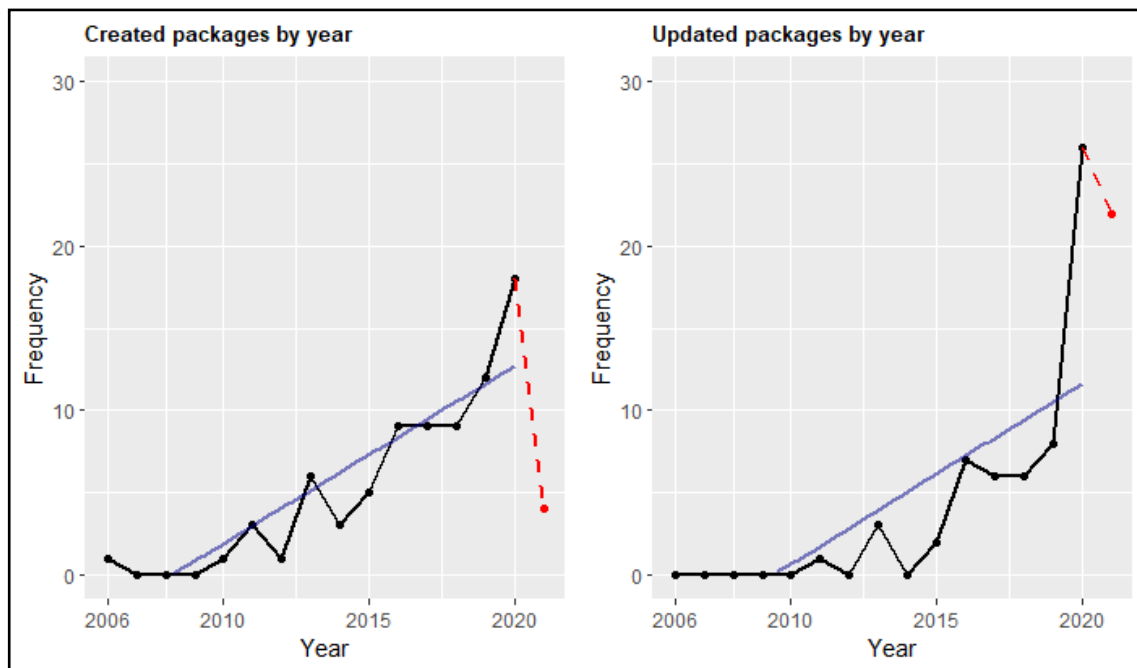
184

185

186



187 **Figure 2:** Number of packages created (left) and updated (right) by year from 2006 to  
 188 2021. The year 2021 is represented in red because data was collected until 02/18/2021.

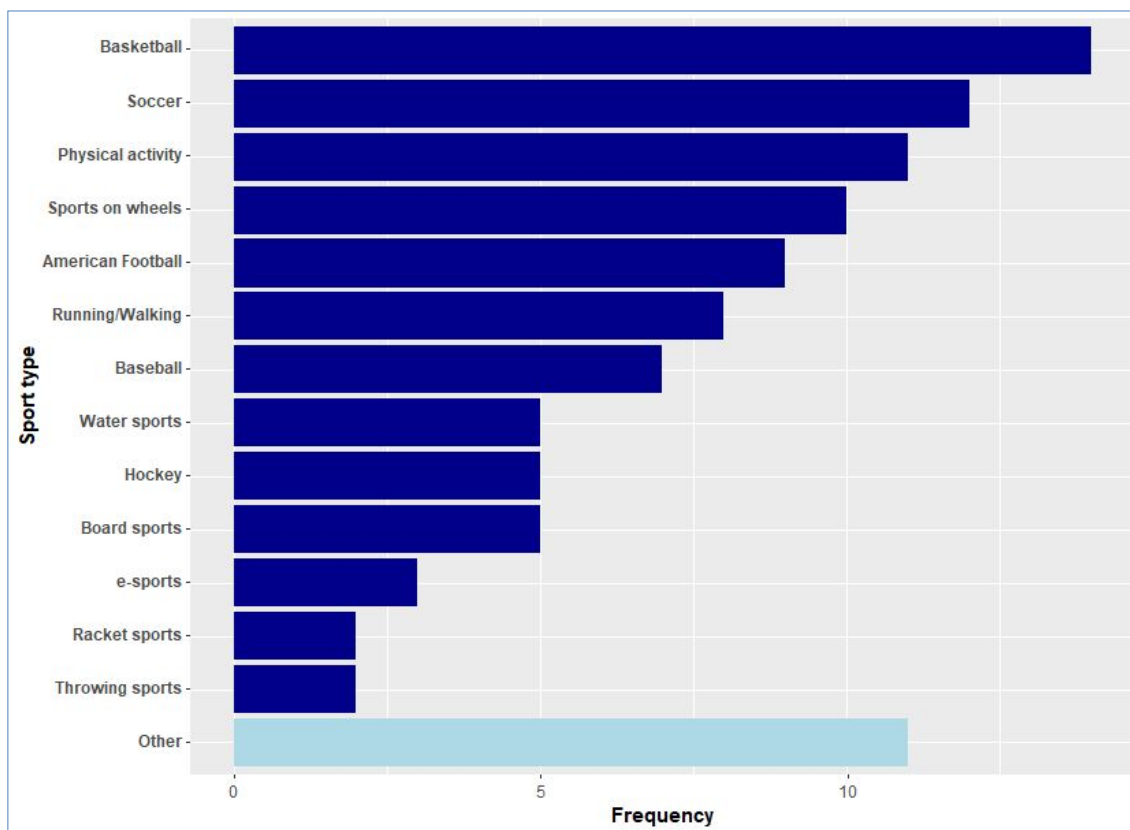


189  
 190 Regarding the sports characteristics, we found packages that are related to more than one  
 191 discipline ( $n=11$ , 13.6%), and therefore have multiple responses. For this reason, the total  
 192 sum of the number of sports ( $n = 104$ ) exceeds the number of packages ( $n = 81$ ). The sport  
 193 with the highest representation in the collected packages is basketball ( $n=14$ , 17.3%),  
 194 followed by soccer ( $n=12$ , 14.8%) and packages related to physical activity ( $n=11$ ,  
 195 13.6%). In Figure 3, we group some sports with similar characteristics. Sports with little  
 196 representation (arbitrage, Australian football, cricket, sport injuries, softball, and  
 197 volleyball) without being grouped were merged into the *Other* category ( $n=11$ , 13.6%)  
 198 (Figure 3).

199 There are no sports packages directed solely for the female gender, while there are 43  
 200 packages (53.1%) focused on both genders and the remaining 38 (46.9%) are intended  
 201 solely for male sports. On the other hand, 48 (59.3%) libraries are focused on professional  
 202 athletes.

1  
2  
3 203 In the same way, as with the *R Journal*, we also looked at which packages appeared in  
4  
5 204 articles in JQAS. Only 2 (2.5%) packages are mentioned in this journal. Fifty packages  
6  
7 205 (61.7%) are related to the *Sports performance* analysis category, followed by the category  
8  
9 206 *Sports technology* (n=15, 18.5%). Regarding the data availability, 56 packages (69.1%)  
10  
11 207 include a dataset, and 3 packages (3.7%) use databases from other libraries. Regarding  
12  
13 208 the statistical analysis methodology, 41 packages (51.6%) do not contain methodology  
14  
15 209 (e.g, one package that contains colors of different equipment to create graphics or  
16  
17 210 interfaces to access data).  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

224 **Figure 3:** Frequency of each sport type in collected R sports packages.

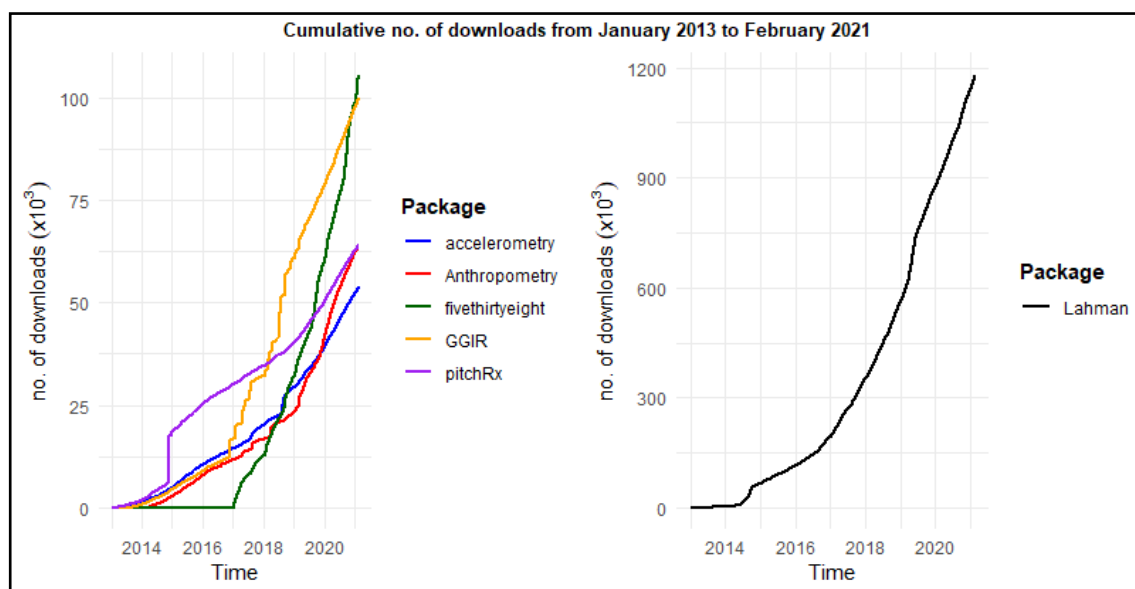


225

226 For the variable type of methodology, there are packages that use several statistical  
 227 analysis methods, so here we have again a multiple response. Web scraping (n=43,  
 228 53.1%) was the main functionality in the collected packages followed by Compute or  
 229 Wrangling Data (n=17, 21%) and data processing (n=14, 17.2%). Prediction and  
 230 Visualization can also be highlighted (n=8, 9.9% each one).

231 Figure 4 shows the popularity of R packages related to sports measured as the number of  
 232 downloads from the CRAN repository. The figure informs about the cumulative  
 233 downloads sum to get an idea about the rate of increase in download numbers since 2013.  
 234 The top sport package downloaded as of January 2013 was *Lahman* (Friendly, et al.  
 235 2021), followed by another baseball package (*pitchRx*), and sports science and  
 236 accelerometer related packages such as *accelerometry* and *GGIR*.

237 **Figure 4:** Cumulative number of downloads for the six most successful packages since  
 238 2013. At the right, *Lahman* was the most downloaded package followed by the  
 239 represented packages at the left. Vertical axes differ in the two panels.



240

## 241 Discussion

242 The R packages selected in this review showed that since 2010 there has been a growing  
 243 trend in the creation of sports-related R packages. The first sports package currently  
 244 active in CRAN was published in 2006, and anecdotally, the book *Moneyball*<sup>20</sup> was  
 245 published in 2003. As Mora-Cantalops et al pointed, evolution of the CRAN repository  
 246 in the last 20 years has been growing in number of packages and there is an overall  
 247 tendency to continuously updating the ecosystem and its functionalities<sup>21</sup>. However, if  
 248 we are interested in comparing the growth of R packages in sports with the growth in R  
 249 packages overall, we can observe the trend in R packages overall seems slightly higher  
 250 (Supplemental material; Figure 5). The number of packages created per year (Figure 2)  
 251 shows an important increase in 2020, probably influenced by the effects of COVID-19  
 252 and the stay-at-home safety measures implemented across the world<sup>22-27</sup>.

1  
2  
3 253 Our review shows that there is room for improvement in offering more packages for  
4  
5 254 female or amateur sports, more homogeneity between sports, and above all, more  
6  
7  
8 255 possibilities for statistical methodology.  
9

10  
11 256 The packages reviewed were predominantly in the field of sports performance analysis  
12  
13 257 or sports technology. This result is consistent with the review of the late Keith Lyons,  
14  
15 258 who was a pioneer in education and analysis of performance in sport and the interest in  
16  
17  
18 259 open sharing in courses such as “Sport Informatics and Analytics”<sup>28</sup>. The phenomenon  
19  
20 260 of Sports Analytics, better known in the United States, has begun to bridge the gap  
21  
22 261 between academia and industry in sports. We are facing a new data-oriented culture that  
23  
24 262 indirectly requires changes to the current management model of organizations, both in  
25  
26 263 academia (teaching and research) and in industry. Technology or data science companies  
27  
28 264 in sports are knocking on the door to make communication between the needs of sports  
29  
30 265 professionals a reality through statistical and computational thinking. For this, data  
31  
32 266 scientists or statisticians have an important role. The lack of packages created related to  
33  
34 267 health and sports is surprising, given the importance of the sport and exercise medicine  
35  
36 268 field and its journals, and also the influence of sports biostatistics and epidemiology in  
37  
38  
39 269 recent years<sup>4</sup>.  
40  
41  
42  
43

44 270 Out of all packages included in this review, eleven refer to physical activity, and they are  
45  
46 271 preceded by mostly team sports, not individual sports. Other sports such as cricket,  
47  
48 272 Australian football, or tennis for example, also well known in the scientific community  
49  
50 273 for different academic works, have fewer packages. Other less common sports or only  
51  
52 274 popular in specific countries do not yet have R packages in CRAN but in the future, we  
53  
54 275 will probably see more eSports packages or other popular sports such as CrossFit, etc.  
55  
56 276 This study suggests that there are still many possibilities to create future R packages in  
57  
58 277 the CRAN repository to aid domain research and statistical education.  
59  
60

1  
2  
3 278 Another aspect of improvement in terms of the characteristics of sports-related packages  
4  
5 279 is that very few have been published, for instance in R journal, they do not have a vignette  
6  
7  
8 280 or tutorial, and there is no data available. In this sense, a few packages have also been  
9  
10 281 found presented in one of the most popular sports analytics journals such as JQAS.  
11  
12 282 Nowadays, more people have more access to data than ever before. But a comparative  
13  
14 283 lack of analytical skills has resulted in scientific findings that are neither replicable nor  
15  
16  
17 284 reproducible <sup>29</sup>. It seems that in sports sciences the crisis of data reproducibility is also  
18  
19 285 beginning to be taken into account <sup>30,31</sup>. It is time to invest in statistical education.  
20  
21  
22 286 With regard to the above, the ability to use real work data examples can help motivating  
23  
24 287 future students in data analysis to try to use statistical thinking or decision making with  
25  
26  
27 288 data. Different scientific works and books have focused on statistical education from sport  
28  
29 289 as a motivational focus <sup>32-36</sup>. Regardless, this research points out the fact that there is an  
30  
31 290 important opportunity to enhance statistical education from data and real problems of  
32  
33  
34 291 sport through R packages. Motivating sport analytic students to learn statistics with  
35  
36 292 examples and specific data sets from the field of work or studies can be key to the  
37  
38 293 evolution of oneself and of the scientific field itself. The common datasets of titanic  
39  
40 294 (titanic), iris (datasets), mosaic (mosaic) among others that are used in many courses,  
41  
42  
43 295 probably do not contribute much to the engagement of students in the field of sport for  
44  
45 296 example <sup>37</sup>. The packages available in this review could be useful to educate and use  
46  
47 297 statistical techniques for descriptive, predictive, or causal statistics, apart from having  
48  
49  
50 298 other functionalities such as the development of publications, books, applications,  
51  
52 299 dashboards and presentations or web scraping. It is surprising that the main methodology  
53  
54 300 used in the revised packages is web scraping or wrangling data, given the evolution of  
55  
56  
57 301 statistics or data science as a science, even in a field such as sports. Regarding the lack of  
58  
59 302 statistical methodologies, there are packages that use data from the sports analytics field  
60

1  
2  
3 303 (among others) such as BradleyTerry2, Elo as well as other packages that are very often  
4  
5 304 applied to sports modeling<sup>38,39</sup>. These have not been taken into account as they are not R  
6  
7  
8 305 packages specifically focused on sports.

9  
10 306 Horton and Hardin pointed out that the leading thinkers of the next decade will be those  
11  
12  
13 307 who seamlessly unite tools from both statistics and computation and that our way of  
14  
15 308 thinking about statistics will be based on complementary computational thinking<sup>40</sup>. The  
16  
17  
18 309 creation of new R packages and CRAN Task view in the world of sports and the creation  
19  
20 310 of new functions and methodologies could bring us closer to improving these statistical  
21  
22 311 and computational thinking skills.

### 23 24 25 312 **Limitations of the study**

26  
27  
28  
29 313 One of the limitations of our study could be that the number of identified packages was  
30  
31 314 not high, despite the review of many packages.

32  
33  
34 315 Another possible limitation of our review is the potential bias to disregard packages that  
35  
36 316 are found in other repositories and software. This review was carried out in the repository  
37  
38  
39 317 of the Comprehensive R Archive Network (CRAN) website ([https://cran.r-](https://cran.r-project.org/web/packages/available_packages_by_name.html)  
40  
41 318 [project.org/web/packages/available\\_packages\\_by\\_name.html](https://cran.r-project.org/web/packages/available_packages_by_name.html)), although other existing  
42  
43 319 sources such as GitHub packages seem to be updated parallelly to CRAN packages. For  
44  
45 320 example, since April there have been some packages (not in R CRAN) directed at female  
46  
47 321 sports (e.g. <https://github.com/adror1/nwslR>). It is known that if authors wish for their  
48  
49 322 packages to have significant traction in the R community, they should submit it to CRAN.  
50  
51 323 Submitting to CRAN involves more work than just providing a version on GitHub, but  
52  
53 324 the vast majority of R users do not install packages from GitHub, because CRAN provides  
54  
55 325 discoverability, ease of installation and a stamp of authenticity<sup>8</sup>. Another of the most  
56  
57 326 used programming software in sports analytics (and especially in soccer) is Python<sup>41</sup>.

1  
2  
3 327 Although the GitHub source and Python software have not been included in this review,  
4  
5 328 we believe that the sample of packages presented is representative of all packages in any  
6  
7  
8 329 repository.

9  
10  
11 330 One of the strongest aspects of this study is the originality and innovation, as it is (to our  
12  
13 331 knowledge) the first systematic review carried out that about R packages in the field of  
14  
15 332 sport. The only existing similar review is the work of the late Keith Lyons mentioned  
16  
17  
18 333 above, or Ben Baumer's recent GitHub repository ([https://github.com/beanumber/ctv-](https://github.com/beanumber/ctv-sportsanalytics)  
19  
20 334 [sportsanalytics](https://github.com/beanumber/ctv-sportsanalytics)) which shows a list of existing packages related to Sports Analytics  
21  
22 335 attending to the classification by type of sports<sup>42</sup>. In a similar way, there are initiatives  
23  
24 336 focused for instance on public sport science datasets  
25  
26  
27 337 ([https://github.com/josedv82/public\\_sport\\_science\\_datasets](https://github.com/josedv82/public_sport_science_datasets)). Moreover, this work  
28  
29 338 includes a R Shiny web application for readers to explore the list of R sport packages in  
30  
31 339 CRAN and provides a platform to dynamically filter and interact with this dataset. Our  
32  
33  
34 340 hope is that more R sport packages are included in this application as they become  
35  
36 341 available.

## 39 342 **Conclusions**

40  
41  
42  
43 343 In recent years, the use of sports-related R packages has increased, especially in  
44  
45 344 professional sports such as basketball or soccer, and in the category of Sports performance  
46  
47 345 analysis. Our review shows that there is room for improvement in offering more packages  
48  
49 346 for female or amateur sports, more homogeneity between sports, and above all, more  
50  
51  
52 347 possibilities for statistical methodology.

53  
54  
55 348 Our review hopes to also show the future possibilities in R packages in this field and  
56  
57  
58 349 especially in statistical education.



1  
2  
3 350 **Funding**  
4  
5

6 351 This research was funded by the Ministerio de Ciencia e Innovación (Spain) (PID2019-  
7  
8 352 104830RB-I00) and the Departament d'Economia i Coneixement de la Generalitat de  
9  
10 353 Catalunya (Spain) (2017 SGR 622 (GRBIO)).  
11  
12

13  
14 354 **References**  
15

- 16  
17 355 1. Albert J, Glickman M, Swartz T, et al. *Handbook of statistical methods and*  
18  
19 356 *analyses in sports*, CRC Press, Boca Raton, FL, 2017.  
20  
21  
22  
23 357 2. Swartz TB. Where Should I Publish My Sports Paper? *Am Stat* 2020; 74: 103–  
24  
25 358 108.  
26  
27  
28  
29 359 3. Casals M. BIOSTATISTICS IN SPORTS SCIENCE AND SPORTS  
30  
31 360 MEDICINE. *Barça Innovation Hub*,  
32  
33 361 <https://barcainnovationhub.com/biostatistics-in-sports-science/> (2018, accessed  
34  
35 362 28 January 2022).  
36  
37  
38  
39 363 4. Casals M, Finch CF. Sports Biostatistician: a critical member of all sports science  
40  
41 364 and medicine teams for injury prevention. *Inj Prev* 2017; 23: 423–427.  
42  
43  
44  
45 365 5. Casals M, Bekker S, Finch CF. Infographic: Sports Biostatisticians as a critical  
46  
47 366 member of all sports science and medical teams for injury prevention. *Br J Sports*  
48  
49 367 *Med* 2019; 53(7):408-409.  
50  
51  
52  
53 368 6. Thomas H. Davenport DJP. Data Scientist: The Sexiest Job of the 21st Century.  
54  
55 369 *Harvard Business University*, 2012, [https://hbr.org/2012/10/data-scientist-the-](https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century)  
56  
57 370 [sexiest-job-of-the-21st-century](https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century) (2012, accessed 28 January 2022).  
58  
59  
60

- 1  
2  
3 371 7. Core T. R: A language and environment for statistical computing. *R Foundation*  
4  
5 372 *for Statistical Computing*,  
6  
7 373 [https://hero.epa.gov/hero/index.cfm/reference/details/reference\\_id/7469942](https://hero.epa.gov/hero/index.cfm/reference/details/reference_id/7469942)  
8  
9 374 (2021, accessed 28 January 2022).
- 10  
11  
12  
13 375 8. Wickham H. *R packages: organize, test, document, and share your code*,  
14  
15 376 O'Reilly Media, Inc, 2015.
- 16  
17  
18  
19 377 9. Hornik K. The comprehensive R archive network. *Wiley Online Libr* 2012; 4:  
20  
21 378 394–398.
- 22  
23  
24 379 10. Ferdi. What is the difference between a library and a package in R? - Stack  
25  
26 380 Overflow. *StackOverflow*, [https://stackoverflow.com/questions/26900228/what-](https://stackoverflow.com/questions/26900228/what-is-the-difference-between-a-library-and-a-package-in-r)  
27  
28 381 [is-the-difference-between-a-library-and-a-package-in-r](https://stackoverflow.com/questions/26900228/what-is-the-difference-between-a-library-and-a-package-in-r) (2015, accessed 28  
29  
30 382 January 2022).
- 31  
32  
33  
34 383 11. Training Team. Functions in R | Introduction to R - ARCHIVED. *Harvard Chan*  
35  
36 384 *Bioinformatics Core*, [https://hbctraining.github.io/Intro-to-R/lessons/03\\_introR-](https://hbctraining.github.io/Intro-to-R/lessons/03_introR-functions-and-arguments.html)  
37  
38 385 [functions-and-arguments.html](https://hbctraining.github.io/Intro-to-R/lessons/03_introR-functions-and-arguments.html) (2022, accessed 28 January 2022).
- 39  
40  
41  
42 386 12. Jovanovic M. Statistical Modelling for Sports Scientists: Practical Introduction  
43  
44 387 Using R (part 1), <https://osf.io/preprints/sportrxiv/dnq3m/> (2019, accessed 25  
45  
46 388 January 2022).
- 47  
48  
49  
50 389 13. Andres A. SABR101x. Sabermetrics 101: Module 4, R and RStudio. *edX and*  
51  
52 390 *Boston University*,  
53  
54 391 [https://open.bu.edu/bitstream/handle/2144/10585/SABR101x\\_M4\\_Stats\\_RepLvl](https://open.bu.edu/bitstream/handle/2144/10585/SABR101x_M4_Stats_RepLvl_I.srt?sequence=22)  
55  
56 392 [\\_I.srt?sequence=22](https://open.bu.edu/bitstream/handle/2144/10585/SABR101x_M4_Stats_RepLvl_I.srt?sequence=22) (2022, accessed 25 January 2022).
- 57  
58  
59  
60

- 1  
2  
3 393 14. Downie T. Analyzing Baseball Data with R. *ournal Stat Softw* 2019; 90: 1–4.  
4  
5  
6 394 15. Baruffa O. 26 Sport analytics | Big Book of R. *Big Book of R*,  
7  
8 395 <https://www.bigbookofr.com/sport-analytics.html> (2022, accessed 28 January  
9  
10 396 2022).
- 11  
12  
13  
14 397 16. Benjamin S. Baumer, Daniel T. Kaplan NJH. *Modern Data Science with R*. 2021.  
15  
16  
17 398 17. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to  
18  
19 399 the PRISMA Statement for Reporting Literature Searches in Systematic Reviews.  
20  
21 400 *Syst Rev* 2021; 10: 39.
- 22  
23  
24  
25 401 18. Delgado R, Tibau XA. Why Cohen’s Kappa should be avoided as performance  
26  
27 402 measure in classification. *PLoS One*; 14. Epub ahead of print 1 September 2019.  
28  
29 403 DOI: 10.1371/JOURNAL.PONE.0222916.
- 30  
31  
32  
33 404 19. Michael Friendly, Chris Dalzell MM and DM. *Package ‘Lahman’ Type Package*  
34  
35 405 *Title Sean ‘Lahman’ Baseball Database*, [https://cran.r-](https://cran.r-project.org/package=Lahman)  
36  
37 406 [project.org/package=Lahman](https://cran.r-project.org/package=Lahman) (2021, accessed 28 January 2022).
- 38  
39  
40  
41 407 20. Lewis M. *Moneyball: The art of winning an unfair game*, WW Norton &  
42  
43 408 Company, 2004.
- 44  
45  
46  
47 409 21. Mora-Cantallops M, Sicilia MÁ, García-Barriocanal E, et al. Evolution and  
48  
49 410 prospects of the Comprehensive R Archive Network (CRAN) package  
50  
51 411 ecosystem. *J Softw Evol Process*; 32. Epub ahead of print 1 November 2020.  
52  
53 412 DOI: 10.1002/SMR.2270.
- 54  
55  
56  
57 413 22. Horbach SPJM. Pandemic publishing: Medical journals strongly speed up their  
58  
59 414 publication process for COVID-19. *Quant Sci Stud* 2020; 1: 1056–1067.  
60

- 1  
2  
3 415 23. Kousha K, Thelwall M. COVID-19 publications: Database coverage, citations,  
4  
5 416 readers, tweets, news, Facebook walls, Reddit posts. *Quant Sci Stud* 2020; 1:  
6  
7 417 1068–1091.  
8  
9  
10  
11 418 24. Casado-Aranda L-A, Sánchez-Fernández J, Viedma-del-Jesús MI. Analysis of  
12  
13 419 the scientific production of the effect of COVID-19 on the environment: A  
14  
15 420 bibliometric study. *Environ Res* 2021; 193: 110416.  
16  
17  
18  
19 421 25. Vancini RL, Andrade MS, Viana RB, et al. Physical exercise and COVID-19  
20  
21 422 pandemic in PubMed: Two months of dynamics and one year of original  
22  
23 423 scientific production. *Sport Med Heal Sci* 2021; 3: 80–92.  
24  
25  
26  
27 424 26. Evans AB, Blackwell J, Dolan P, et al. Sport in the face of the COVID-19  
28  
29 425 pandemic: towards an agenda for research in the sociology of sport. *Taylor Fr*  
30  
31 426 2020; 17: 85–95.  
32  
33  
34  
35 427 27. Parasher V, Shekhawat KK, Bhati S, et al. Behavioural Changes in the Personal  
36  
37 428 and Professional Lives of Doctors during COVID-19 Pandemic: A Cross-  
38  
39 429 sectional Observational Study. *J Clin DIAGNOSTIC Res*. Epub ahead of print  
40  
41 430 2021. DOI: 10.7860/JCDR/2021/48153.14645.  
42  
43  
44  
45 431 28. WikiEducator. Sport Informatics and Analytics/Pattern Recognition/Using R -  
46  
47 432 WikiEducator. *WikiEducator*,  
48  
49 433 [https://wikieducator.org/Sport\\_Informatics\\_and\\_Analytics/Pattern\\_Recognition/](https://wikieducator.org/Sport_Informatics_and_Analytics/Pattern_Recognition/Using_R)  
50  
51 434 [Using\\_R](https://wikieducator.org/Sport_Informatics_and_Analytics/Pattern_Recognition/Using_R) (2018, accessed 28 January 2022).  
52  
53  
54  
55 435 29. Peng R. The reproducibility crisis in science: A statistical counterattack.  
56  
57 436 *Significance* 2015; 12: 30–32.  
58  
59  
60

- 1  
2  
3 437 30. Whiteley R, Steventon-Lorenzen N, Maniar N, et al. You Can't Replicate What  
4  
5 438 You Can't See: A Call for Researchers to Share Their Data and Avoid "Fragile"  
6  
7 439 Correlations. *J Orthop Sport Phys Ther* 2021; 51: 556–558.  
8  
9  
10  
11 440 31. Abt G, Jobson S, Morin J-B, et al. Raising the bar in sports performance research.  
12  
13 441 *J Sports Sci* 2022; 1–5.  
14  
15  
16 442 32. Albert J. A Baseball Statistics Course. *J Stat Educ*; 10. Epub ahead of print  
17  
18 443 January 2002. DOI: 10.1080/10691898.2002.11910663.  
19  
20  
21  
22 444 33. Kvam P, Education JS-IT on, 2004 undefined. Teaching statistics with sports  
23  
24 445 examples. *pubsonline.informs.org* 2004; 5: 75–87.  
25  
26  
27  
28 446 34. Bickel JE. Teaching Decision Making with Baseball Examples. *INFORMS Trans*  
29  
30 447 *Educ* 2004; 5: 2–9.  
31  
32  
33 448 35. Jian L. A Preliminary Study and Practice on the Teaching Reforme of 'Sport  
34  
35 449 Statistics'-- 《Sport Science and Technology》 2004年02期. *Sport Sci Technol*; 2,  
36  
37 450 [https://en.cnki.com.cn/Article\\_en/CJFDTTotal-TYKJ200402022.htm](https://en.cnki.com.cn/Article_en/CJFDTTotal-TYKJ200402022.htm) (2004,  
38  
39 451 accessed 28 January 2022).  
40  
41  
42  
43  
44 452 36. Barneva RP, Brimkov VE. Teaching Data Analytics to Sport Management  
45  
46 453 Majors: A Practical Approach. *Commun Comput Inf Sci* 2019; 1011: 408–417.  
47  
48  
49  
50 454 37. Kaplan D, Horton NJ, Pruum R, et al. The mosaic Package: Helping Students to  
51  
52 455 Think with Data Using R. *researchgate.net*. Epub ahead of print 2017. DOI:  
53  
54 456 10.32614/RJ-2017-024.  
55  
56  
57  
58 457 38. Heinzen E. elo: Ranking Teams by Elo Rating and Comparable Methods. *R*  
59  
60 458 *package version 3.0.1*, <https://cran.r-project.org/package=elo> (2022).

- 1  
2  
3 459 39. Turner H, Firth D. Bradley-Terry Models in R : The BradleyTerry2 Package. *J*  
4  
5 460 *Stat Softw*; 48. Epub ahead of print 2012. DOI: 10.18637/jss.v048.i09.  
6  
7  
8  
9 461 40. Horton NJ, Hardin JS. Integrating Computing in the Statistics and Data Science  
10  
11 462 Curriculum: Creative Structures, Novel Skills and Habits, and Ways to Teach  
12  
13 463 Computational Thinking. *J Stat Data Sci Educ* 2021; 29: S1–S3.  
14  
15  
16 464 41. Haaren J Van. Soccer Analytics 2020 Review · Jan Van Haaren. *Jan Van*  
17  
18 *Haaren*, <https://janvanhaaren.be/2020/12/30/soccer-analytics-review-2020.html>  
19 465  
20  
21 466 (2020, accessed 28 January 2022).  
22  
23  
24 467 42. Baumer B.S NQ, G.J. and M. CRAN Task View: Sports Analytics. *CRAN Task*  
25  
26  
27 468 *View: Sports Analytics* 2022; Version 2022-05-30.  
28  
29  
30 469  
31  
32

### 33 470 **Table Captions**

- 34  
35  
36 471 Table 1: Variable characteristics by package  
37  
38  
39 472 Table 2: Description of the collected variables  
40  
41  
42

43 473

### 44 45 46 474 **Figure Captions**

- 47  
48  
49 475 Figure 1: PRISMA: Flowchart of the systematic review of R packages  
50  
51  
52 476 Figure 2: Number of packages created (left) and updated (right) by year from 2006 to  
53  
54  
55 477 2021.  
56  
57  
58 478 Figure 3: Frequency of each sport type in collected R sports packages.  
59  
60

1  
2  
3 479 Figure 4: Cumulative number of downloads for the six most successful packages since  
4  
5 480 2013. At the right, *Lahman* was the most downloaded package followed by the  
6  
7 481 represented packages at the left. Vertical axes differ in the two panels.  
8  
9

10  
11 482  
12

13  
14 483  
15

16  
17 484  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

For Peer Review

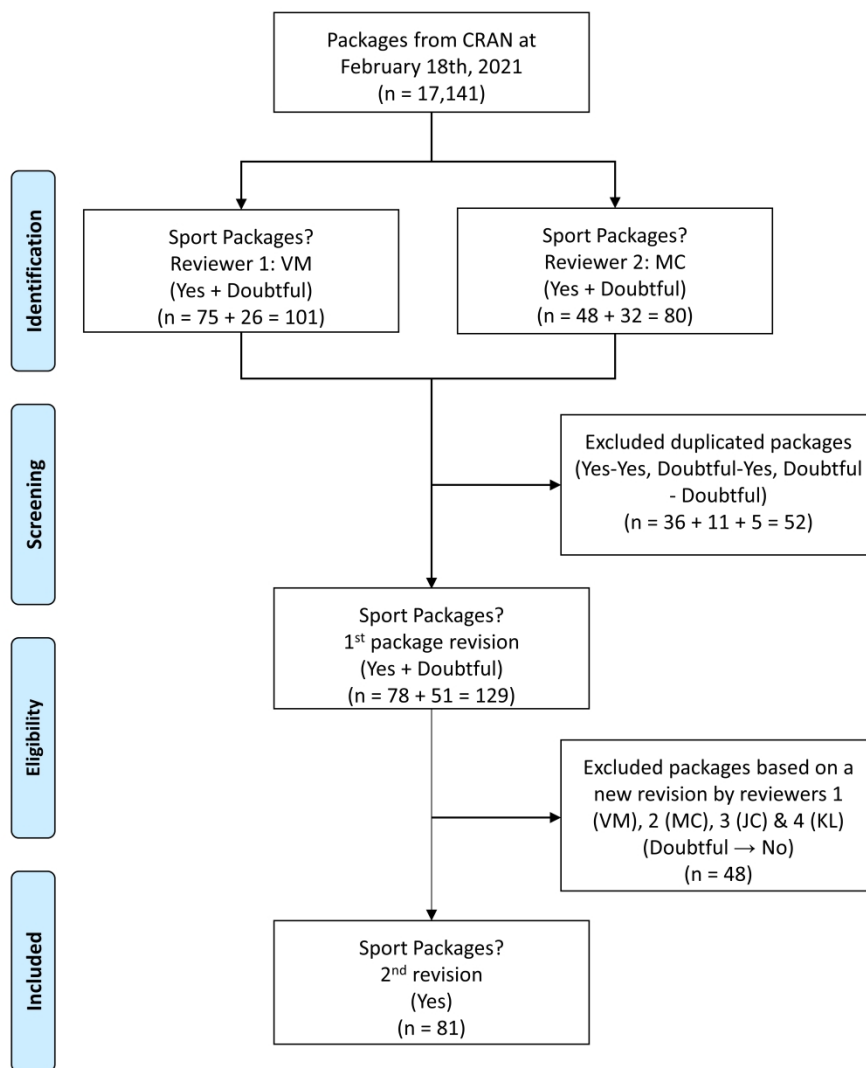


Figure 1: PRISMA: Flowchart of the systematic review of R packages

484x533mm (302 x 302 DPI)



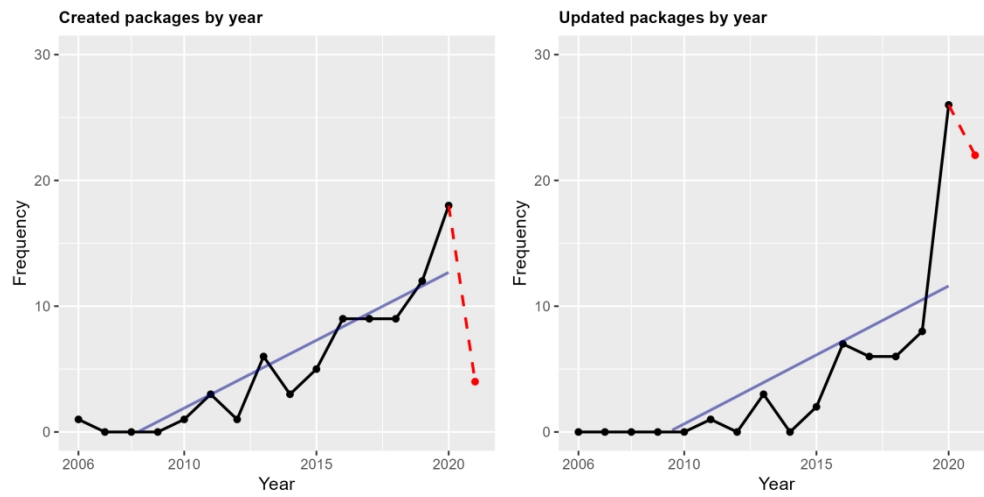


Figure 2: Number of packages created (left) and updated (right) by year from 2006 to 2021.

542x271mm (118 x 118 DPI)

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

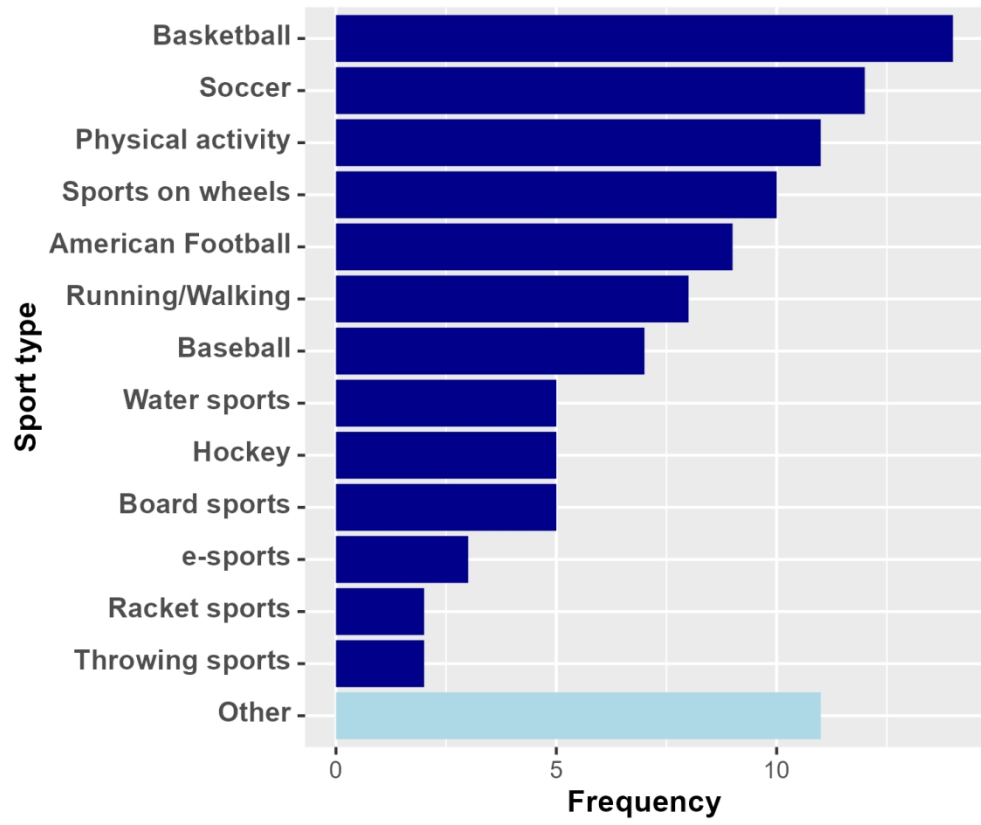


Figure 3: Frequency of each sport type in collected R sports packages.

322x271mm (118 x 118 DPI)

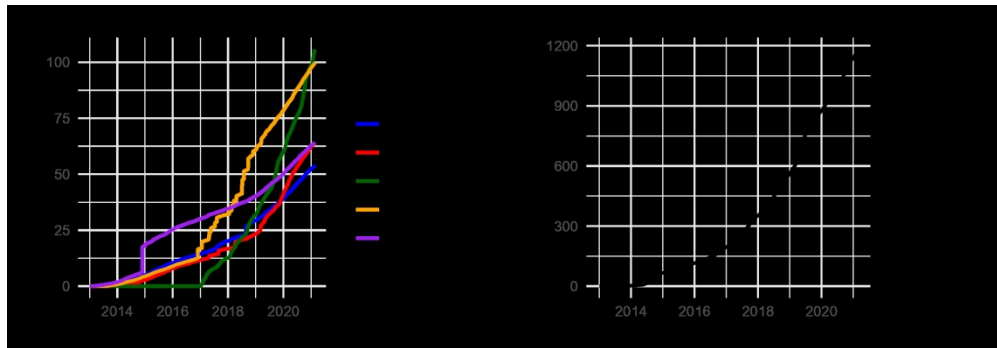
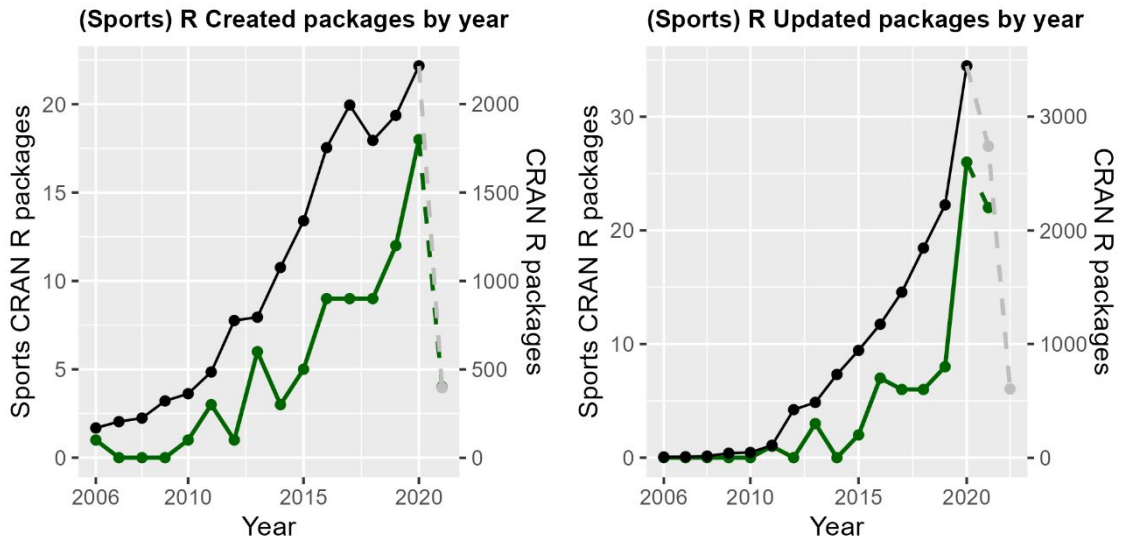


Figure 4: Cumulative number of downloads for the six most successful packages since 2013. At the right, Lahman was the most downloaded package followed by the represented packages at the left. Vertical axes differ in the two panels.

542x187mm (118 x 118 DPI)

Supplemental material

Figure 5: Sports R created and updated packages by year



Peer Review