International Journal of Sports Physiology and Performance, 2023, 18, 1213-1218
https://doi.org/10.1123/ijspp.2023-0007

Human Kinetics
BRIEF REPORT

Synthetic Data as a Strategy to Resolve Data Privacy and Confidentiality Concerns in the Sport Sciences: Practical Examples and an *R Shiny* Application

Mitchell Naughton,[1] Dan Weaving,[1,2] Tannath Scott,[2,3] and Heidi Compton[1]

[1]Applied Sports Science and Exercise Testing Laboratory, University of Newcastle, Ourimbah, NSW, Australia;
[2]Carnegie Applied Rugby Research Centre, Leeds Beckett University, Leeds, United Kingdom;
[3]School of Health Sciences and Social Work, Griffith University, Gold Coast, QLD, Australia

Purpose: There has been a proliferation in technologies in the sport performance environment that collect increasingly larger quantities of athlete data. These data have the potential to be personal, sensitive, and revealing and raise privacy and confidentiality concerns. A solution may be the use of synthetic data, which mimic the properties of the original data. The aim of this study was to provide examples of synthetic data generation to demonstrate its practical use and to deploy a freely available web-based *R Shiny* application to generate synthetic data. ***Methods:*** Openly available data from 2 previously published studies were obtained, representing typical data sets of (1) field- and gym-based team-sport external and internal load during a preseason period (n = 28) and (2) performance and subjective changes from before to after the posttraining intervention (n = 22). Synthetic data were generated using the *synthpop* package in R Studio software, and comparisons between the original and synthetic data sets were made through Welch *t* tests and the distributional similarity standardized propensity mean squared error statistic. ***Results:*** There were no significant differences between the original and more synthetic data sets across all variables examined in both data sets ($P > .05$). Further, there was distributional similarity (ie, low standardized propensity mean squared error) between the original observed and synthetic data sets. ***Conclusions:*** These findings highlight the potential use of synthetic data as a practical solution to privacy and confidentiality issues. Synthetic data can unlock previously inaccessible data sets for exploratory analysis and facilitate multiteam or multicenter collaborations. Interested sport scientists, practitioners, and researchers should consider utilizing the *shiny* web application (SYNTHETIC DATA—available at https://assetlab.shinyapps.io/SyntheticData/).

Keywords: sport performance, technology, hypothesis generation, data analysis, simulation

There has been a proliferation in technologies in sport that gather large quantities of information regarding athlete and team performance. Examples of these include global positioning systems, global navigation satellite systems (GNSS), force plate technologies, and inertial measurement units, among others.[1,2] Indeed, a recent Australian Academy of Science report investigating data governance in sport noted that there has been an "... explosion in the amount of data being generated and in the number of parties who have taken an interest ..." in sporting data.[3] Such data are of interest to support staff (eg, sports scientists, strength and conditioning coaches) as it can inform knowledge of player physical fitness, match and training player movement tracking, and fatigue/recovery.[4] Other entities have also begun to seek access to performance data including third parties (eg, sports betting companies, sports technology vendors, and broadcasters) who view such data as a potentially monetizable asset. Moreover, data that were traditionally collected and stored locally within the team environment is now being collected and stored electronically on systems which often exist in (or are backed up to) servers in remote locations or on systems owned by device manufacturers. Collectively, these developments raise potential privacy and confidentiality concerns, as the types of data that are routinely collected could be identifiable and the data have the potential to be personal, sensitive, and intimately revealing if disclosed.[3]

A potential solution to some of these issues in sport is the use of synthetic data,[5] originally proposed by Rubin,[6] with the goal of producing data which could be used for inference in place of real data. Synthetic data sets contain simulated data, which replace some or all the original observed values, with values which are sampled from the underlying distribution of the data.[7] This ensures that the essential features (eg, distributional shape, skewness, missing data, etc) of the original data set are replicated.[5,8] As the individual data points which could be de-identifiable in the original data set are replaced, privacy and confidentiality concerns are alleviated, with the synthetic data set able to then be shared freely. Research in the biobehavioral sciences has indicated that there is a high degree of similarity between original and synthetic data sets, which is maintained with data sets of varying size, skewness, and quantity of missing data.[8] As these properties are maintained, relationships (eg, predictor and response) between variables, and

Weaving https://orcid.org/0000-0002-4348-9681
Scott https://orcid.org/0000-0003-4336-2370
Compton https://orcid.org/0000-0002-5818-4450
Naughton (mitch.naughton@newcastle.edu.au) is corresponding author, https://orcid.org/0000-0002-3389-6275

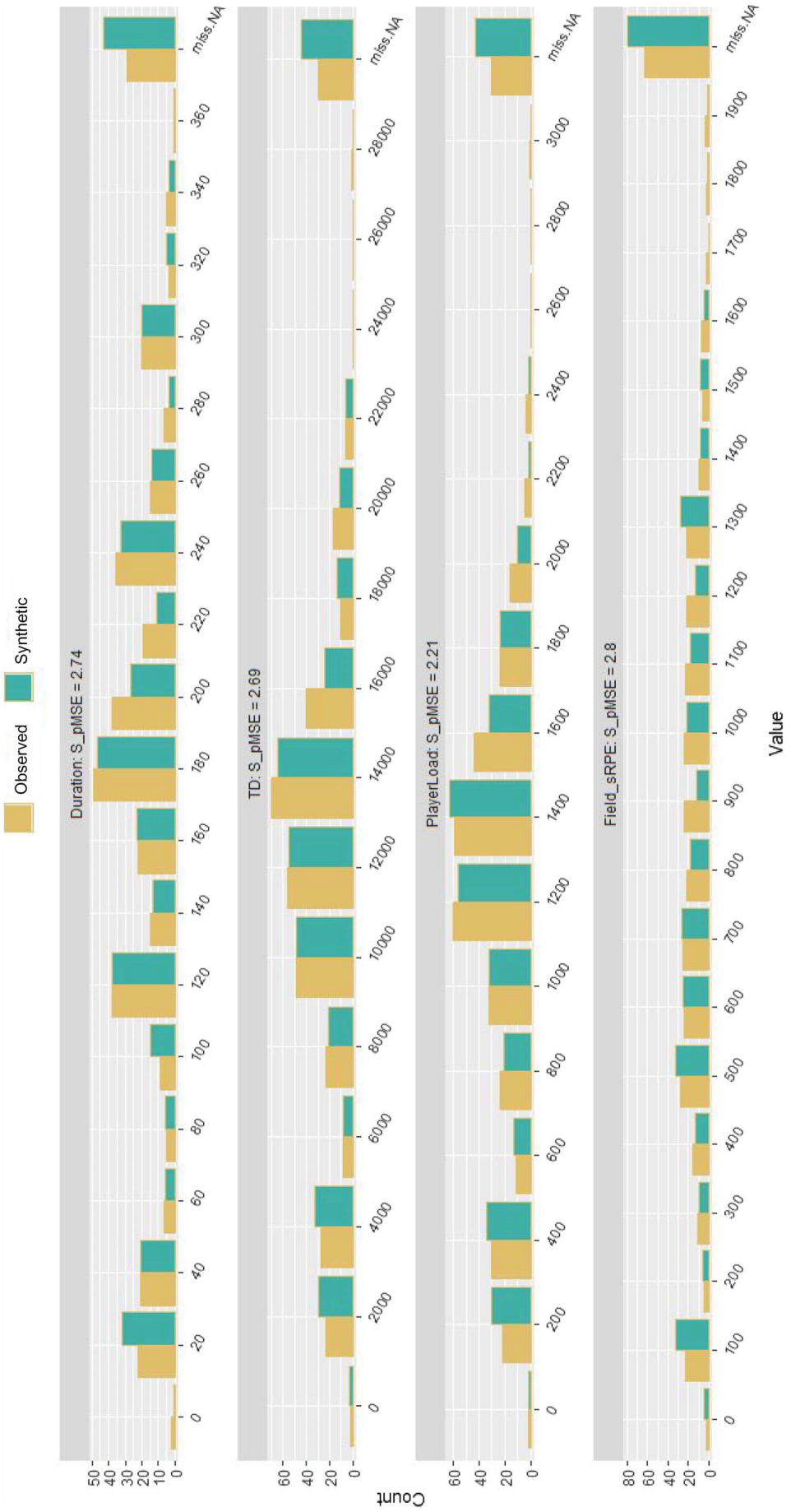

Figure 1 — Comparison between the original observed (yellow bars, left-hand bar in the pair for each value) and synthetic (blue bars, right-hand bar in the pair for each value) data sets across a range of weekly summated field-training variables including duration (in minutes), TD (in meters), PlayerLoad (AU), and sRPE (AU). The column "miss.NA" includes counts of missing data from both observed and synthetic data sets. The S_pMSE statistic quantifies the distributional similarity with values closer to zero indicated a better fit. S_pMSE indicates standardized propensity mean squared error; sRPE, session rating of perceived exertion; TD, total distance; AU, arbitrary units.

inferences drawn from the data are also expected to remain constant.

Therefore, the aim of this study is to provide practical examples of synthetic data generation to demonstrate how it can be used in practice, and to develop and deploy a freely available web-based *R Shiny* application that allows users to upload their data sets and be provided with a synthetic data set which mimics the original data sets properties. The goal of this was to provide practitioners with a freely available resource, which they could use to aid in the sharing of data in a manner which does not violate player privacy and confidentiality.

Methods

Design

This study involved 2 previously published open data sets, which were used for synthetic data generation (see "Subjects" section for details).

Subjects

Data set 1 included male academy rugby league players (n = 28) who were tracked throughout their preseason with data collected on weekly gym-based resistance training volume, field-based external load via GNSS (Catapult OptimEye S5, Catapult Innovations), and gym- and field-based internal load by session rating of perceived exertion.[2] See Moore et al[2] for further details on the methods of data collection.

Data set 2 included indicators of fatigue and recovery in male and female team-sport players (n = 22) before (Pre), immediately after (Post 1), and 72 hours after (Post 2) a 6-day high-intensity interval training program.[9] See Wiewelhove et al[9] for further details on the methods of data collection.

Statistical Analysis

All statistical analysis was performed using R Studio (version 4.1.1, R Core Development Team). Packages used included *synthpop*,[10] *tidyverse*,[11] and *shiny* (see "*R Shiny* Application Development" section below). All synthetic data were produced using the *synthpop* package, using the Classification and Regression Tree approach.[10] The advantage of using Classification and Regression Tree is that it requires no assumptions regarding the distribution of values and can therefore handle data that are highly skewed and multimodal, while also performing well on data sets with missing data.[12]

Synthetic data sets were compared with the original observed data sets through general utility and specific utility. General utility (ie, the distributional similarity) is measured by the standardized propensity mean squared error (S_pMSE) statistic, with values closer to zero indicating a better utility (eg, fit). Specific utility (eg, statistical comparison[s]) is compared through Welch *t* tests of original and synthetic data sets.[7,13] Statistical significance was set at $P < .05$.

R Shiny Application Development

The app (SYNTHETIC DATA) was developed with the *shiny* package for R Studio (version 2021.09.1) and runs online inside an internet browser. The app, instructions on how to use the app, and structure the data for use, as well as other documentation can be found on the landing page (https://assetlab.shinyapps.io/SyntheticData/). The app source code is permanently available in the repository (https://github.com/heidithornton/Sport-science-synthetic-data) and is released under a GNU General Public License version 3 (https://www.r-project.org/Licenses/GPL-3), which ensures that the app is free; open source; and can be used, modified, or contributed to. Finally, to remove the risk that synthetic data are accidently viewed as real data, each individual column in the synthetic data set output is amended with notation indicating that it is synthetic and not original data ("SYNTH_"), as recommended by Nowok et al.[7]

Results

Data Set 1—Moore et al[2]

Comparison between the original observed and synthetic data set for counts of weekly summated field session variables is shown in Figure 1. The S_pMSE values were close to 0 (Figure 1), indicating better utility (eg, fit) between observed and synthetic data sets.

The original observed and synthetic data sets were compared across a range of gym-based training variables and are shown in Table 1. There were no statistically significant differences ($P > .05$) between observed and synthetic data sets across these variables. This figure represents the typical output available from the comparison plots in the SYNTHETIC DATA *R Shiny* application.

Data Set 2—Wiewelhove et al[9]

Comparison between the original and synthetic data sets for changes in subjective soreness and countermovement jump performance is shown (Figure 2). The S_pMSE values ranged between 0.90 and 3.2.

Table 1 Comparison Between the Original Observed and Synthetic Data Sets Across a Range of Weekly Gym Training Variables Including Average Session Duration, Upper-Body Weight Lifted, Lower-Body Weight Lifted, and Rating of Perceived Exertion for Each Session

	Descriptive data, mean (SD)		Statistical comparison (Welch *t* test)		
Variable	Observed	Synthetic	Mean difference (95% CI)	*T* statistic	*P*
Session duration (min)	159 (53)	155 (54)	4.0 (–4.1 to 12.7)	1.01	.312
Upper-body weight lifted (kg)	383 (156)	370 (145)	13.0 (–12.4 to 38.4)	1.01	.315
Lower-body weight lifted (kg)	485 (265)	483 (267)	2.0 (–41.2 to 45.0)	0.09	.932
Rating of perceived exertion (AU)	13.4 (5.1)	12.8 (5.3)	0.6 (–0.2 to 1.4)	1.38	.168

Abbreviation: AU, arbitrary units.

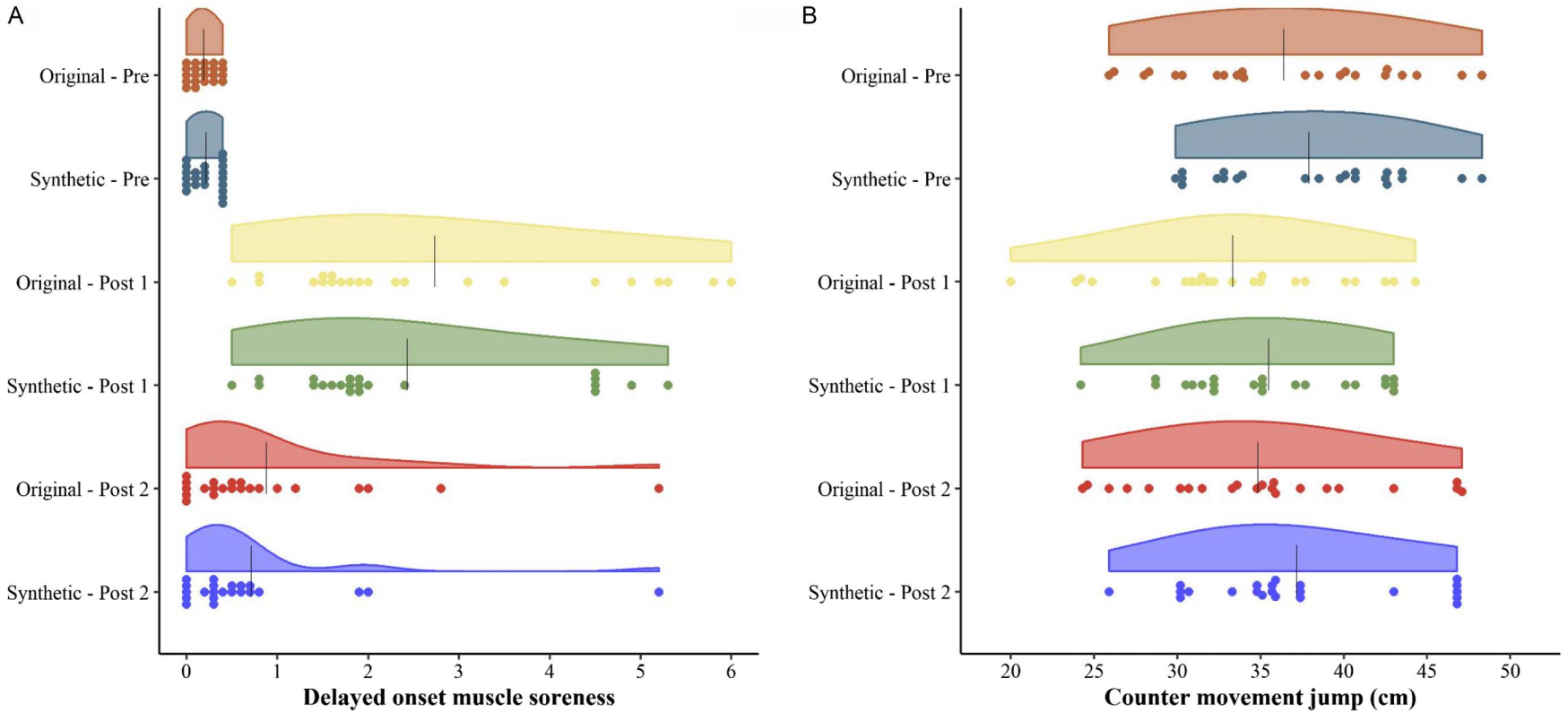

Figure 2 — Rain-cloud plot comparison between the original observed and synthetic data sets for (A) delayed-onset muscle soreness and (B) countermovement jump height before (Pre), immediately after (Post 1), and 72 hours after (Post 2) the completion of a 6-day training intervention. Each dot represents an individual data point, and the single vertical line indicates the mean.

For specific utility, comparison was made with the original data sets across different running and jumping performance measures (Table 2). There were no statistically significant differences between observed and synthetic data sets for any of the studied variables ($P > .05$).

The synthetic data set output for *Study 1* and *Study 2* are provided in the Supplementary Material (available online). The SYNTHETIC DATA *R Shiny* application is available at https://assetlab.shinyapps.io/SyntheticData/.

Discussion

The aim of this study was to provide practical examples of synthetic data generation, which demonstrate how it can be used in practice, and to develop and deploy a freely available web-based SYNTHETIC DATA *shiny* application, which allows users to undertake the process of synthetic data generation. With the examples provided in this study, synthetic data can be used in the sports sciences in the place of original observed data in several different use cases. Sports scientists, practitioners, and researchers should consider the application of synthetic data, and the SYNTHETIC DATA *R Shiny* web application (available at https://assetlab.shinyapps.io/SyntheticData/), when the sharing of original observed data sets is inappropriate.

There are several potential applications for synthetic data sets and the associated SYNTHETIC DATA *R Shiny* app, which has been developed specifically for this study. First, it allows practitioners who wish to collaborate with researchers to explore relationships in their data can do so while maintaining the confidentiality of their athletes/participants. In this way, synthetic data might unlock previously inaccessible data sets for exploratory analysis and hypothesis generation,[5] and facilitate larger scale multiteam or multicenter collaborations.[14] Further, it may be a solution for organizations retaining historical data, whereby anonymizing and synthesizing data may alleviate concerns of retaining identifiable data, while still providing use for further analysis (eg, positional match demand trends). Second, it could allow for the cataloging of freely available data sets, which can be used for teaching and educational purposes. This could assist students, who would otherwise not have access, in developing their data cleaning, analysis, and visualization skills on data sets, which are collected in real performance environments with the associated considerations (eg, missing data, variability).[4] A collection of different synthetic data sets from various sports and common use cases would be a foreseeable outcome for this application. Finally, by developing and deploying the synthetic data production in a freely available web *R Shiny* application, we have removed the necessity for practitioners to have the necessary expertise to produce synthetic data by developing their own scripts.

There are limitations which may be associated with the use of synthetic data. First, as the original data points are replaced, the ability to condition the analysis on individual player responses in more complex models (eg, mixed-models, repeated-measures correlations) may be restricted. This analysis may be possible with the additional functionality of the *synthpop* package, which compares the overlap and lack-of-fit of the z values between original and observed statistical models.[10] In practice, this means that, at present, synthetic methods are limited to inference based on linear models.[8] Examining the merit of synthetic data in more complex analyses (eg, mixed-effects models, principal component analysis, or statistical parametric mapping) is necessary to examine its suitability in these situations. Second, for cases when there are variables with few observations or with extreme outliers, users should be aware that these may increase the disclosure risk, even in the synthetic data sets. An example of this may be a data set that includes a world-record holder or world-leading performance, which would be identifiable in a synthetic data set if the data set containing the original performance was able to be linked. In these disclosure risk instances, the identification risk profile of the synthetic data set can be calculated.[15] Further, there remains uncertainty on data ownership following the creation of a synthetic data set, as there are no conclusive ownership rights of athlete data under the legislations such as Australian Privacy Act 1988 (Cth) and the European Union's General Data Protection Regulation 2016 despite the data not being "reasonably identifiable." Finally, the use of synthetic data may not diminish the risk of third parties accessing the original data sets should they have system access, which is not limited by the legal frameworks surrounding data ownership/custody in a given jurisdiction. Collectively, these limitations need to be considered in the context of using synthetic data and the associated SYNTHETIC DATA *R Shiny* application.

Practical Applications

This report provides practical examples of approaches to synthetic data generation, which mimic the original data sets, in both their distributional and statistical properties. A freely available web application has been developed and deployed, which will allow

Table 2 Comparison Between Original Observed and Synthetic Data Sets for 20-m Sprint Time and Repeated-Sprint Ability Across the 3 Different Testing Time Points (Pre, Post 1, and Post 2)

		Descriptive data, mean (SD)		Statistical comparison, Welch *t* test		
Variable	Time	Observed	Synthetic	Mean difference (95% CI)	*T* statistic	*P*
20-m sprint (s)	Pre	3.28 (0.24)	3.30 (0.32)	–0.02 (–0.19 to 0.15)	–0.22	.830
	Post 1	3.40 (0.31)	3.36 (0.25)	0.04 (–0.13 to 0.20)	0.40	.689
	Post 2	3.35 (0.24)	3.31 (0.23)	0.04 (–0.10 to 0.18)	0.59	.560
Repeated-sprint ability (s)	Pre	5.02 (0.52)	5.11 (0.45)	–0.09 (–0.39 to 0.21)	–0.62	.540
	Post 1	4.84 (0.56)	4.96 (0.55)	–0.12 (–0.46 to 0.21)	–0.73	.470
	Post 2	4.97 (0.56)	5.07 (0.53)	–0.10 (–0.43 to 0.23)	–0.59	.561

Note: Pre, before the training intervention; Post 1, immediately after the end of the 6-day training intervention. Post 2, 72 hours after the end of the 6-day training intervention.

users to upload a data set and receive a corresponding synthetic data set, which they can then freely share. This approach allows practitioners or users without the necessary coding expertise to be able to produce their own synthetic data sets.

Conclusions

Synthetic data generation can be considered a potential solution to the data privacy and confidentiality concerns that have been identified in accessing and sharing data in sport. The examples described here, and the freely available web application, allow practitioners to produce and share data, which should promote greater collaboration with researchers, exploratory analysis, and hypothesis generation.

References

1. van den Tillaar R, Nagahara R, Gleadhill S, Jiménez-Reyes P. Step-to-step kinematic validation between an inertial measurement unit (IMU) 3D system, a combined laser+IMU system and force plates during a 50 m sprint in a cohort of sprinters. *Sensors*. 2021;21(19):6560. doi:10.3390/s21196560
2. Moore DA, Jones B, Weakley J, Whitehead S, Till K. The field and resistance training loads of academy rugby league players during a pre-season: comparisons across playing positions. *PLoS One*. 2022;17(8):e0272817. doi:10.1371/journal.pone.0272817
3. Australian Academy of Science. *Getting Ahead of the Game*. 2022.
4. Thornton HR, Delaney JA, Duthie GM, Dascombe BJ. Developing athlete monitoring systems in team sports: data analysis and visualization. *Int J Sports Physiol Perform*. 2019;14(6):698–705. doi:10.1123/ijspp.2018-0169
5. Warmenhoven J, Harrison A, Quintana D, Hooker G, Gunning E, Bargary N. Unlocking sports medicine research data while maintaining participant privacy via synthetic datasets. *SportRxiv*. August 20, 2020. doi:10.31236/osf.io/f3rz7
6. Rubin DB. Statistical disclosure limitation. *J Off Stat*. 1993;9(2):461–468.
7. Nowok B, Raab GM, Dibben C. Providing bespoke synthetic data for the UK Longitudinal Studies and other sensitive data with the synthpop package for R. *Stat J IAOS*. 2017;33(3):785–796. doi:10.3233/SJI-150153
8. Quintana DS. A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation. *Elife*. 2020;9:275. doi:10.7554/eLife.53275
9. Wiewelhove T, Raeder C, Meyer T, Kellmann M, Pfeiffer M, Ferrauti A. Markers for routine assessment of fatigue and recovery in male and female team sport athletes during high-intensity interval training. *PLoS One*. 2015;10(10):e0139801. doi:10.1371/journal.pone.0139801
10. Nowok B, Raab GM, Dibben C. Synthpop: bespoke creation of synthetic data in R. *J Stat Softw*. 2016;74(11):1–26. doi:10.18637/jss.v074.i11
11. Wickham H, Averick M, Bryan J, et al. Welcome to the Tidyverse. *J Open Source Softw*. 2019;4(43):1686. doi:10.21105/joss.01686
12. Lewis RJ. *An Introduction to Classification and Regression Tree (CART) Analysis*. Citeseer; 2000.
13. Delacre M, Lakens D, Leys C. Why psychologists should by default use Welch's t-test instead of Student's t-test. *Int Rev Soc Psychol*. 2017;30(1):82. doi:10.5334/irsp.82
14. Slattery K, Crowcroft S, Coutts AJ. Innovating together: Collaborating to impact performance. *Int J Sports Physiol Perform*. 2021;16(10):1383–1384. doi:10.1123/ijspp.2021-0389
15. Hornby R, Hu J. Identification risks evaluation of partially synthetic data with the identificationriskcalculation R package. *arXiv*. 2020;14(1):37–52.