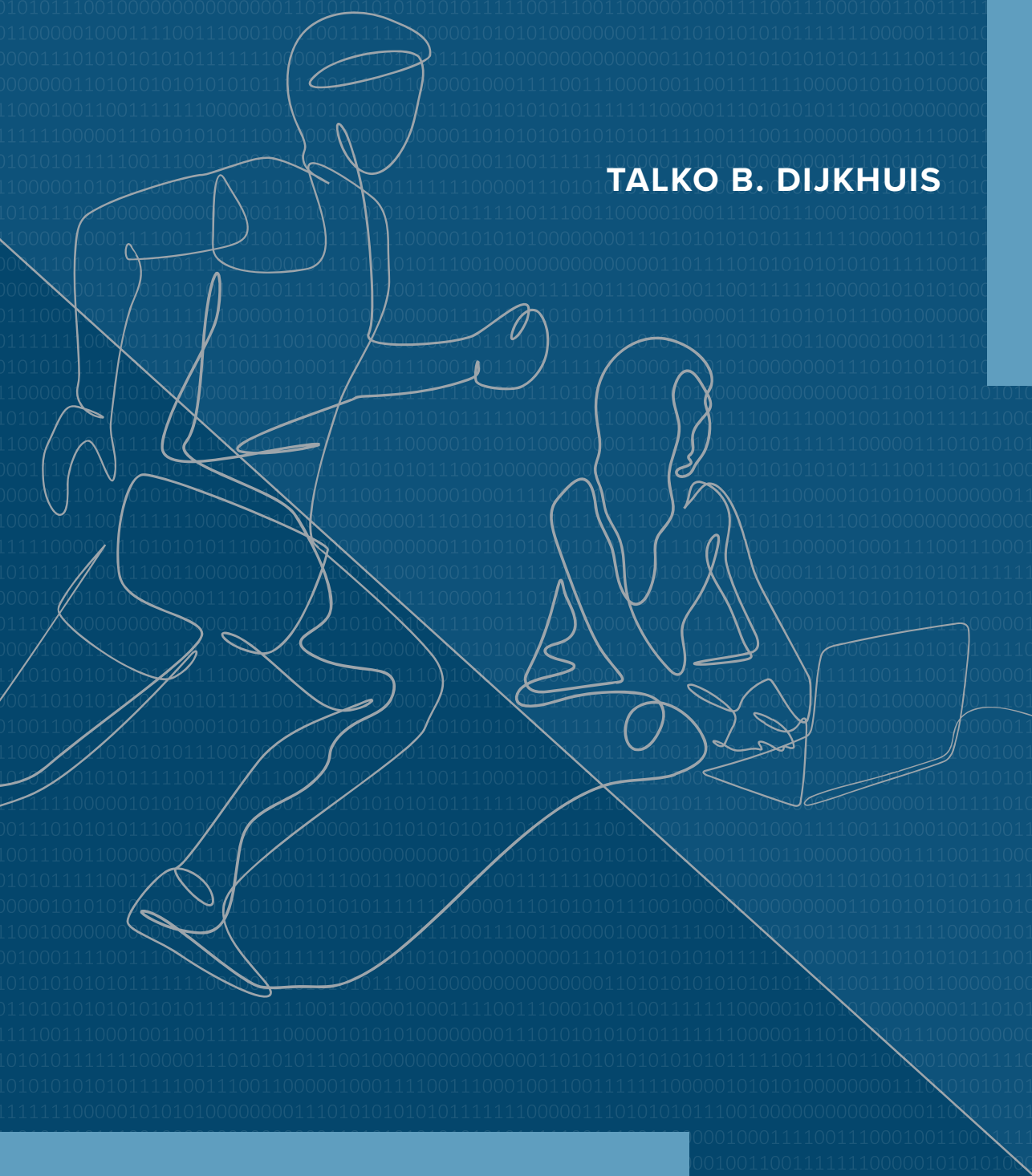
PHYSICAL PERFORMANCE IN DAILY LIFE AND SPORTS:

BRIDGING THE DATA ANALYTICS GAP

TALKO B. DIJKHUIS



Physical performance in daily life and sports: bridging the data analytics gap

Talko B. Dijkhuis

Cover: Ilse Modder | www.ilsemodder.nl
Layout: Ilse Modder | www.ilsemodder.nl
Printed by: Ipskamp Printing | www.ipskampprinting.nl

university of groningen

Physical performance in daily life and sports: bridging the data analytics gap

PhD thesis

to obtain the degree of PhD at the
University of Groningen
on the authority of the
Rector Magnificus Prof. J.M.A. Scherpen
and in accordance with
the decision by the College of Deans.

This thesis will be defended in public on

Wednesday 7 February 2024 at 16.15 hours

by

Talko Bernhard Dijkhuis

born on 28 February 1969

Supervisors
Prof. K.A.P.M. Lemmink
Prof. M. Aiello

Co-supervisor
Dr. H. Velthuijsen

Assessment Committee
Prof. M. Biehl
Prof. J.E.W.C. van Gemert-Pijnen
Prof. J.N. Kok

Paranymphs
Dr. Frank J. Blaauw
Dr. Wico Mulder

TABLE OF CONTENTS

CHAPTER 1

General introduction

BACKGROUND

Running a marathon at the Olympic level represents the ultimate challenge for an elite athlete, whereas walking independently to the supermarket may be an outstanding achievement for a frail older person. While both engage in physical activity, 'any bodily movement produced by skeletal muscle that requires energy expenditure' [1], the difference in the level of physical performance between an Olympic athlete and a frail older person illustrates that valuing physical performance depends on an individual's physical capacities and context [2], [3]. Zehr categorises the physical performance of humans as a continuum from the lowest category, humans being hindered from performing basic activities due to illness or injury, through a middle category of humans being able to perform in an everyday day setting to the highest category of humans that can perform at an athletic level [4]. In this thesis, we concentrate our research on humans functioning in daily life and elite sports.

In everyday life, being physically active helps to maintain general functioning as long as it is done regularly and with sufficient duration and intensity [5]. Physical activity in daily life can be undertaken in many ways, such as walking, cycling, gardening, household activities, sports, and work. Regular physical activity generally prevents chronic or non-communicable diseases such as heart disease, stroke, and diabetes and supports mental health, quality of life, and well-being [5]. Inversely, insufficient physical activity may lead to deterioration of physical performance in daily life and higher risks of health problems, illness, and lower life expectancy [6]–[8]. Consequently, it is essential to monitor physical activity and to intervene to enhance physical activity when the physical activity is insufficient to maintain general functioning.

In elite sports, for example, high-level running or professional soccer, humans require many years of targeted physical activity (i.e., training). During these years of training, a delicate balance between training (referred to as 'load'), recovery, and individual capacity must be maintained to increase physical performance and prevent injuries and illnesses [9]. A short-term increase in training frequency, duration, or intensity provokes a short-term decrement in performance, termed functional overreaching (FOR) [10]. FOR is considered a necessary component of a training programme to improve physical performance [11]–[13]. On the other hand, high training loads without sufficient recovery over a longer period may lead to overtraining, causing a decrease in physical performance and a higher risk of sustaining an injury [14]–[19]. Therefore, balancing load, recovery, and capacity is crucial for elite athletes to maintain or increase physical performance and prevent injuries. Hence, monitoring the balance and adjust the balance is essential for elite athletes to maintain and increase physical performance.

The frequency, duration, and intensity of physical activity in daily life or elite sports can be expressed in physical performance measures, such as the number of steps in a day, the covered distance at different running intensities in a soccer match, or the perceived exertion of a training session in elite runners. An intriguing question is whether physical activity or performance can be predicted based on earlier physical activity or physical performance, such as predicting the total number of steps on a day based on the number of steps in the morning or predicting the covered distance at the end of a soccer match based on the covered distance at the beginning of a match. Also, the potential of physical activity or physical performance measures to predict other outcomes, such as the risk of sustaining injuries [14]–[19] or the rate of improvement in the performance of running a marathon [20], is of interest. The combination of monitoring physical activity and physical performance, along with predictive analytics of physical activity, performance, and injuries, presents an opportunity to intervene in physical activity patterns throughout the day, during training sessions, races, or matches. This intervention can help optimize physical performance and reduce the risk of injuries.

MONITORING AND ANALYTICS

In the last decade, options for measuring physical activity and physical performance have dramatically increased [21], [22]. This increase is mainly based on technological developments (but not limited to), such as wearable sensor devices, automatic tracking systems, video-based motion analysis, and Global Positioning Systems (GPS) [21], [22]. These systems increase the ability to provide insight into physical activity patterns in daily life and sports [23]. For example, wearable sensor devices, such as the activity trackers of Fitbit or Garmin, can monitor physical activity in daily life. The global shipment volume of wearable sensor devices was 266.3 million in 2020 and is expected to reach 776.23 million units by 2026 [24]. In the last decade, researchers have taken advantage of Fitbit's public appeal, prominence, and relatively low cost by incorporating these devices into their studies on physical activity [25].

In sports, computer-aided tracking technology has developed substantially to monitor athletes' physical activity and physical performance during training and match play. Monitoring systems are used in all kinds of team sports, such as soccer [26], Australian football [27], basketball [28], hockey [29], or individual sports like speed skating [30] and running [31]. Monitoring systems using tracking technology have evolved considerably. For example, Van Gool et al. were the first to track a soccer match in the eighties filming at 5hz and processing it afterwards [32]. Nowadays, technology can quickly record and process the data of all athletes' physical activity throughout an entire match or training session [33]. These monitoring systems have become commonplace in professional

sports [34], enabling automatic analysis of physical activity during races or match-play [35].

Wearable sensor devices and monitoring systems provide an immense amount of physical activity and physical performance data [27]–[29], which present opportunities to develop knowledge in scientific areas like behavioural science, human movement, and sports science [23], [36] and to translate this knowledge to daily practice. For example, this data can be utilised in lifestyle interventions, rehabilitation programmes, or training programmes for elite athletes. In addition, training logs kept by athletes and coaches provide a wealth of data on physical activity, physical performance, psychological well-being, injury, and recovery.

Data analytics, a field that encompasses techniques for working with data, such as machine learning and advanced statistics [37], [38], enables the extraction of insights and predictions from the collected data. However, the use of data analytics in behavioural, human movement and sport sciences is limited so far [23], [35]. Also, the application of artificial intelligence (AI) and machine learning (ML) based on wearable data and monitoring systems data in sports is still in its preliminary stage [38][39].

Although data analytics offers opportunities, there are various problems in extracting actionable and meaningful insights and predictions based on physical activity and physical performance data. Four problems can be identified in data analytics of physical activity and physical performance data, creating a **data analytics gap**. To address these problems, we propose potential solutions for each of the identified problems.

The **first problem** is the limited **use of individualised prediction based on personalised data** [39]–[41], limiting the provision of meaningful insights and predictions for individuals, athletes and coaches. In order to provide individualised insights and predictions, a boundary condition is that the data contains sufficient personal information. **A potential solution to the first problem** to enable performing data analytics at the individual level using personalised data **is to use datasets containing personalised data** from wearable sensor devices such as Fitbit or Garmin or optical tracking systems, such as SportsVU, monitoring each individual during their daily life, training or match play, or individual test and exercise log data collected by smart apps such as Sports Tracker or Runkeeper. The **second problem** is **the vast amount and complexity of data**. Several authors acknowledged the complexity of data analysis and data analytics in the past decade, given the vast amount of data. For example, Silver's book on prediction, published in 2012, indicated that realising a correct prediction model, among others in basketball, is challenging because the amount of meaningful information relative to the increasing overall amount of data is declining [41]. In 2014, Davenport concluded that the amount

of data in sports is considerably more than the ability to extract meaningful insights from it [39]. Six years later, in 2020, as highlighted in a survey on the use of wearables producing GPS data in English professional soccer (2020) [40], it was posed that there is still a struggle to distil predictions due to the amount and complexity of the data and the outcomes are often inconclusive [40]. **A potential solution to the second problem is using more sensitive performance measures and applying various machine learning algorithms**. The more sensitive a physical performance measure the more responsive or reactive it is to changes in the underlying system being measured [42]. The fundamental consideration in machine learning is not only which algorithm is superior but, rather, the conditions (e.g. physical performance measures used) under which an algorithm can outperform others [43]. Therefore, to determine these physical performance measures' relative applicability and sensitivity in prediction models, we examine the effectiveness of converting raw physical activity data from daily life and sports into various physical performance measures. We combine these physical performance measures with multiple machine learning algorithms to identify the optimal combination enabling meaningful predictions. The **third problem** is **the use of simplified models and assumptions.** When using data from wearables and monitoring systems applying models to derive meaningful insights is complex. The complexity arises from the fact that extracting insights using traditional statistical models often entails unrealistic assumptions on the underlying reality and often reduces complex questions to simple statistical assumptions using a fixed set of parameters [41], [44]–[47]. For example, the often-used Generalised Linear Models depend on the linearity and completeness of the parameters. However, this is seldom the case and yields suboptimal results [48]. **A potential solution to the third problem is to apply a causal roadmap in combination with a causal model.** The causal roadmap is a framework for designing and evaluating causal inference studies [49]. It provides a structured approach for identifying potential sources of bias and confounding in observational studies and selecting appropriate methods for adjusting for these sources of bias. The causal roadmap demands a well-defined causal model of reality. A causal model, as proposed by Judea Pearl in the Book of Why [50], is a mathematical framework for understanding the relationships between variables in a system and how changes in one variable can cause changes in another while identifying the potential sources of bias and confounding. Pearl's causal model allows for the analysis and manipulation of complex systems. **A fourth problem** is **the absence of confounding variables.** For example, contextual variables in soccer, such as match location (home or away), score (win, draw or lose), and rival level, and individual characteristics [51], such as the athlete's psychological well-being or social support, might influence physical performance during a match. Although monitoring systems and log data may include some of these variables, the availability and integration of all relevant variables in sports is a specific problem for the prediction of physical performance or injury risk [52]. The inability to include all relevant variables in sports analytics, is

known as the endogeneity problem [53]. **A potential solution to the fourth problem is using statistical methods that account for the absence of confounding variables.** One such method involves utilising an ensemble of machine learning techniques in conjunction with Targeted Maximum Likelihood Estimation (TMLE). The goal of this strategy is to minimise the impact of the missing variables by using TMLE, which is known to be more robust to inaccuracies in modelling the underlying reality compared to traditional statistical methods [44], [45], [54]. **As an alternative potential solution to the fourth problem, we take a two-way approach by applying traditional statistical analysis and machine learning techniques to the same incomplete data set**. The use of traditional statistical analysis on the one hand and machine learning on the other provides insight into their relative performance when dealing with incomplete data [55]. This approach allows us to understand the applicability and discuss the strengths of traditional statistical analysis and machine learning.

AIM AND OUTLINE

This thesis aimed to reduce the data analytics gap represented by the four identified problems while examining the associated potential solutions to enable meaningful insights and predictions related to physical activity and physical performance. These insights and predictions can be used to make more informed decisions regarding physical activity and physical performance interventions.

In Figure 1, we present an outline of the thesis and visualise how each chapter is mapped to one or more problems and solutions to reduce the identified data analytics gap.

The structure of this thesis is as follows:
In Chapter 2, we investigated the possibility of predicting the daily physical performance of employees based on wearable data. The study involved coaching Hanze University of Applied Science employees to increase their physical activity during the daytime and monitor their steps using Fitbits. These steps were subsequently transformed into physical performance measures such as number of steps per hour and total number of steps until a particular hour. In addition, we applied machine learning to predict whether an employee would achieve his or her overall daily step goal during the working day. Through automated analysis of physical activity and physical performance, timely detection of anomalies in behaviour and identifying effective coaching strategies may become feasible.

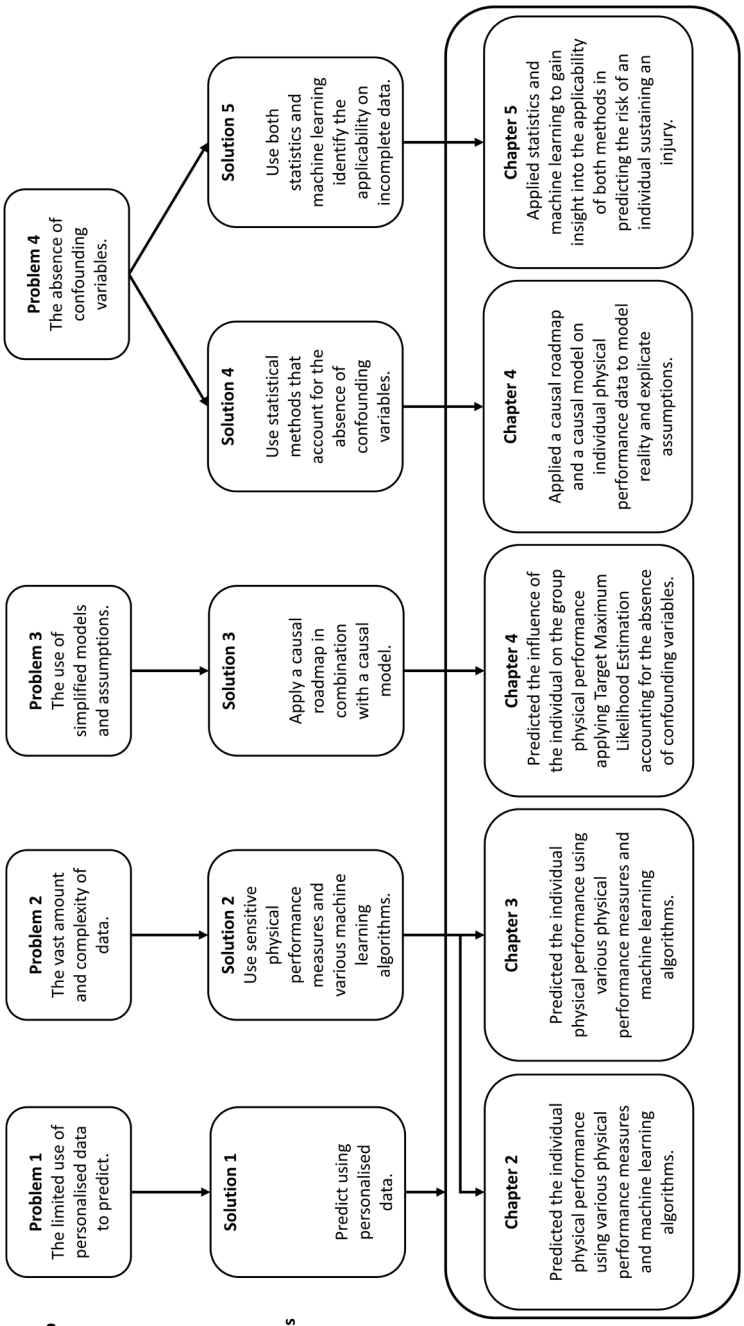


Figure 1 Mapping of the problems creating a data analytics gap, the proposed potential solutions, and the outline of the thesis.

In Chapter 3, we studied the predictability of physical performance in elite soccer matches using various physical performance measures and machine learning techniques. Match data was collected from 302 matches in elite soccer throughout one season. Semi-automatic multiple-camera video technology, the SportsVU optical tracking system, recorded each player's position over time. The individual positions in time translated into three increasingly sensitive physical performance measures, i.e. distance covered, distance in speed category, and energy expenditure in power category. These physical performance measures were used in different machine learning models to identify and predict the physical performance of individual players throughout an elite soccer match.

In Chapter 4, we investigated the effect of substitution in soccer on a team's physical performance and the suitability of following the causal roadmap and a causal model in this context. A causal model of the relationships between substitution variables and the soccer team's physical performance was created as an essential step in following the causal roadmap. The causal model included variables such as the number of substitutions made by a team, the moment of the substitution and the soccer team's total distance covered. We used the causal model to identify the assumptions needed to infer causality from the data and the potential sources of bias and confounding that may affect the causal effect estimates.

We also provided an in-depth analysis of statistical methods. We evaluated the accuracy of estimating the impact of substitutions on a football team's physical performance using synthetically generated position and substitution data and data from the SportsVU optical tracking system. We compared the accuracy of the TMLE and generalized linear model using the complete data set versus the accuracy of these models when a crucial variable was removed from the data set. The difference in accuracy between the two methods indicated whether a more robust statistical method such as TMLE could provide more accurate insight into the effect of substitutions on football team physical performance when a crucial variable is absent, compared to a traditional generalized linear model.

In **Chapter 5, we investigated the relationship between training characteristics and injuries in competitive runners while evaluating the feasibility of using statistical analysis and exploring the potential of machine learning.** The dataset comprised test, training, and injury log data collected from individual competitive runners and their coach. We used the log data on multiple physical load measures, such as training intensity and rate of perceived exertion, to construct the physical performance measure acute:chronic workload ratio. Traditionally running injuries have a complex origin, and datasets lack relevant confounding variables that can provide insight into injury risk

factors [56]. We aimed to assess the viability of using statistical analysis and prediction in injury risk and evaluate the potential of using machine learning to predict injuries using a dataset known for its absence of several relevant variables.

Finally, **Chapter 6** provides a general discussion of the results, an overall conclusion and ends with some concrete ideas on practical applications and possible directions for further investigations.

REFERENCES

[1] World Health Organisation, "Global recommendations on physical activity for health," 2010.

[2] L. Holsbeeke, M. Ketelaar, M. M. Schoemaker, and J. W. Gorter, "Capacity, Capability, and Performance: Different Constructs or Three of a Kind?," *Arch. Phys. Med. Rehabil.*, vol. 90, no. 5, pp. 849–855, 2009, doi: 10.1016/j.apmr.2008.11.015.

[3] C. Dunn, W., & Brown, "The Ecology of Human Framework for Considering the Effect of Context," *Am. J. Occup. Ther.*, vol. 48, pp. 595–607, 1994, doi: 10.5014/ajot.48.7.595.

[4] E. P. Zehr, "The potential transformation of our species by neural enhancement," *J. Mot. Behav.*, vol. 47, no. 1, pp. 73–78, 2015, doi: 10.1080/00222895.2014.916652.

[5] World Health Organisation, *Global action plan on physical activity 2018-2030: more active people for a healthier world.* 2019.

[6] I. Min-Lee *et al.*, "Effect of physical inactivity on major non-communicable diseases worldwide: An analysis of burden of disease and life expectancy," *Lancet*, vol. 380, no. 9838, pp. 219–229, 2012, doi: 10.1016/S0140-6736(12)61031-9.

[7] U. Ekelund *et al.*, "Does physical activity attenuate, or even eliminate, the detrimental association of sitting time with mortality? A harmonised meta-analysis of data from more than 1 million men and women," *Lancet*, vol. 388, no. 10051, pp. 1302–1310, 2016, doi: 10.1016/S0140-6736(16)30370-1.

[8] E. Losina, H. Y. Yang, B. R. Deshpande, J. N. Katz, and J. E. Collins, "Physical activity and unplanned illness-related work absenteeism: Data from an employee wellness program," *PLoS One*, vol. 12, no. 5, pp. 1–8, 2017, doi: 10.1371/journal.pone.0176872.

[9] R. T. A. Otter, "Monitoring endurance athletes," 2016.

[10] P. Bellinger *et al.*, "Muscle fiber typology is associated with the incidence of overreaching in response to overload training," *J. Appl. Physiol.*, vol. 129, no. 4, pp. 823–836, 2020, doi: 10.1152/japplphysiol.00314.2020.

[11] S. L. Halson, "Monitoring Training Load to Understand Fatigue in Athletes," *Sports Medicine*, vol. 44, no. Suppl 2. Springer, pp. 139–147, Nov. 2014, doi: 10.1007/s40279-014-0253-z.

[12] S. L. Halson and A. E. Jeukendrup, "Does overtraining exist? An analysis of overreaching and overtraining research.," *Sport. Med.*, vol. 34, no. 14, pp. 967–981, 2004, [Online]. Available: http://search.ebscohost.com/login.aspx?direct=true&db=ccm&AN=106594041&site=ehost-live.

[13] R. Meeusen *et al.*, "Prevention, diagnosis, and treatment of the overtraining syndrome: Joint consensus statement of the european college of sport science and the American College of Sports Medicine," *Med. Sci. Sports Exerc.*, vol. 45, no. 1, pp. 186–205, 2013, doi: 10.1249/MSS.0b013e318279a10a.

[14] A. Esmaeili, W. G. Hopkins, A. M. Stewart, G. P. Elias, B. H. Lazarus, and R. J. Aughey, "The individual and combined effects of multiple factors on the risk of soft tissue non-contact injuries in elite team sport athletes," *Front. Physiol.*, vol. 9, no. SEP, pp. 1–16, 2018, doi: 10.3389/fphys.2018.01280.

[15] N. B. Murray, T. J. Gabbett, A. D. Townshend, B. T. Hulin, and C. P. Mclellan, "Individual and combined effects of acute and chronic running loads on injury risk in elite Australian footballers," *Scand. J. Med. Sci. Sport.*, no. 2007, pp. 1–9, 2016, doi: 10.1111/sms.12719.

[16] J. D. Ruddy, C. W. Pollard, R. G. Timmins, M. D. Williams, A. J. Shield, and D. A. Opar, "Running exposure

is associated with the risk of hamstring strain injury in elite Australian footballers," *Br. J. Sports Med.*, p. bjsports-2016-096777, 2016, doi: 10.1136/bjsports-2016-096777.

[17] B. Hulin *et al.*, "Spikes in acute workload are associated with increased injury risk in elite cricket fast bowlers," *Artic. Br. J. Sport. Med.*, 2013, doi: 10.1136/bjsports-2013-092524.

[18] B. T. Hulin, T. J. Gabbett, D. W. Lawson, P. Caputi, and J. a Sampson, "The acute:chronic workload ratio predicts injury: high chronic workload may decrease injury risk in elite rugby league players," *Br. J. Sports Med.*, vol. 50, no. 4, pp. 231–236, 2016, doi: 10.1136/bjsports-2015-094817.

[19] A. Jaspers, J. P. Kuyvenhoven, F. Staes, W. G. P. Frencken, W. F. Helsen, and M. S. Brink, "Examination of the external and internal load indicators' association with overuse injuries in professional soccer players," *J. Sci. Med. Sport*, vol. 21, no. 6, pp. 579–585, 2018, doi: 10.1016/j.jsams.2017.10.005.

[20] V. L. Billat, "Current perspectives on performance improvement in the marathon: From universalisation to training optimisation," *New Stud. Athl.*, vol. 20:3, pp. 21–39, 2005.

[21] A. Rossi, L. Pappalardo, P. Cintia, F. M. Iaia, J. Fernàndez, and D. Medina, "Effective injury forecasting in soccer with GPS training data and machine learning," *PLoS One*, vol. 13, no. 7, pp. 1–15, 2018, doi: 10.1371/journal.pone.0201264.

[22] M. Herold, F. Goes, S. Nopp, P. Bauer, C. Thompson, and T. Meyer, "Machine learning in men's professional football: Current applications and future directions for improving attacking play," *Int. J. Sport. Sci. Coach.*, vol. 14, no. 6, pp. 798–817, 2019, doi: 10.1177/1747954119879350.

[23] B. Caulfield, B. Reginatto, and P. Slevin, "Not all sensors are created equal: a framework for evaluating human performance measurement technologies," *npj Digit. Med.*, vol. 2, no. 1, 2019, doi: 10.1038/s41746-019-0082-4.

[24] "Industry report on smart wearables market." https://www.mordorintelligence.com/industry-reports/smart-wearables-market.

[25] R. G. St Fleur, S. M. St George, R. Leite, M. Kobayashi, Y. Agosto, and D. E. Jake-Schoffman, "Use of fitbit devices in physical activity intervention studies across the life course: Narrative review," *JMIR mHealth uHealth*, vol. 9, no. 5, 2021, doi: 10.2196/23411.

[26] J. Castellano, D. Alvarez-Pastor, and P. S. Bradley, "Evaluation of research using computerised tracking systems (amisco® and prozone®) to analyse physical performance in elite soccer: A systematic review," *Sport. Med.*, vol. 44, no. 5, pp. 701–712, 2014, doi: 10.1007/s40279-014-0144-3.

[27] S. Ryan, T. Kempton, and A. J. Coutts, "Data reduction approaches to athlete monitoring in professional Australian football," *Int. J. Sports Physiol. Perform.*, vol. 16, no. 1, pp. 59–65, 2021, doi: 10.1123/IJSPP.2020-0083.

[28] A. Heishman *et al.*, "Associations Between Two Athlete Monitoring Systems Used to Quantify External Training Loads in Basketball Players," *Sports*, vol. 8, no. 3, p. 33, 2020, doi: 10.3390/sports8030033.

[29] T. Kim, J. H. Cha, and J. C. Park, "Association between in-game performance parameters recorded via global positioning system and sports injuries to the lower extremities in elite female field hockey players," *Cluster Comput.*, vol. 21, no. 1, pp. 1069–1078, 2016, doi: 10.1007/s10586-016-0690-6.

[30] T. Purevsuren, B. Khuyagbaatar, K. Kim, and Y. H. Kim, "Investigation of Knee Joint Forces and Moments during Short-Track Speed Skating Using Wearable Motion Analysis System," *Int. J. Precis. Eng. Manuf.*, vol. 19, no. 7, pp. 1055–1060, 2018, doi: 10.1007/s12541-018-0125-9.

[31] T. Haugen and M. Buchheit, "Sprint Running Performance Monitoring: Methodological and Practical Considerations," *Sport. Med.*, vol. 46, no. 5, pp. 641–656, 2016, doi: 10.1007/s40279-015-0446-0.

[32] D. Van Gool, D. Van Gerven, and J. Boutmans, "The physiological load imposed on soccer players during real match-play.," in *Science and football*, W. J. Reilly, T.; Lees, A.; Davids, K.; Murphy, Ed. London: Spon, 1988, pp. 51–59.

[33] M. Buchheit, A. Allen, T. K. Poon, M. Modonutti, W. Gregson, and V. Di Salvo, "Integrating different tracking systems in football: multiple camera semi-automatic system, local position measurement and GPS technologies," *J. Sports Sci.*, vol. 32, no. 20, pp. 1844–1857, 2014, doi: 10.1080/02640414.2014.942687.

[34] L. Torres-Ronda, E. Beanland, S. Whitehead, A. Sweeting, and J. Clubb, "Tracking Systems in Team Sports: A Narrative Review of Applications of the Data and Sport Specific Analysis," *Sport. Med. - Open*, vol. 8, no. 1, 2022, doi: 10.1186/s40798-022-00408-z.

[35] S. Barris and C. Button, "A review of vision-based motion analysis in sport," *Sport. Med.*, vol. 38, no. 12, pp. 1025–1043, 2008, doi: 10.2165/00007256-200838120-00006.

[36] Statista Technology & Telecommunications, "Connected wearable devices worldwide 2016-2022," *Statista*, 2021. https://www.statista.com/statistics/487291/global-connected-wearable-devices/ (accessed Jun. 30, 2021).

[37] T. A. Runkler, *Data Analytics*, 3rd ed. Wiesbaden: Springer Fachmedien Wiesbaden GmbH, 2020.

[38] W. E. Nagel and T. Ludwig, "Data Analytics," *Informatik-Spektrum*, vol. 42, no. 6, pp. 385–386, 2020.

[39] T. H. Davenport, "Analytics in sports: The new science of winning," *Int. Inst. Anal.*, vol. 2, no. February, pp. 1–28, 2014.

[40] P. Nosek, T. E. Brownlee, B. Drust, and M. Andrew, "Feedback of GPS training data within professional English soccer: a comparison of decision making and perceptions between coaches, players and performance staff," *Sci. Med. Footb.*, vol. 5, no. 1, pp. 35–47, 2021, doi: 10.1080/24733938.2020.1770320.

[41] N. Silver, *The Signal and the Noise: Why So Many Predictions Fail--but Some Don't*. Penguin Press, 2012.

[42] M. Buchheit and B. M. Simpson, "Player-Tracking Technology : Half-Full or Half-Empty Glass ?," *Int. J. Sports Physiol. Perform.*, vol. 12, no. S2, pp. 35–41, 2017.

[43] F. J. Osisanwo, J. E. T. Akinsola, O. Awodele, J. O. Hinmikaiye, O. Olakanmi, and J. Akinjobi, "Supervised Machine Learning Algorithms: Classification and Comparison," *Int. J. Comput. Trends Technol.*, vol. 48, no. 3, pp. 128–138, 2017, doi: 10.14445/22312803/ijctt-v48p126.

[44] M. J. Van der Laan and Rose., *Targeted learning*. 2012.

[45] M. L. Petersen and M. J. Van Der Laan, "Causal models and learning from data: Integrating causal modeling and statistical estimation," *Epidemiology*, vol. 25, no. 3, pp. 418–426, 2014, doi: 10.1097/EDE.0000000000000078.

[46] M. J. van der Laan and S. Rose, *Targeted Learning*, vol. 20. New York, NY: Springer-Verlag New York, 2011.

[47] M. L. Petersen, "Applying a Causal Road Map in Settings with Time-dependent Confounding," *Empidemiology*, vol. 25, no. 6, pp. 898–901, 2014, doi: 10.1117/12.2549369.Hyperspectral.

[48] A. S. Benjamin *et al.*, "Modern machine learning outperforms GLMs at predicting spikes," *bioRxiv*, pp. 1–13, 2017, doi: 10.1101/111450.

[49] M. L. Petersen and M. J. Van Der Laan, "Causal models and learning from data: Integrating causal modeling and statistical estimation," *Epidemiology*, vol. 25, no. 3, pp. 418–426, 2014, doi: 10.1097/

EDE.0000000000000078.

[50] J. Pearl and D. Mackenzie, *The Book of Why*. New York: Basic Books, 2018.

[51] C. Lago, L. Casais, E. Dominguez, and J. Sampaio, “The effects of situational variables on distance covered at various speeds in elite soccer,” *Eur. J. Sport Sci.*, vol. 10, no. 2, pp. 103–109, 2010, doi: 10.1080/17461390903273994.

[52] J. G. Claudino, D.-O. Capanema, T.-V. De-Souza, J. C. Serrão, A.-C. Machado Pereira, and G.-P. Nassis, "Current Approaches to the Use of Artificial Intelligence for Injury Risk Assessment and Performance Prediction in Team Sports: a Systematic Review," *Sport. Med. - Open*, vol. 5, no. 1, 2019.

[53] E. Morgulev, O. H. Azar, and R. Lidor, “Sports analytics and the big-data era,” *Int. J. Data Sci. Anal.*, vol. 5, no. 4, pp. 213–222, 2018, doi: 10.1007/s41060-017-0093-7.

[54] M. J. van der Laan and S. Rose, *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. 2018.

[55] D. Bzdok, N. Altman, and M. Krzywinski, “Points of Significance: Statistics versus machine learning,” *Nature Methods*, vol. 15, no. 4. Nature Publishing Group, pp. 233–234, Apr. 03, 2018, doi: 10.1038/nmeth.4642.

[56] D. van Poppel *et al.*, “Risk factors for overuse injuries in short- and long-distance running: A systematic review,” *J. Sport Heal. Sci.*, vol. 10, no. 1, pp. 14–28, Jan. 2021, doi: 10.1016/J.JSHS.2020.06.006.

CHAPTER 2

Personalized physical activity coaching: a machine learning approach

Based on
"Personalized Physical Activity Coaching:
A Machine Learning Approach."

Talko B. Dijkhuis
Frank J. Blaauw
Miriam W. van Ittersum
Hugo Velthuijsen
Marco Aiello

2018. Sensors 18(2):1–20.
doi: 10.3390/s18020623.

ABSTRACT

Living a sedentary lifestyle is one of the major causes of numerous health problems. To encourage employees to lead a less sedentary life, the Hanze University started a health promotion program. One of the interventions in the program was the use of an activity tracker to record participants' daily step count. The daily step count served as input for a fortnightly coaching session. In this paper, we investigate the possibility of automating part of the coaching procedure on physical activity by providing personalized feedback throughout the day on a participant's progress in achieving a personal step goal. The gathered step count data was used to train eight different machine learning algorithms to make hourly estimations of the probability of achieving a personalized, daily steps threshold. In 80% of the individual cases, the Random Forest algorithm was the best performing algorithm (mean accuracy = 0.93, range = 0.88–0.99, and mean F1-score = 0.90, range = 0.87–0.94). To demonstrate the practical usefulness of these models, we developed a proof-of-concept Web application that provides personalized feedback about whether a participant is expected to reach his or her daily threshold. We argue that the use of machine learning could become an invaluable asset in the process of automated personalized coaching. The individualized algorithms allow for predicting physical activity during the day and provides the possibility to intervene in time.

Keywords

Physical activity; machine learning; coaching; sedentary lifestyle.

1. INTRODUCTION

Unhealthy lifestyles lead to increased premature mortality and are a risk factor for sustaining noncommunicable diseases (NCDs) such as cardiovascular diseases, cancers, chronic respiratory diseases, and diabetes [1]. NCDs caused 63% of all deaths that occurred globally in 2008 [1]. There are four behavioural factors that have a significant influence on the prevention of NDCs: healthy nutrition, not smoking, maintaining a healthy body weight, and sufficient physical activity. Insufficient physical activity is one of the leading risk factors for the major NCDs and not meeting the recommended level of physical activity is associated with approximately 5.3 million deaths that occurred globally in 2008 [2].

A high amount of sedentary time without sufficient daily physical activity leads to a higher rate of all-cause mortality [3]. Besides the increased risk of premature mortality in the long term, the short-term quality of life, being able to work, and social participation is also threatened by insufficient physical activity [4]. Fortunately, these risks are eliminated when this sedentary time is compensated for with sufficient physical activity of moderate intensity [3].

In Western civilization, living a sedentary lifestyle is the rule rather than the exception, as many people work in office environments. In pursuance of preventing the negative effects of insufficient physical activity in the workplace, the Hanze University of Applied Sciences Groningen (HUAS), a large university in the northern part of the Netherlands, started a novel initiative named (in Dutch): 'Het Nieuwe Gezonde Werken' (The New Healthy Way of Working; HNGW). With HNGW, the HUAS aims to promote a healthy lifestyle and physical activity during the workday. HNGW consists of providing participants with educational group meetings, food boxes with healthy recipes, and individual coaching sessions supplemented with an activity tracker. Despite the fact that participants are coached every two weeks and measured continuously, it remains difficult for a coach to provide timely personalized feedback. The manual task of creating personalized feedback is time consuming, and as such it is not always possible for the participants to get in-depth and timely daily feedback on their progression. Furthermore, current activity trackers do not provide a prediction for reaching the daily goal.

In order to fill this gap, we propose a novel, personalized, and flexible machine-learning-based procedure that can automate a part of the coaching process and serve as a source of information on a participant's progress with physical activity during the day. The personalized model provides, throughout the day, information on the probability of the participant meeting his or her daily physical activity goal. We demonstrate the accuracy and effectiveness of this solution in practice by training different machine learning

algorithms and evaluating their performance using a train-test split dataset from the HNGW data. We apply techniques like grid search and cross-validation to optimize each model in order to find their best configuration. To show the applicability of this research in practice, we developed a proof-of-concept Web application, which has, to the best of our knowledge, not been done before. With the personalized actionable information, the application provides, we demonstrate that machine learning automating is feasible as a part of the coaching process. The techniques described in this work could serve two goals in the field of personalized coaching. Firstly, we envision how coaches can use such applications and how these applications can provide them with detailed insight about the participants' activity during the day. Secondly, the tool could be used as a self-support tool, in which the participants' engagement with their lifestyle might increase as a result of the extra feedback.

2. RELATED WORK

A number of studies have been performed on physical activity over days, where the sources of variance in activity is related to the subject, the day of the week, the season, and occupational and non-occupational days [5]. Tudor-Locke et al. (2005) showed that the individual is the main source of variability in physical activity next to the difference between the Sunday and the rest of the week [6]. Another study identified physical inactivity being lower on weekend days, and Saturday was the most active day of the week for both men and women [5].

To reduce sedentary time and increase physical activity levels, individuals need to change their behaviour and daily routines. This is hard to achieve because of various reasons and requires interventions and coaching strategies that use well-established techniques to induce a behaviour change. A review by Gardner et al. (2016) found that self-monitoring, problem solving, and restructuring the social or physical environment were the most promising behaviour change strategies, and—although the evidence base is quite weak—advises environmental restructuring, persuasion, and education to enhance self-regulatory skills [7]. Interventions aimed at increasing physical activity levels or reducing sedentary time varies widely in content and in effectiveness. For example, studies focusing on exercise training and behavioural approaches have demonstrated conflicting results, whereas interventions focusing on reducing sedentary time seem to be more promising [8]–[12]. The use of active video games seems to be effective in increasing physical activity, but has inconsistent findings on whether they are suitable to meet the recommended levels [13]. Also, interventions targeting recreational screen time reduction might be effective when using health promotion curricula or counselling [14]. Web- or app-based interventions to improve diet, physical activity, and sedentary

behaviour can be effective. Multi-component interventions appear to be more effective than stand-alone app interventions, although the optimal number and combination of app features and level of participant contact needed remain to be confirmed [15], [16]. The workplace is often used for health promotion interventions. Recent reviews on workplace interventions for reducing sitting at work found initial evidence that the use of alternative workstations (sit-stand desks or treadmills) can decrease workplace sitting by thirty minutes to two hours. In addition, one review found that interventions promoting stair use and personalized behavioural interventions increase physical activity, while the other found no considerable or inconsistent effects of various interventions [17], [18].

Step counters provide an objective measure of activity levels and enable self-monitoring. Furthermore, most modern consumer-based activity trackers already contain several behaviour change models or theories [19], [20]. Therefore, based on the aforementioned, using activity trackers in interventions to promote healthy lifestyles is promising. From meta-analyses by Qiu et al. and Stephenson et al. it was concluded that step counter use was indeed associated with small but significant effects in reducing sedentary time [21], [22]. Adding an activity tracker to physical therapy or counselling was effective in some populations [23]–[25]. Besides collecting activity data for therapy or counselling, it is known that the Fitbit itself also serves as an intervention mechanism [26]. The mere fact of wearing an activity tracker (even without any form of coaching) could motivate physical activity and improve health-related quality of life [27], [28]. On the other hand, studies on workplace interventions using activity trackers report conflicting results [29]–[33].

There are several studies that use sensor or activity tracker data to build a custom-made application to support research. An example is the social computer game, Fish'n'Steps, which connects the daily steps of an employee to the growth and activity of the individual avatar fish in a virtual fish tank. The more one is active, the faster the fish grows and prospers [34]. Another example is the study on increased physical activity as the effect of social support groups using pedometers and an app [35].

Although applying machine learning to coaching is new, machine learning techniques in combination with sensors have been applied before to identify the type of activity. Identifying human activity using machine learning and sensor data have been studied, for example, by Wang et al. for recognizing human daily activities from an accelerometer signal [36], by Li et al. on the quantification of the lifetime circadian rhythm of physical activity [37], or by Catal et al. on the use of an ensemble of classifiers for accelerometer-based activity recognition [38]. Only a few studies have investigated the use of actionable, data-driven predictive models. A study on creating a predictive physical fatigue model based on sensors identified relevant features for predicting physical fatigue, however

the model was not proven to be predictive enough to be applied [39].

In order to improve physical activity in combination with activity trackers, a coaching feature is helpful, but only when the messages are personal and placed in context [40]. Perceiving the coaching information as personal and relevant is crucial for the effectiveness of (e)Coaching [41]. Such tailored (e)Coaching has many aspects, two of which are personalization and timing [42]. Timeliness of information is important for participants to be able to process the information and apply the advice while it is still relevant for them. In order to provide such advice, access to real-time predictions is vital, as it allows for timing the moment of coaching, either virtual or in real life and as flexible as needed. To the best of our knowledge, no studies exist about the use of sensor data combined with machine learning techniques for creating validated and individualized predictive models on physical activity. The individualized models could help the coach and the participant in the process of behaviour change and increased physical activity.

3. MATERIALS AND METHODS

The present work revolves around the HNGW project. This project was started in 2015 and focuses on promoting a healthy lifestyle. We describe the design of this study and how the resulting data is used in the present work. Next, we describe our analysis pipeline. We describe the conversion of the raw data set into a feature set, the evaluation methods of the predictive models, and the choice of the algorithms. Finally, we shed light on the proof-of-concept application we created to demonstrate how these techniques could be used in practice.

3.1. Study Design

The goal of the workplace health promotion intervention HNGW at the HUAS was to increase physical activity during workdays, by improving both physical and mental health, and several work-related variables. In the study, several performance-based tests and self-reported questionnaires were used to assess its effectiveness on a group level.

Forty-eight eligible participants from the HUAS were randomized into two groups, stratified according to age, gender, BMI, and baseline self-reported health. One group followed a twelve-week workplace health promotion intervention; the other served as a control during the first twelve weeks and thereafter received the twelve-week workplace health promotion intervention.

During the study, minutely step count data of the participants was collected. Step count was measured using a wrist-worn activity tracker, the Fitbit Flex. The Fitbit Flex has

been shown to be a reliable and valid device for step count and suitable for health enhancement programs [13]. Further details of the trial design on HNGW at the HUAS are represented in the manuscript of van Ittersum et al.[43].

3.2. Data Set

The anonymized data used in the present study was collected from participants during their participation in the HNGW health promotion program. All participants provided informed consent for participation in the HNGW study and for the use of their anonymized data for research purposes.

We used the steps per minute of each participant, resulting in a total of 349,920 measurements across all participants. We only considered the step data collected during the intervention period. That is, for both the intervention and the control group, we used the last twelve weeks of available step data. By focusing on the intervention period, we have a more homogeneous sample than we would have when including both the intervention and control data.

While the Fitbit platform provides us with several minutely measures (e.g., steps, metabolic equivalent of tasks [METs], calories, and distance), in our analysis we only included the steps variable. We used the steps variable as we expect it to be the most accurate and relevant, as all other variables are by-products derived using approximation algorithms.

3.3. Data Processing, Transformation, and Performance

To prepare the available minutely step data as input for training the algorithms, we first performed a data cleaning, reformatting, and pre-processing step. First, we removed incomplete days from the data set. We also removed all days with zero steps and weekend days. We then converted all provided variables in a format that could be used by our algorithms, by augmenting our initial data set with several new augmented variables, such as hour of the workday, the number of steps for that hour, and a cumulative sum of the number of steps till that hour.

Note that we define a workday as the weekdays Monday to Friday. The normal working hours at the university are between 8:00 AM and 5:00 PM. The HNGW tried to motivate the participants to walk at least a part of the distance they commute daily. As a consequence, the hours of interest are the combination of the working hours and the period of commuting. Therefore, we only considered the number of steps per hour between 7:00 AM and 6:00 PM. As features for training the algorithms, we used the hour per workday (ranged from 7:00 AM to 6:00 PM), the number of steps of that hour, and the cumulative sum of the number of steps till that hour.

As the outcome measure, we calculated the average number of steps for all workdays over all weeks. That is, for each individual, we calculated one average for all workdays. We considered the number of steps between 7:00 AM and 6:00 PM. Note that this outcome measure is not used as input in the training process. We constructed a binary outcome variable represented by the indicator variable $\boldsymbol{Y_j = \mathbf{I}(s_j \geq \theta_j)}$, in which $\boldsymbol{s_j}$ refers to the number of steps on a workday for individual $\boldsymbol{j}$, and $\boldsymbol{\theta_j}$ refers to the specific step goal for that $\boldsymbol{j}$. The indicator function returns one (the `true' label) when the inside condition holds, and zero (the 'false' label) otherwise.

Three days of repeated measures are necessary to represent adults' usual activity levels with an 80% confidence [6]. Forty-four participants met the criteria. The processing and transformation for these forty-four participants resulted in a total of 120,480 data blocks (for the number of steps, mean = 9,031, median = 8,543, range = 0- 47,121). The total number of positives when the threshold is met at 6:00 PM, is 1528. The total number of negatives when the threshold is not met at 6:00 PM, is 1,879.

Note that we did not include any of the group level/baseline variables like age or gender, as we only considered personalized models. Although these variables might affect the outcome, they do not vary within the individual and as such do not add information.

3.4. Evaluation of the Performance of Algorithms and Models

We trained eight different machine learning algorithms. To compare their performance, we used a method known as `confusion matrices'. The confusion matrices give an overview of the true positives (TP; the model predicted a `true' label and the actual data contained a `true' label), true negatives (TN; the model predicted a `false` label and the actual data turned out to have a `false' label), false positives (FP; the model predicted a `true' label, but the actual data contained a `false' label), and false negatives (FN; the model predicted a `false' label, but in fact the data contained a `true' label) of a model. An example of a confusion matrix is provided in Table 1. These confusion matrices served as a basis for the calculation of two other performance measures: The accuracy and the F1-score [15].

Table 1. Confusion matrix.

		True class	
		Yes	**No**
Predicted class	**Yes**	True Positives (TP)	False Negatives (FN)
	No	False Positives (FP)	True Negatives (TN)

True Positive: the threshold of daily steps was met and predicted; True Negative: the threshold of daily steps was not met and predicted; False Negative: the threshold of daily steps was met and not predicted; False Positive: the threshold of daily steps was not met and not predicted.

Accuracy is a metric to determine the nearness of the prediction to the true value. A value of the accuracy close to one indicates the best performance. It calculates the ratio between the correctly classified cases and all cases as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}.$$

Besides the accuracy metric, we calculated the F1-score for each model. Similar to the accuracy metric, the F1-score takes its values from between zero and one, one corresponding to the best performance. To calculate the F1-score, we use two other metrics known as the precision and the recall of the model. Precision is the proportion of the true positives and the false negatives and is calculated as

$$\text{Precision} = \frac{TP}{TP + FN}.$$

Recall is the true positive rate, which is calculated as

$$\text{Recall} = \frac{TP}{TP + FP}.$$

Using these definitions of precision and recall, the F1-score can be calculated as

$$\text{F1-score} = 2 \times \frac{Precision \times Recall}{Precision + Recall}.$$

3.5. Computing the Personalized Predictive Model

We aim to predict (throughout the day) whether or not an individual will meet his or her daily step goal. Prediction of meeting a set goal is a supervised two-class classification problem. Nowadays, many different algorithms for performing such classifications are available. Unfortunately, it is generally considered impossible to determine *a priori* which algorithm will perform best on any given data set [44]. Although distinct algorithms are better suited for different types of data and problems, the type of algorithm is merely an indication of the most suitable algorithm. Currently, the preferred way to find the best-performing algorithm is by empirically testing each of them [45]. Nevertheless, there exist general guidelines to direct the search for specific algorithms for the problem at hand. One of the leading organizations on open source machine learning library, scikit-learn.org, offers a flowchart about which algorithms can be chosen in which situation [46]. Also, Microsoft provides a 'cheat sheet' on their Azure machine learning platform [47]. The flow chart and ´cheat sheet´ served as a basis for our selection process and we chose the following machine learning classification algorithms: (i) AdaBoost (ADA), (ii) Decision Trees (DT), (iii) KNeighborsClassifier (KNN), (iv) Logistic Regression (LR), (v) Neural Networking(NN), (vi) Stochastic Gradient Descent (SGD), (vii) Random Forest (RF), and (viii) Support Vector Classification (SVC). The performance of each of these

algorithms was first determined for seventy percent of the whole dataset including five-fold cross-validation with scaling of the factors for KNN, NN, SGD, and SVC. Subsequently, for every participant we individualized the algorithms with five-fold cross-validation and grid search on selected hyperparameters. Seventy percent of the available individual data was used as training data. After training the algorithms, the algorithms were turned into persistent predictive models per participant. We used the individual models to construct confusion matrices, which in turn served as a basis for the F1-score and the accuracy per individual predictive model. To compare the performance of the machine learning models, we included a baseline model. This baseline model checks the cumulative step count. If this cumulative step count equals or exceeds the average personalized goal, the model returns true and false otherwise. We ranked all machine learning models (including the baseline model) using the average of the F1-score and the accuracy.

3.6. Proof of Concept

We designed and implemented a Web application to demonstrate how the personalized prediction based on machine learning and activity tracker data could be used in practice. We developed this application as a Web application, which can be accessed on http://personalized-coaching.compsy.nl/. In this application, the user can input the values `Hour of the day', `Steps previous hour', `Total steps till the Hour', combined with the participant's ID and the algorithm to use. The Web application then uses the individualized model and input data to predict the outcome together with the probability thereof.

3.7. Implementation Details

We used scikit-learn (v0.18, [48]) to establish the best predictive model for the individual. Scikit-learn is an open-source Python module integrating a wide range of machine learning algorithms. Scikit-learn was integrated in Anaconda (v4.2.13, [49]) and Jupyter Notebooks (4.0.6, [50]) was used in combination with Python (v3.5.2, [49]) for creating the data processing and machine learning pipeline. Jupyter Notebooks is an interactive method to write and run various programming languages, such as Python. The participants, their physical activity data, and the results of the performance of the algorithms and models were saved in an Oracle database (v11g2 XE; [51]). The Oracle database management system is a widely used SQL-based system for persisting data. The source code and corresponding notebooks of the machine learning procedure is available as open-source software on Github (https://github.com/compsy/personalized-coaching-ml).

For the Web application, we used Flask (Version 0.10.1, [52]), a Python-based Web application microframework for developing Web applications. We used a PostgreSQL database to store information regarding the models and the participants. The machine learning models resulting from the pipeline are exported as Python Pickle files, which were imported into the Web application. The infrastructure-as-a-service provider

Heroku is used to host a demo version of the Web application. This Web application is available at http://personalized-coaching.compsy.nl. The Web application is available as open-source software on Github https://github.com/compsy/personalized-coaching-app).

4. RESULTS

After optimizing our machine learning models by applying grid search in combination with cross-validation, we assessed the models using the test set. The results are presented here.

4.1. Accuracy and F1-Score on Group Level

Table 2 presents the F1-score and accuracy of the eight different algorithms at the group level. The top three group algorithms based on the mean accuracy and F1-score are: AdaBoost, Neural Networking, and Support Vector Classifier.

Table 2. Algorithms and their scores for the whole dataset.

Algorithm Name	Mean Accuracy (standard deviation)	Mean F1 (standard deviation)	Rank
AdaBoost (ADA)	0.776623 (0.002080)	0.854157 (0.001626)	1
Neural Networking (NN)	0.777774 (0.001545)	0.852797 (0.002938)	2
Support Vector Classifier (SVC)	0.770728 (0.002505)	0.856341 (0.002405)	3
Stochastic Gradient Descent (SGD)	0.767623 (0.005490)	0.853575 (0.004574)	4
KNeighborsClassifier (KNN)	0.749171 (0.005683)	0.829826 (0.005544)	5
Logistic Regression (LR)	0.742125 (0.009821)	0.825725 (0.008487)	6
Random Forest (RF)	0.737451 (0.003210)	0.819065 (0.003840)	7
Decision Tree (DT)	0.720535 (0.004787)	0.804220 (0.003006)	8

We visualized the accuracy and F1-score per algorithm using boxplots in Figures 1 and 2. The box represents the second and third quartile groups and the red line indicates the median. The upper whisker visualizes the fourth quartile group and the lower whisker visualizes the first quartile group. Finally, the plus sign indicates outliers on either side of both whiskers.

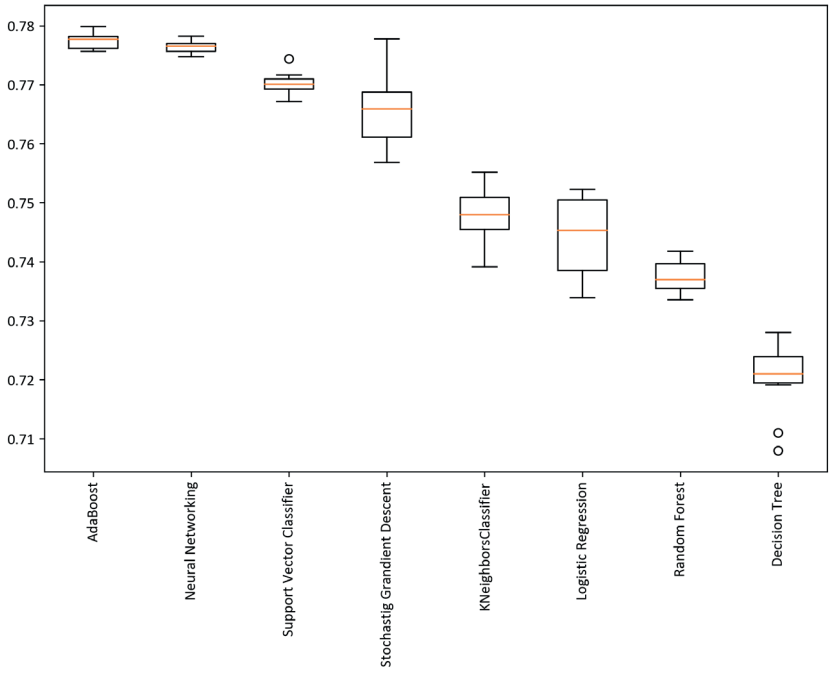


Figure 1. Algorithm accuracy comparison.

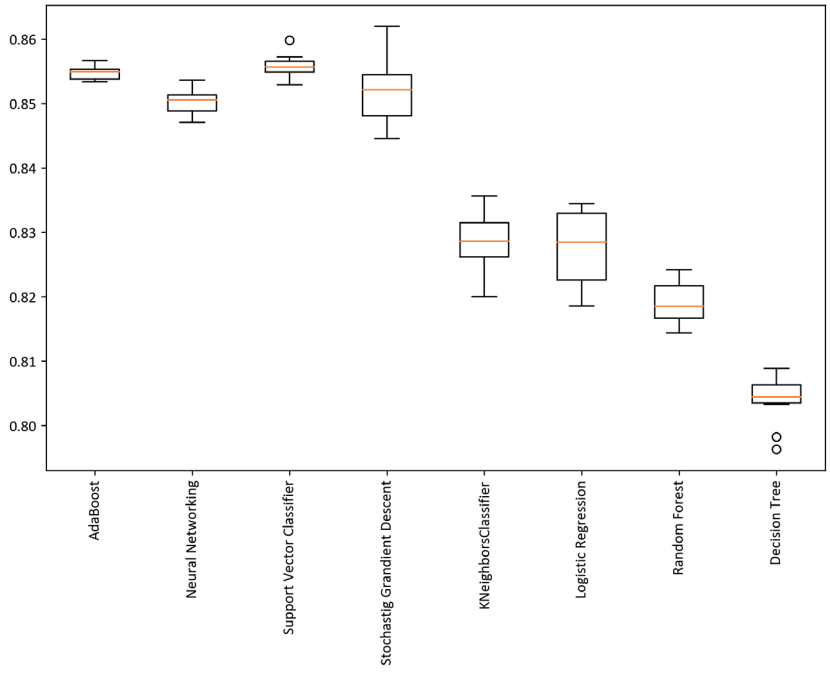


Figure 2. Algorithm F1-score comparison.

4.2. Individual Algorithms

We trained all algorithms on the training set of each individual and performed cross-validation to tune the hyperparameters. Table 3 lists the used machine learning algorithms, the set of tested hyperparameters, and the selected grid search values.

Table 3. Algorithms, used parameters, and grid search values.

Algorithm name	Hyperparameters	Values
AdaBoost (ADA)	n_estimators: number of decision trees in the ensemble	[10,50]
	learning rate: the shrink of the contribution of each successive decision tree in the ensemble	[0.1, 0.5, 1.0, 10.0]
Decision Tree (DT)	criterion: the algorithm to use to decide on split	['gini', 'entropy']
	max_features: the number of features to consider when to split	['auto','sqrt','log2']
KNeighborsClassifier (KNN)	metrics: the distance metric to use	['minkowski','euclidean', 'manhattan']
	weights: weight function used	['uniform','distance']
	n_neighbors: number of neighbors to use for queries	[5, 6, 7, 8, 9]
Neural Networking (NN)	learning_rate_init: the control of the step-size in updating the weights	['constant', 'invscaling', 'adaptive']
	activation: the activation function for the hidden layer	['identity', 'logistic', 'tanh', 'relu']
	learning_rate: the rate for the weight of the updates	[0.01, 0.05, 0.1, 0.5, 1.0]
Logistic Regression (LR)	C: regularization strength	[0.001, 0.01, 0.1, 1, 10, 100, 1000]
	penalty: whether to use Lasso (L1) or Ridge (L2) regularization	['l1', 'l2']
	fit_intercept: whether or not to compute the intercept of the linear classifier	[True, False]
Stochastic Gradient Descent (SGD)	fit_intercept: whether or not the intercept should be computed	[True, False]
	l1_ratio: the penalty is set to L1 or L2	[0,0.15,1]
	loss: quantification of the loss	['log','modified_huber']
Support Vector Classifier (SVM)	kernel: the kernel type to be used in the algorithm	['linear','rbf']
Random Forest (RF)	n_estimators:number of decision trees	[10, 50, 100, 500]
	max_features: the number of features to consider when to split	[0.1, 0.25, 0.5, 0.75, 'sqrt', 'log2', None]
	criterion: which algorithm should be used to decide on split	['gini', 'entropy']

The accuracy and F1-score of the individual algorithms differ. Figure 3 visualizes the results of the average of the individual scores.

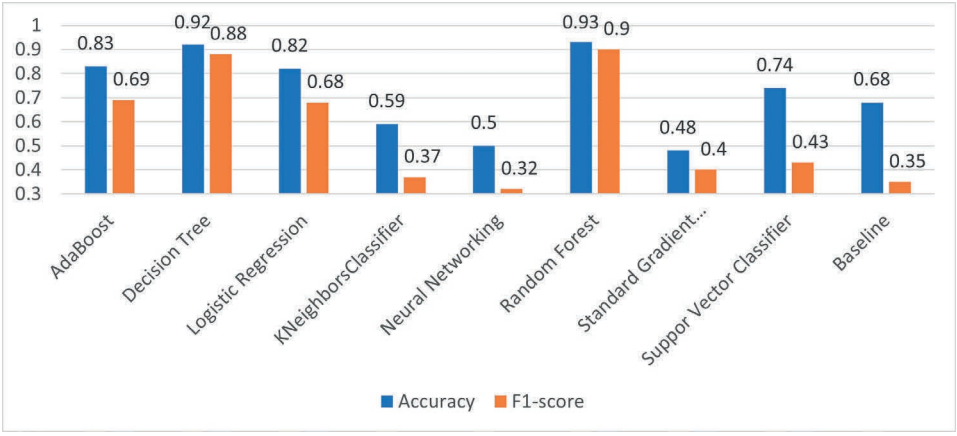


Figure 3. Average accuracy and F1-score per model.

For thirty-five subjects, the best-performing individual model was the Random Forest algorithm, in eight cases this was the Decision Tree algorithm, and for one subject the AdaBoost algorithm performed best. The average accuracy of the Random Forest algorithm is 0.93 (range 0.88–0.99). Thus, in terms of accuracy, the individual Random Forest models score better than its counterpart that was generalized over all individuals (mean personalized accuracy = 0.93 versus mean generalized accuracy = 0.82). The average accuracy of the Decision Tree model is 0.93 (range 0.91–0.97) and outperforms the generalized, group-based Decision Tree accuracy of 0.75. The accuracy of the single AdaBoost model is 0.98, which outperforms the group accuracy of 0.85.

The mean F1-score of the Random Forest model is 0.90 (range 0.87–0.94). The mean F1-score of the Decision Tree model based on the eight best performing models is 0.90 (range 0.87–0.93). Finally, the best AdaBoost model has an F1-score of 0.92, while the group accuracy for the AdaBoost algorithm was 0.77.

The use of grid search to tune the hyperparameters of the algorithms led to several optimized models per individual. To demonstrate the difference this optimization operation can have, we present an example of two individual models with different hyperparameter configurations in Table 4. Table 5 gives an overview of the number of occurrences of a value for the Random Forest hyperparameters.

Table 4. Example of different tuned personalized Random Forest models.

Participant	Parameters	Values
1119	criterion	gini
	max_features	sqrt
	n_estimators	50
1121	criterion	entropy
	max_features	log2
	n_estimators	50

Table 5. The number of different values per Random Forest hyperparameter.

Hyperparameter	Value	Number of Occurrences
criterion	entropy	7
	gini	37
max_features	0.1	4
	0.25	5
	0.5	7
	0.75	15
	log2	2
	sqrt	2
	null	9
n_estimators	10	3
	100	17
	50	16
	500	6

The accuracy and F1-score of the various machine learning algorithms increase slightly during the day. The size of this increase differs slightly per machine learning algorithm. For instance, the F1-score of Random Forest increases with 10% during the day, starting with an F1-score of 0.89 at 7:00 AM and ending with an F1-score of 0.97 at 6:00 PM. Both Figures 4 and 5 also show the increase in accuracy and F1-score of the baseline algorithm during the day. Its accuracy starts with 0.55 and ends at 1 at the end of the workday, while the F1-score starts at 0 and ends at 1. The accuracy increases for Random Forest, Logistic Regression, and AdaBoost, whereas the accuracy of Neural Networking is best at 11:00 AM and Stochastic Gradient Descent remains the same.

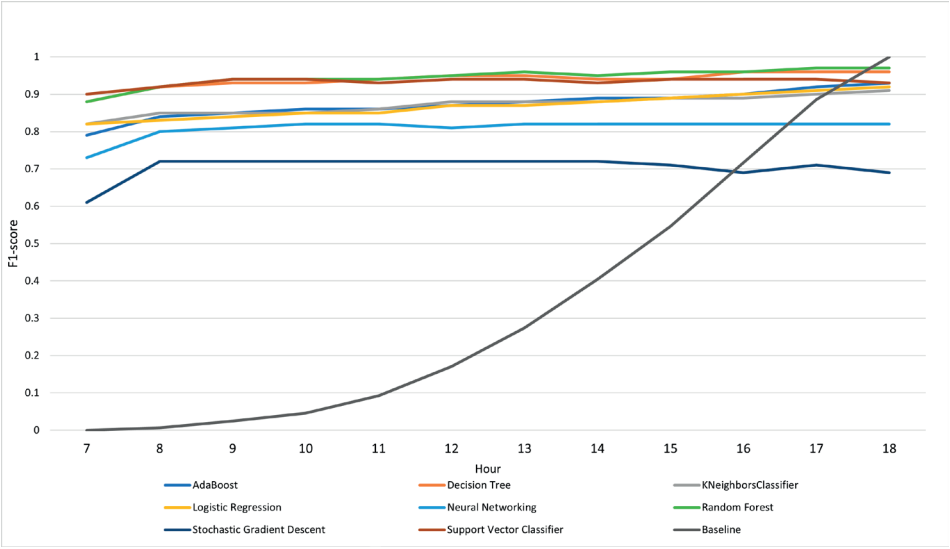


Figure 4. Average F1-Score per algorithm, per hour based on the individual scores.

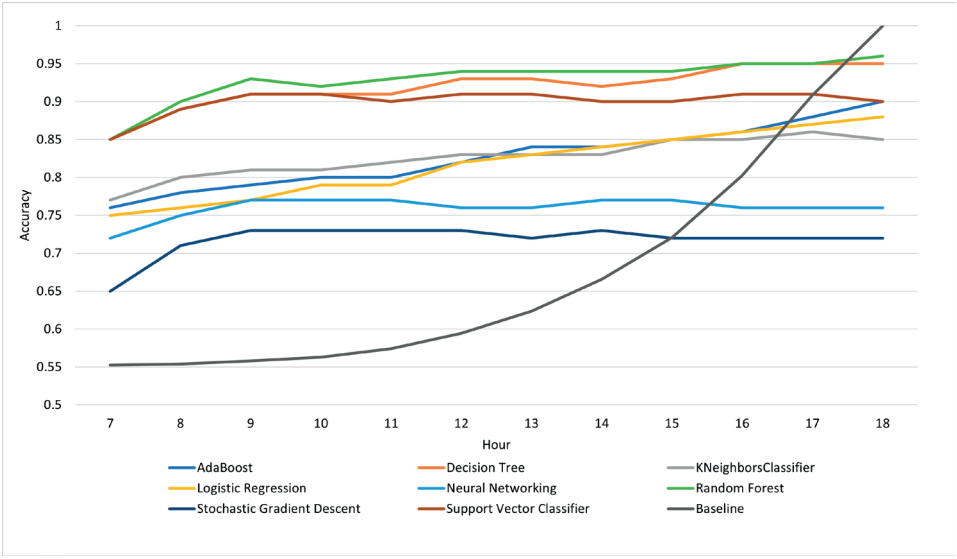


Figure 5. Average accuracy per algorithm, per hour based on the individual scores.

4.3. The Web Application

The Web application is a demonstration of how the aforementioned machine learning techniques could be used in practice, from the perspective of both the coach and the participant. The application allows the user to determine whether a participant will achieve his or her goal for the day, during the day, by applying the individualized algorithms. The procedure for predicting this goal is as follows. First, the user selects a participant identifier from the dropdown menu. After this selection had been made, the application selects the best and personalized machine learning algorithm for this specific participant. Then the user can fill out a form, providing the necessary details to base the prediction on (hour of day, the number of steps so far, and the number of steps in the past hour). Finally, when the user submits the form, the application returns advice personalized for the individual selected from the dropdown menu. The demo application is available at http://personalized-coaching.compsy.nl/. Figure 6 provides a screenshot of both the input fields of the application and the generated prediction and advice.

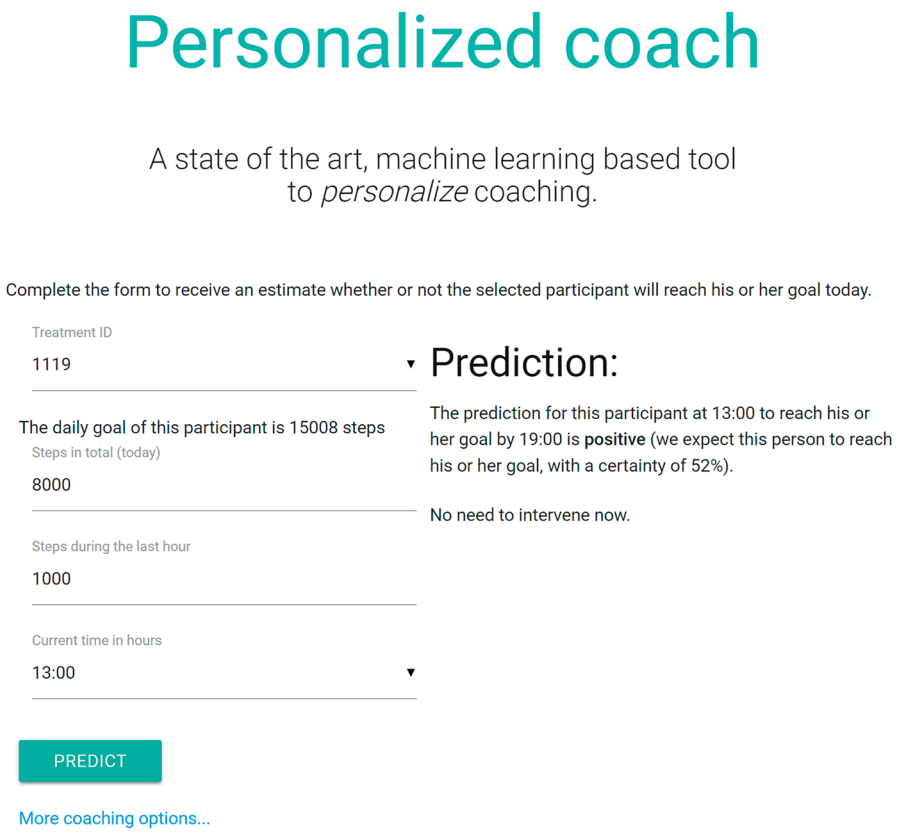


Figure 6. Screenshot of the Personalized Coach Web Application.

5. DISCUSSION

We investigated machine learning as a means to support personalized coaching on physical activity. We demonstrated that for our particular data sets, the tree algorithms and tree-based ensemble algorithms performed especially well. To demonstrate how the results of machine learning techniques could be used in practice, an application was used to aid the coaching of the physical activity process. Furthermore, the analysis shows that selecting the right algorithm, using the dataset of the individual participant, and tuning its individual algorithm parameters, can lead to significant improvements in predictive performance and is a critical step in machine learning application. All source code, including the different notebooks and the proof-of-concept Web application is available online as open-source software. The source code can serve as a blueprint for other researchers when aiming to apply machine learning for coaching.

Although Random Forest outperformed most of the other algorithms, it is problematic to provide a generalized recommendation for specific algorithms, parameters, or parameter settings [44]. Presumably due to individually different physical activity patterns, different algorithms and parameters have to be considered. As a starting point, we selected the algorithms based on well-established sources [41], [42], applied cross-validation, and grid-searched the values of the selected parameters. Nevertheless, it's important to note that these algorithms, parameters, and grid search values might not work best on all individual physical patterns, and the algorithms, parameters, and grid search values should only be used as starting points. Future work might consist of investigating the underlying mechanisms to be able to choose the best algorithm beforehand.

We based the prediction solely on the hour of the day and the number of steps. These steps are naturally increasing over the day, and as such, not independent from each other. By including the cumulative number of steps for each block of data, and by including the number of steps made in the past hour, we assume each block to be independent from the other blocks, and as such, are still able to use the regular machine learning methods. A limitation of the present work is that all participants included in this study participated in an intervention. This intervention might have made the participants more aware and engaged with the project, and as such, the individualized models might be biased towards the best scenario. When people are not extrinsically motivated to meet their daily physical activity goal, and lower their physical activity, the predictive power of the models and therefor the effect of automated intervention will lessen. On the other hand, when an intervention like the health promotion program ends, the individualized models check the participant on his or her performance as if the program is supporting the participant.

As presented in the state of the art literature, the total number of steps differ significantly between Sunday and rest of the weekdays [5], [6]. Within this health promotion program, the focus was on improving physical activity during working hours and commuting. Therefore, the machine learning models were trained based on the normal workweek. Only one model per participant, based on the five weekdays, is adequate to predict whether or not a participant will meet his or her threshold. It may be necessary to conduct different models for the weekend and weekdays when a health promotion program is expanded to weekends. A reason to establish more than one or two models per participant is found in the variances between weekdays [5]. Examples of different factors that could influence the level of physical activity are weekly sport obligations, weekly meetings, or lunch walks on certain days. Constructing a model per weekday might establish an even more personalized and precise prediction.

In the present work, we only train our machine learning algorithms on variables provided by the activity tracker, extending this set of variables with other (exogenous) variables from other data sources. For example, the data can be extended to include information on the changes in the weather conditions and/or season, which are known to correlate with the day-to-day activity [5], [53], or non-working time during weekdays like national holidays and free time, or part-time working schedule, for the activity level differs between non-occupational and occupational time, or the influence and effectiveness of coaching and interventions. Adding the mentioned factors to the dataset might improve the predictive accuracy of the model and might increase the effectiveness of the coaching process.

To apply the personalized machine learning models effectively, they have to become a part of a larger ecosystem. An ideal coaching process is fully tailored to the individual participant. One of the most important characteristics of the personalization of a coaching strategy consists in the timing and ease to execute triggers to change behaviour [54]. To support these two aspects of coaching, timely information on the participant and the effectiveness of the coaching strategy are needed. Coaching might not be limited to a personal real-life coach but also may include virtual coaching. An example of a possible use of the system is: at the moment the participant doesn't score a 'yes' for two hours in a row on the prediction of meeting his threshold, a notification is sent out to both the participant and the coach. On the basis of this notification, the participant and the coach can take action; the coach can timely intervene to stimulate his client to become physically active and the participant can become instantly more active. Blok et al. proposed a system which combines the real-time analysis of activity tracker data and other personal streaming data as well as the evaluation of virtual coaching strategies, which enables it to tune the coaching to the person [55]. The present work could serve as a central component of a virtual coach system like that envisioned by Blok et al. [55].

To make the information even more personal and relevant, a promising direction for future work is to include a prediction of the actual number of steps at the end of the day. Adding more (and personalized) information might strengthen the effectiveness of the system. To do so, we could apply a similar procedure to the one presented in this study, but instead replace our classification algorithms with regression machine learning algorithms. The predicted number of steps could be a valuable extension in addition to the currently implemented classification of the step goal.

To conclude, machine learning is a viable asset to automate personalized daily physical activity prediction. Coaching can provide accurate and timely information on the participants' physical activity, even early in the day. This is the result of applying machine learning to the behaviour of the individual participant as precisely and frequently measured by wearable sensors. The prediction of the participant meeting his goal in combination with the probability of such achievement allows for early intervention and can be used to provide support for personalized coaching. Also, the motivation for self-coaching might be increased, while every model is personalized, and the results are better fitted to the situation. Furthermore, machine learning techniques empower automated coaching and personalization.

Acknowledgments

We thank the Hanze University Health Program for providing the physical activity data of the Health Program and all the participants in the experiment.

REFERENCES

[1] WHO, "Global action plan for the prevention and control of noncommunicable diseases 2013-2020.," *World Heal. Organ.*, p. 102, 2013, doi: 978 92 4 1506236.

[2] I. Min-Lee *et al.*, "Effect of physical inactivity on major non-communicable diseases worldwide: An analysis of burden of disease and life expectancy," *Lancet*, vol. 380, no. 9838, pp. 219–229, 2012, doi: 10.1016/S0140-6736(12)61031-9.

[3] U. Ekelund *et al.*, "Does physical activity attenuate, or even eliminate, the detrimental association of sitting time with mortality? A harmonised meta-analysis of data from more than 1 million men and women," *Lancet*, vol. 388, no. 10051, pp. 1302–1310, 2016, doi: 10.1016/S0140-6736(16)30370-1.

[4] E. Losina, H. Y. Yang, B. R. Deshpande, J. N. Katz, and J. E. Collins, "Physical activity and unplanned illness-related work absenteeism: Data from an employee wellness program," *PLoS One*, vol. 12, no. 5, pp. 1–8, 2017, doi: 10.1371/journal.pone.0176872.

[5] C. E. Matthews *et al.*, "Sources of Variance in Daily Physical Activity Levels in the Seasonal Variation of Blood Cholesterol Study," *Am. J. Epidemiol.*, vol. 153, no. 10, pp. 987–995, 2001.

[6] C. Tudor-Locke, L. Burkett, J. P. Reis, B. E. Ainsworth, C. A. Macera, and D. K. Wilson, "How many days of pedometer monitoring predict weekly physical activity in adults?," *Prev. Med. (Baltim).*, vol. 40, no. 3, pp. 293–298, 2005, doi: 10.1016/j.ypmed.2004.06.003.

[7] B. Gardner, L. Smith, F. Lorencatto, M. Hamer, and S. J. Biddle, "How to reduce sitting time? A review of behaviour change strategies used in sedentary behaviour reduction interventions among adults," *Health Psychol. Rev.*, vol. 10, no. 1, pp. 89–112, Jan. 2016, doi: 10.1080/17437199.2015.1082146.

[8] P. R. A. Baker, D. P. Francis, J. Soares, A. L. Weightman, and C. Foster, "Community wide interventions for increasing physical activity," *Sao Paulo Medical Journal*, vol. 129, no. 6. Associação Paulista de Medicina, pp. 436-437, Dec. 2011, doi: 10.1590/S1516-31802011000600013.

[9] D. E. Conroy *et al.*, "Lifestyle intervention effects on the frequency and duration of daily moderate–vigorous physical activity and leisure screen time.," *Heal. Psychol.*, vol. 36, no. 4, pp. 299–308, 2017, doi: 10.1037/hea0000418.

[10] L. W. Cindy Ng, J. Mackney, S. Jenkins, and K. Hill, "Does exercise training change physical activity in people with COPD? A systematic review and meta-analysis," *Chron. Respir. Dis.*, vol. 9, no. 1, pp. 17–26, Feb. 2012, doi: 10.1177/1479972311430335.

[11] V. Cleland *et al.*, "Effectiveness of interventions to promote physical activity and/or decrease sedentary behaviour among rural adults: a systematic review and meta-analysis," *Obes. Rev.*, vol. 18, no. 7, pp. 727–741, Jul. 2017, doi: 10.1111/obr.12533.

[12] S. A. Prince, T. J. Saunders, K. Gresty, and R. D. Reid, "A comparison of the effectiveness of physical activity and sedentary behaviour interventions in reducing sedentary time in adults: A systematic review and meta-analysis of controlled trials," *Obes. Rev.*, vol. 15, no. 11, pp. 905–919, Nov. 2014, doi: 10.1111/obr.12215.

[13] C. Höchsmann, M. Schüpbach, and A. Schmidt-Trucksäss, "Effects of Exergaming on Physical Activity in Overweight Individuals," *Sport. Med.*, vol. 46, no. 6, pp. 845–860, Jun. 2016, doi: 10.1007/s40279-015-0455-z.

[14] L. Wu, S. Sun, Y. He, and B. Jiang, "The effect of interventions targeting screen time reduction: A systematic review and meta-analysis.," *Medicine (Baltimore).*, vol. 95, no. 27, p. e4029, Jul. 2016, doi: 10.1097/MD.0000000000004029.

[15] S. Schoeppe *et al.*, "Efficacy of interventions that use apps to improve diet, physical activity and sedentary behaviour: A systematic review," *Int. J. Behav. Nutr. Phys. Act.*, vol. 13, no. 1, p. 127, Dec. 2016, doi: 10.1186/s12966-016-0454-y.

[16] C. R. L. Beishuizen *et al.*, "Web-Based Interventions Targeting Cardiovascular Risk Factors in Middle-Aged and Older People: A Systematic Review and Meta-Analysis.," *J. Med. Internet Res.*, vol. 18, no. 3, p. e55, Mar. 2016, doi: 10.2196/jmir.5218.

[17] N. Shrestha, K. Kt, V. Jh, S. Ijaz, V. Hermans, and S. Bhaumik, "Workplace interventions for reducing sitting at work (Review)," *Cochrane Database Syst Rev.*, no. 3, 2016, doi: 10.1002/14651858.CD010912.pub3. www.cochranelibrary.com.

[18] D. A. Commissaris, M. A. Huysmans, S. E. Mathiassen, D. Srinivasan, L. L. Koppes, and I. J. Hendriksen, "Interventions to reduce sedentary behavior and increase physical activity during productive work: a systematic review.," *Scand. J. Work. Environ. Health*, vol. 42, no. 3, pp. 181–91, May 2016, doi: 10.5271/sjweh.3544.

[19] K. Mercer, M. Li, L. Giangregorio, C. Burns, and K. Grindrod, "Behavior Change Techniques Present in Wearable Activity Trackers: A Critical Analysis.," *JMIR mHealth uHealth*, vol. 4, no. 2, p. e40, Apr. 2016, doi: 10.2196/mhealth.4461.

[20] M. Duncan *et al.*, "Activity Trackers Implement Different Behavior Change Techniques for Activity, Sleep, and Sedentary Behaviors," *Interact. J. Med. Res.*, vol. 6, no. 2, p. e13, Aug. 2017, doi: 10.2196/ijmr.6685.

[21] S. Qiu *et al.*, "Step Counter Use and Sedentary Time in Adults: A Meta-Analysis.," *Medicine (Baltimore).*, vol. 94, no. 35, p. e1412, Sep. 2015, doi: 10.1097/MD.0000000000001412.

[22] A. Stephenson, S. M. McDonough, M. H. Murphy, C. D. Nugent, and J. L. Mair, "Using computer, mobile and wearable technology enhanced interventions to reduce sedentary behaviour: a systematic review and meta-analysis," *Int. J. Behav. Nutr. Phys. Act.*, vol. 14, no. 1, p. 105, Dec. 2017, doi: 10.1186/s12966-017-0561-4.

[23] H. J. de Vries, T. J. M. Kooiman, M. W. van Ittersum, M. van Brussel, and M. de Groot, "Do activity monitors increase physical activity in adults with overweight or obesity? A systematic review and meta-analysis," *Obesity*, vol. 24, no. 10, pp. 2078–2091, 2016, doi: 10.1002/oby.21619.

[24] L. C. Li, E. C. Sayre, H. Xie, C. Clayton, and L. M. Feehan, "A Community-Based Physical Activity Counselling Program for People With Knee Osteoarthritis: Feasibility and Preliminary Efficacy of the Track-OA Study," *JMIR mHealth uHealth*, vol. 5, no. 6, p. e86, Jun. 2017, doi: 10.2196/mhealth.7863.

[25] M. Miyauchi *et al.*, "Exercise Therapy for Management of Type 2 Diabetes Mellitus: Superior Efficacy of Activity Monitors over Pedometers," *J. Diabetes Res.*, vol. 2016, pp. 1–7, Sep. 2016, doi: 10.1155/2016/5043964.

[26] L. A. Cadmus-Bertram, B. H. Marcus, R. E. Patterson, B. A. Parker, and B. L. Morey, "Randomized Trial of a Fitbit-Based Physical Activity Intervention for Women.," *Am. J. Prev. Med.*, vol. 49, no. 3, pp. 414–8, Sep. 2015, doi: 10.1016/j.amepre.2015.01.020.

[27] S. Mansi, S. Milosavljevic, S. Tumilty, P. Hendrick, C. Higgs, and D. G. Baxter, "Investigating the effect of

a 3-month workplace-based pedometer-driven walking programme on health-related quality of life in meat processing workers: A feasibility study within a randomized controlled trial," *BMC Public Health*, vol. 15, no. 1, 2015, doi: 10.1186/s12889-015-1736-z.

[28] Z. H. Lewis, E. J. Lyons, J. M. Jarvis, and J. Baillargeon, "Using an electronic activity monitor system as an intervention modality: A systematic review.," *BMC Public Health*, vol. 15, p. 585, Jun. 2015, doi: 10.1186/s12889-015-1947-3.

[29] R. Freak-poli, M. Cumpston, A. Peeters, and S. Clemes, "Workplace pedometer interventions for increasing physical activity (Review)," *Cochrane database Syst. Rev.*, vol. 4, no. 4, p. CD009209, 2013, doi: 10.1002/14651858.CD009209.pub2.www.cochranelibrary.com.

[30] S. Compernolle, C. Vandelanotte, G. Cardon, I. De Bourdeaudhuij, and K. De Cocker, "Effectiveness of a web-based, computer-tailored, pedometer-based physical activity intervention for adults: a cluster randomized controlled trial.," *J. Med. Internet Res.*, vol. 17, no. 2, p. e38, Feb. 2015, doi: 10.2196/jmir.3402.

[31] S. M. Slootmaker, M. J. M. Chinapaw, A. J. Schuit, J. C. Seidell, and W. Van Mechelen, "Feasibility and Effectiveness of Online Physical Activity Advice Based on a Personal Activity Monitor: Randomized Controlled Trial," *J. Med. Internet Res.*, vol. 11, no. 3, p. e27, Jul. 2009, doi: 10.2196/jmir.1139.

[32] J. Poirier *et al.*, "Effectiveness of an Activity Tracker- and Internet-Based Adaptive Walking Program for Adults: A Randomized Controlled Trial," *J. Med. Internet Res.*, vol. 18, no. 2, p. e34, Feb. 2016, doi: 10.2196/jmir.5295.

[33] E. A. Finkelstein *et al.*, "Effectiveness of activity trackers with and without incentives to increase physical activity (TRIPPA): a randomised controlled trial," *Lancet Diabetes Endocrinol.*, vol. 4, no. 12, pp. 983–995, Dec. 2016, doi: 10.1016/S2213-8587(16)30284-4.

[34] L. Mamykina *et al.*, "Fish'n'Steps: Encouraging Physical Activity with an Interactive Computer Game," in *Ubicomp 2006:Ubiquitous Computing*, 2006, vol. 4206, no. August 2015, doi: 10.1007/11853565.

[35] T. Toscos, A. Faber, K. Connelly, and A. M. Upoma, "Encouraging physical activity in teens. Can technology help reduce barriers to physical activity in adolescent girls?," in *Pervasive Computing Technologies for Healthcare, 2008. PervasiveHealth 2008*, 2008, vol. 3, no. Group 3, pp. 218–221.

[36] J. Wang, R. Chen, X. Sun, M. F. H. She, and Y. Wu, "Recognizing human daily activities from accelerometer signal," *Procedia Eng.*, vol. 15, pp. 1780–1786, 2011, doi: 10.1016/j.proeng.2011.08.331.

[37] X. Li *et al.*, "Digital Health: Tracking Physiomes and Activity Using Wearable Biosensors Reveals Useful Health-Related Information," *PLOS Biol.*, vol. 15, no. 1, p. e2001402, 2017, doi: 10.1371/journal.pbio.2001402.

[38] C. Catal, S. Tufekci, E. Pirmit, and G. Kocabag, "On the use of ensemble of classifiers for accelerometer-based activity recognition," *Appl. Soft Comput. J.*, vol. 37, pp. 1018–1022, 2015, doi: 10.1016/j.asoc.2015.01.025.

[39] Z. Sedighi Maman, M. A. Alamdar Yazdi, L. A. Cavuoto, and F. M. Megahed, "A data-driven approach to modeling physical fatigue in the workplace using wearable sensors," *Appl. Ergon.*, vol. 65, pp. 515–529, 2017, doi: 10.1016/j.apergo.2017.02.001.

[40] J. S. Mollee, A. Middelweerd, S. J. Te Velde, and M. C. A. Klein, "Evaluation of a personalized coaching system for physical activity: user appreciation and adherence," Accessed: Jan. 10, 2018. [Online].

Available: https://research.vu.nl/ws/files/39950778/ITCH_v18_Revision.pdf.

[41] M. Gerdes, S. Martinez, and D. Tjondronegoro, "Conceptualization of a Personalized eCoach for Wellness Promotion," 2017.

[42] H. op den Akker, V. M. Jones, and H. J. Hermens, "Tailoring real-time physical activity coaching systems: a literature survey and model," *User Model. User-adapt. Interact.*, vol. 24, no. 5, pp. 351–392, Dec. 2014, doi: 10.1007/s11257-014-9146-y.

[43] M. W. van; Ittersum, H. K. E. . Oldenhuis, H. . Velthuijsen, and M. de. Groot, "Self-Tracking-Supported Health Promotion: A Randomized Trial among Dutch Employees.," *Eur J Public Heal.*, vol. under revi, 2017.

[44] D. H. Wolpert, "The Lack of A Priori Distinctions Between Learning Algorithms," *Neural Comput.*, vol. 8, no. 7, pp. 1341–1390, 1996, doi: 10.1162/neco.1996.8.7.1341.

[45] S. Raschka and V. Mirjalili, *Python Machine Learning*. Packt Publishing Ltd, 2015.

[46] "scikit learn,choosing the right estimator." http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html (accessed Nov. 01, 2017).

[47] "Machine learning algorithm cheat sheet for Microsoft Azure Machine Learning Studio." https://docs.microsoft.com/en-us/azure/machine-learning/studio/algorithm-cheat-sheet (accessed Nov. 01, 2017).

[48] "scikit-learn v0.18." http://scikit-learn.org/0.18/documentation.html (accessed Nov. 01, 2017).

[49] "Anaconda." www.anaconda.com (accessed Nov. 01, 2017).

[50] "Jupyter Notebooks." https://jupyter.org/ (accessed Nov. 01, 2017).

[51] "Oracle Express Edition 11g2." http://www.oracle.com/technetwork/database/database-technologies/express-edition/overview/index.html (accessed Nov. 01, 2017).

[52] "Flask." http://flask.pocoo.org/ (accessed Nov. 01, 2017).

[53] C. B. Chan, D. A. Ryan, and C. Tudor-Locke, "Relationship between objective measures of physical activity and weather: a longitudinal study," *Int. J. Behav. Nutr. Phys. Act.*, vol. 3, no. 1, p. 21, 2006, doi: 10.1186/1479-5868-3-21.

[54] B. Fogg, "A behavior model for persuasive design," in *Proceedings of the 4th International Conference on Persuasive Technology - Persuasive '09*, 2009, p. 1, doi: 10.1145/1541948.1541999.

[55] J. Blok, A. Dol, and T. Dijkhuis, "Toward a Generic Personalized Virtual Coach for Self-management : a Proposal for an Architecture," no. c, pp. 105–108, 2017.

CHAPTER 3

Early prediction of physical performance in elite soccer matches- a machine learning approach to support substitutions

Based on
"Early prediction of physical performance in elite soccer matches- a Machine Learning approach to support substitutions"

Talko B. Dijkhuis
Matthias Kempe,
Koen A. P. M. Lemmink

2021. Entropy 23(8):1–19.
doi: 10.3390/e23080952.

ABSTRACT

Substitution is an essential tool of a coach to influence the match. Factors like an injury of a player, required tactical changes, or underperformance of a player initiates substitutions. This study aims to predict the physical performance of individual players in an early phase of the match to provide additional information to the coach for his decision on substitutions. Tracking data of individual players, except for goalkeepers, from 302 elite soccer matches of the Dutch 'Eredivisie' 2018-2019 season were used to enable the prediction of the individual physical performance. The players' physical performance is expressed in the variables distance covered, distance in speed category, and energy expenditure in power category. The individualized normalized variables were used to build machine learning models that predict whether players will achieve 100%, 95%, and 90% of their average physical performance in a match. The tree-based algorithms Random Forest and Decision Tree were applied to build the models. A simple Naïve Bayes algorithm was used as the baseline model to support the superiority of the tree-based algorithms. The machine learning technique Random Forest combined with the variable energy expenditure in power category was the most precise. The combination of Random Forest and energy expenditure in power category resulted in precision in predicting performance and underperformance after 15 minutes in a match were 0.91, 0.88, and 0.92 for the thresholds 100%, 95%, and 90%, respectively. To conclude, it is possible to predict the physical performance of individual players in an early phase of the match. These findings offer opportunities to support the coaches in making more informed decisions on players' substitutions in elite soccer.

Keywords
Fatigue; decision-support; football.

1. INTRODUCTION

Soccer is a highly competitive and physically demanding sport. The physical demand is highlighted by an increase in ball (game) speed by 15% over the last 50 years [1]. A cohesive body of research points out that players' fatigue leads to a decline in their running activities. For instance, in a team participating in the Australian national soccer league, total distance, average speed, high-intensity running distance, and very high-intensity running distance decreased significantly from the first to the second half by 7.92, 9.47, 10.10, and 10.99%, respectively [2]. **In similar fashion, in the Italian A series, a team showed a significant reduction between the first and second half in high-intensity running distance** (-14.9%) [3]. These examples highlight that, players are unable to perform maximally throughout a match [4]. Information on this drop in performance is essential for players and coaches. A recent study showed that running performance parameters (e.g., the number of accelerations or decelerations and the distance covered in different speed categories) affect successful performance soccer for some playing positions [5]. As most soccer matches are often decided by just one goal [6], a drop in physical performance can make the difference between winning and losing. Therefore, teams and coaches need to identify players that physically underperform in a match as early as possible to adapt their style of play or substitute these players. In general, an injury of a player, necessary tactical changes, or underperformance of a player causes substitutions (for an overview, see Hills et al., 2018) [7]. Substitution may be the most powerful tool of a coach to influence the match. Substitutions can minimize or offset the effects of fatigue of the team as substitutes cover more distance and perform more high-intensity actions relative to entire match players [8]. According to the Fédération Internationale de Football Association (FIFA) SARS-COVID-19 2020 rules, a coach has five substitution options in a match, implicating fitness of the individual player and physical performance has more impact on substitution than before COVID-19 [9].

To identify a physically underperforming player, coaches can base their decision on real-time motion data. To record and monitor real-time motion data, multi-camera position tracking systems such as SportVU and TRACAB® system are most commonly used in professional leagues [10]. However, one has to constantly monitor and analyse several physical variables of all eleven players. As highlighted in a survey paper by Nosek, Brownlee, Drust, & Andrew (2020) [11], staff and IT solutions struggle with giving helpful feedback to the coach after training sessions due to the amount and complexity of the data and its often inconclusive communication [11]. In order to enable helpful, timely feedback, Robertson advocated using machine learning approaches decision support for the coach [12]. Decision support provides superior efficacy when the volume of the data is large, and the data is complex [13]. An in-match physical performance prediction

and decision support using machine learning is a novelty that has not yet been realized for team sports.

In order to build an in-match physical performance prediction and decision support, models have to be based on derived time-motion data variables. These derived time-motion variables can be divided into type-1 or type-2 [14]. Type-1 variables include external load measures such as distance covered and distance covered in speed category. Type-2 variables include load measures related to changes in velocity such as accelerations, decelerations, and summarized variables like metabolic power and energy expenditure. Researchers have tried to quantify physical performance decline as a decrease in various type-1 variables. It turns out that during the match, the distance covered, and the distance in the speed category decreases [2], [3], [15]. However, type-2 variables are more sensitive to identifying in-match physical performance decline than type-1 variables [14]. Furthermore, condense variables like metabolic power are specially equipped for identifying in-match performance decline. They hold a more linear relationship with fatigue and include accelerations and decelerations in their calculation [16]. These findings highlight the sensitivity of type-2 variables for physical performance decline. Therefore, we include both the more common type-1 and the more sensitive type-2 variables in our prediction models. Contextual factors like home or away, rank-position, and score show a difference in the overall distance covered [17]. Although we acknowledge the contextual factors such as home or away, rank position and score, we excluded these contextual factors in this proof-of-concept study. Instead, we focused on the individual player in-match motion data.

The study's goal is to predict the in-match physical performance decline of the individual soccer player using machine learning. To our knowledge, no prior study in professional soccer has investigated the in-match physical performance prediction using machine learning techniques enabling decision support for the coach on substitutes. We aim to prove: (1) if physical performance decline can be identified using both type-1 and type-2 variables; (2) if substitutes perform better than entire match players on both type-1 and type-2 variables and (3) if the degree of physical performance of a player can be predicted in an early stage of the match using machine learning models for type-1 and type-2 variables.

2. MATERIALS AND METHODS

2.1 Experimental Approach to the problem

For our study, we retrospectively collected the in-match position tracking data from 302 competitive professional soccer matches between 18 teams during the Dutch 'Eredivisie'

2018-2019 season. For our analysis, two matches with erroneous and missing data were excluded. Also, the extra time at the end of the first and second half and goalkeepers were excluded from the dataset. The effect of substitution on the match was controlled by identifying both entire match players and substitutes. Thus, entire match players played the full match where the substitutes entered the match at a later stage.

2.2 Subjects

Four hundred and eighty players participated in the 300 matches. Four thousand nine hundred thirty-five times, entire match players were identified. In addition, 1533 substitutes were identified. The majority of substitutions happened at half-time (50-minute mark) and between the 60- and 90-minute marks (Figure 2). The number of substitutions in the first half and the 55-minute mark is significantly lower ($P < 0.001$) compared to the second half and between the 60- and 90-minute marks. The Ethics Committee CTc UMCG, of the University Medical Center Groningen, The Netherlands, approved the study, approval number: 201800430.

2.3 Data

The sample includes tracking data of all players of 302 matches. The players' time, position, speed, and acceleration were detected and recorded by SportsVU optical tracking system (SportsVU, STATS LLC, Chicago, IL, USA). Linke et al. (2018) tested the SportsVU optical tracking system and rated the system as being adequately reliable [18].

2.4 Variables

The type-1 variables distance covered, distance in speed category, and the type-2 variable energy expenditure in power category were applied to examine the decline in physical performance [14]. The variables were calculated as (i) distance covered per five minutes, 15 minutes, half and entire match [15], [19], [20] (ii) distance in speed category per five minutes, 15 minutes, half, and entire match: the speed categories were categorized as Very Low Intensity Running (VLIR; 0.7–7.2 km·h-1), Low Intensity Running (LIR; 7.2–14.4 km·h-1), Medium Intensity Running (MIR; 14.4–19.8 km·h-1), High Intensity Running (HIR; 19.8–25.1 km·h-1), and Very High Intensity Running (VHIR; >25.2 km·h-1) [15], [21] (iii) energy expenditure in power category per five minutes,15 minutes, half and entire match, calculated conform Osgnach et al. [22]. The power categories were categorized as Low Power (LP; from 0 to 10 W*kg-1), Intermediate Power (IP; from 10 to 20 W*kg-1), High Power (HP; from 20 to 35 W*kg-1), Elevated Power (EP; from 35 to 55 W*kg-1), and Maximal Power (MP; >55 W*kg-1) [22]. The descriptive statistics of the variables were calculated for entire match players and substitutes and reported as mean ± standard deviation for each variable. The difference between entire match players and substitutes was reported for all variables as well.

2.5 Statistical Analysis

For the statistical analysis, we used the statsmodels package 0.11.1 in Python 3.7.2. The statistical analysis was performed for the variables distance covered, distance covered in speed category, and energy expenditure in power category. First, the normality of the variables was checked for entire match players for the first half, the second half, and the 15-minute periods of both halves. The normality of the variables was checked for substitutes in the second half and 15-minute periods in the second half. The Kolmogorov-Smirnov test determined the normality of the variables. No normal distribution was found for both entire match players and substitutes in the variables (i) the distance covered ($P < 0.001$), (ii) the distance covered in speed category in all speed categories ($P < 0.001$), (iii) the energy expenditure in power category in all power categories ($P < 0.001$). Kruskal-Wallis test evaluated the differences between the different periods and variables. There were significant differences between every period and variable ($P < 0.001$) for both entire match players and substitutes. As a measure of effect size, epsilon squared ($\square^2$) was calculated for the Kruskal-Wallis test, and values from 0 to 1 indicate no relationship to a perfect relationship, respectively [23]. In the event of a significant difference, Conover post-hoc tests were used to identify any localized effects. The variable pairwise comparisons were used to reject the null hypothesis ($P < 0.01$). Statistical significance was set at $p< 0.05$.

The source code, access to the data, and corresponding Jupiter notebooks of the statistics procedure are available as open-source software on Github (https://github.com/dijkhuist/Early-Performance-Prediction-Machine-Learning-in-Soccer).

2.6 Machine learning

To predict the physical performance of individual players machine learning models were constructed for each variable distance covered, distance covered in speed category, and energy expenditure in power category. The physical performance differences between players were eliminated by individualization and normalization of the variables and outcome measures. Variables were calculated per five-minute period of the match. The performance in the current match was compared to the average individual performance of a player over the whole season. In other words, the mean value of the performance variable over the entire season based on all entire matches by an individual player was set as a personal baseline. We further calculated these baseline values for each of the 18 5-minute periods of a match. Given this approach, we could calculate a relative individual performance for each player. All constructed features are presented in Table 1.

Table 1. Variable based constructed features.

Variable: Distance Covered	
Feature	Explanation
Period	The five-minute period indicator of the match (range: 1-18)
Percentage average distance	The percentage of distance covered versus the average percentage of distance covered in the specific five-minute period
Percentage average percentage summed distance	The percentage of the summed distance covered versus the average percentage summed distance covered up to and including the specific the five-minute period
Variable: Distance in Speed Category	
Feature	Explanation
Period	The five-minute period indicator of the match (range: 1-18)
Percentage very low intensity running distance covered versus average percentage very low intensity running distance covered	The percentage of the distance covered in the very low intensity running speed category versus the average percentage distance in the very low intensity running speed category in the specific five-minute period.
Percentage summed very low intensity running distance covered versus average very low intensity running distance covered	The percentage of the summed distance covered in the very low intensity running speed category versus the average percentage summed distance in the very low intensity running speed category in the specific five-minute period.
Percentage low intensity running distance covered versus average percentage low intensity running distance covered	The percentage of the distance covered in the very low intensity running speed category versus the average percentage distance in the very low intensity running speed category in the specific five-minute period.
Percentage summed low intensity running distance covered versus average percentage summed Low intensity running distance covered	The percentage of the summed distance covered in the low intensity running speed category versus the average percentage summed distance covered in the low intensity running category up to and including the specific five-minute period.
Percentage medium intensity running distance covered versus average percentage medium intensity running distance covered	The percentage of the distance covered in the medium intensity running speed category versus the average percentage distance covered in the medium intensity running speed category in the specific five-minute period
Percentage summed medium intensity running distance covered versus average medium intensity running distance covered	The percentage of the summed distance covered in the medium intensity running speed category versus the average percentage summed distance covered in the medium intensity running speed category up to and including the specific five-minute period.
Percentage high intensity running distance covered versus average percentage high intensity running distance covered	The percentage of the distance covered in the high intensity running speed category versus the average percentage distance covered in the high intensity running speed category in the specific five-minute period.
Percentage summed high intensity running distance covered versus average percentage summed high intensity running distance covered	The percentage of the summed distance covered in the high intensity running speed category versus the average percentage summed distance covered in the high intensity running speed category up to and including the specific five-minute period.

Table 1. Continued.

Variable: Distance in Speed Category	
Feature	Explanation
Percentage very high intensity running distance covered versus average percentage very high intensity running distance covered	The percentage of the summed distance covered in the very high intensity running speed category versus the average percentage summed distance covered in the very high intensity running speed category in the specific five-minute period.
Percentage summed very high intensity running distance covered versus average percentage very high intensity running distance covered	The percentage of the summed distance covered in the very high intensity running speed category versus the average percentage summed distance covered in the very high intensity running speed category up to and including the specific five-minute period.
Variable: Energy Expenditure in Power Category	
Feature	Explanation
Period	The five-minute period indicator of the match (values: 1-18)
Percentage low power energy expenditure versus average low power energy expenditure	The percentage of the energy expenditure in the low power category versus the average percentage energy expenditure in the low power category in the specific five-minute period.
Percentage summed low power energy expenditure versus average percentage summed low power energy expenditure	The percentage of the summed energy expenditure in the low power category versus the average percentage summed energy expenditure in the low power category up to and including the specific five-minute period.
Percentage intermediate power energy expenditure versus average percentage low power energy expenditure	The percentage of the energy expenditure in the intermediate power category versus the average percentage energy expenditure in the low power category in the specific five-minute period.
Percentage summed intermediate power energy expenditure versus average percentage summed intermediate power energy expenditure	The percentage of the summed energy expenditure in the low power category versus the average percentage summed energy expenditure in the intermediate power category up to and including the specific five-minute period.
Percentage high power energy expenditure versus average percentage high power energy expenditure	The percentage of the energy expenditure in the high-power category versus the average percentage energy expenditure in the high power category in the specific five-minute period.
Percentage summed high power energy expenditure versus average percentage summed high power energy expenditure	The percentage of the summed energy expenditure in the high-power category versus the average percentage summed energy expenditure in the high power category up to and including the specific five-minute period.
Percentage elevated power energy expenditure versus average elevated power energy expenditure	The percentage of the energy expenditure in the elevated power category versus the average percentage energy expenditure in the elevated power category in the specific five-minute period.
Percentage summed elevated power energy expenditure versus average summed elevated power energy expenditure	The percentage of the summed energy expenditure in the elevated power category versus the average percentage summed energy expenditure in the elevated power category up to and including the specific five-minute period.

Table 1. Continued.

Variable: Energy Expenditure in Power Category	
Feature	Explanation
Percentage maximal power energy expenditure versus average percentage maximal power energy expenditure	The percentage of the energy expenditure in the maximal power category versus the average percentage energy expenditure in the maximal power category in the specific five-minute period.
Percentage summed maximal power energy expenditure versus average percentage summed maximal power energy expenditure	The percentage of the summed energy expenditure in the maximal power category versus the average percentage summed energy expenditure in the maximal power category up to and including the specific five-minute period.

To predict the underperformance of a player during the match, the underperformance was classified as not achieving 100%, 95%, or 90% of the entire season average of the individual player. The outcome measures were: distance (m) (for distance covered and distance in speed category) and energy expenditure (kJ kg-1) (for energy expenditure in power category). The machine learning process is visualized in Figure 1. The tracking data was used to calculate physical performance variables per individual player, as described before, and labelled as underperforming or not. After that, the data set was split into a 70% training set and a 30% test set. Subsequently, the training set was resampled to have an equal division of performing and underperforming labels using the SMOTE method [24]. Machine learning models were generated using the learning algorithms, and the test set was applied to identify the physical performance of the individual player.

Since there is no linear relation in physical performance during the soccer match, tree-based algorithms like the Random Forest algorithm and the Decision Tree algorithm were applied. Conducting the machine learning models was combined with parameter tuning, randomized search, and cross-validation [25]. A simple Naïve Bayes classifier was used as the baseline model to highlight the validity of the tree-based algorithms. As it is common practice for evaluating machine learning approaches, Random Forest and Decision Tree should outperform the simple Naïve Bayes baseline classifier. The following overall performance measures were calculated for each model: accuracy, precision, recall, F1-score, and Area under the curve (AUC). The scikit-learn package 0.23.1 in Python 3.7.2 was used to construct and judge the machine learning models' performance. The source code, access to the data, and corresponding Jupiter notebooks of the machine learning procedure is available as open-source software on Github (https://github.com/dijkhuist/Early-Performance-Prediction-Machine-Learning-in-Soccer).

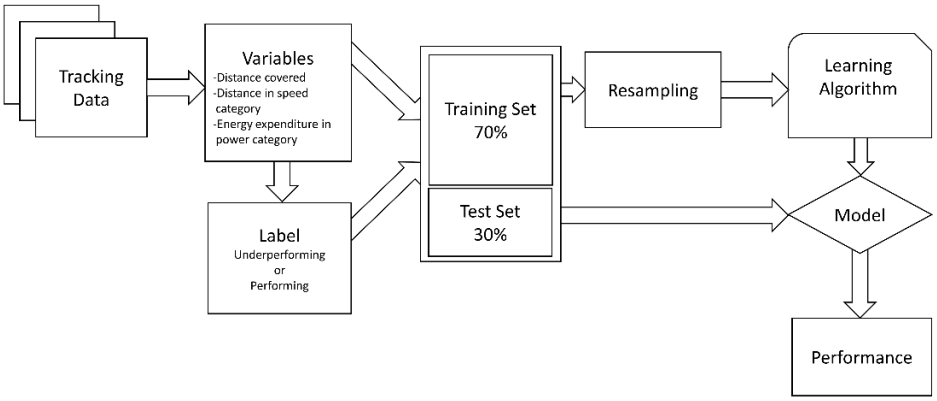


Figure 1. Visualization of the machine learning process.

3. RESULTS

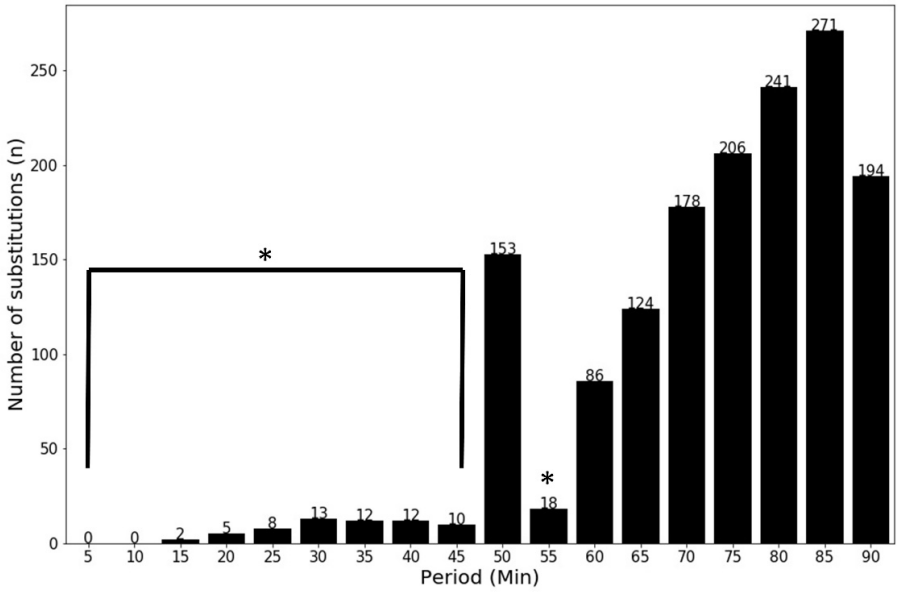


Figure 2. Number of substitutes per 5-minute period.
*Significantly lower number of substitutes ($P < 0.001$).

3.1 Physical performance

3.1.1 Entire match players

In general, the physical performance of players participating in the entire match declined

throughout the match. The visualization of the distance covered is represented in Figure 3. The average distance covered declined over time from 5275 ± 223m in the first half to 4906 ± 225m in the second half ($P < 0.001$, $\varepsilon^2 = 0.43$).

The visualization of the distance in speed category of the entire match players can be found in Figure 4. The distance covered in speed category showed a decline in the average distance in the speed category LIR from 2345 ± 170m in the first half to 2092 ± 148m in the second half ($P < 0.001$, $\varepsilon^2 = 0.43$), MIR from 888 ± 87m in the first half to 792 ± 77m in the second half ($P < 0.001$, $\varepsilon^2 = 0.26$), and HIR from 290 ± 38m in the first half to 268 ± 35m in the second half ($P < 0.001$, $\varepsilon^2 = 0.08$). VLIR increased during the second half from 564 ± 66m in the first 15 minutes to 603 ± 69m in the last 15 minutes ($P < 0.001$, $\varepsilon^2 = 0.05$), and the distance covered in the speed categories HIR and VHIR almost stay stable during the entire match.

The descriptives of the distance in power category of the entire match players are visualized in Figure 5. The energy expenditure in power category shows a decline in the average energy expenditure in the power categories IP from 7.26 ± 1.24kJ kg-1 in the first half to 6.47 ± 1.14kJ kg-1 in the second half ($P < 0.001$, $\varepsilon^2 = 0.75$), HP from 3.52 ± 0.76kJ kg-1 in the first half to 3.13 ± 0.67kJ kg-1 in the second half ($P < 0.001$, $\varepsilon^2 = 0.70$) and EP from 1.47 ± 0.44kJ kg-1 in the first half to 1.32 ± 0.40kJ kg-1 the second half ($P < 0.001$, $\varepsilon^2 = 0.75$). Where energy expenditure in the power categories LP and MP almost stayed stable during the match.

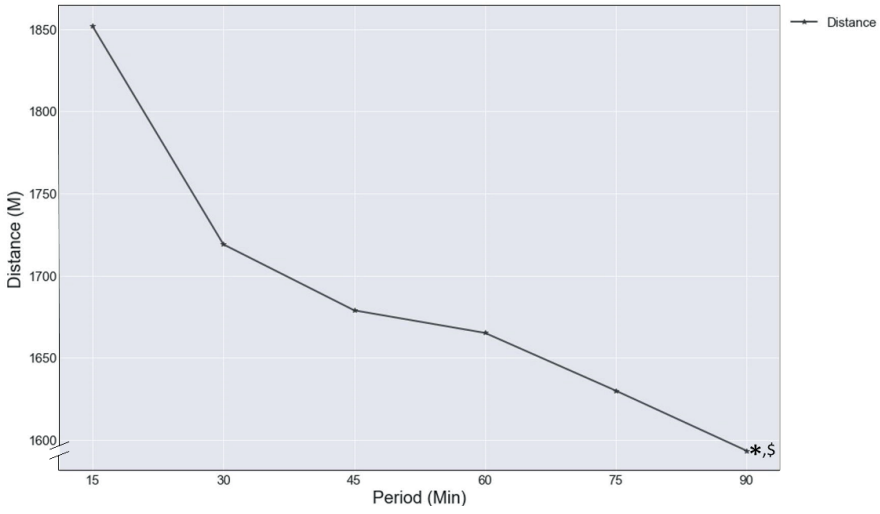


Figure 3. Average distance covered entire match players in 15-minute periods.
* ($P < 0.01$) a significant decline between the first half (15-45Min) and the second half (60-90Min).
$ ($P < 0.01$) a significant decline between the 15-minute periods in the second half (60-90Min).

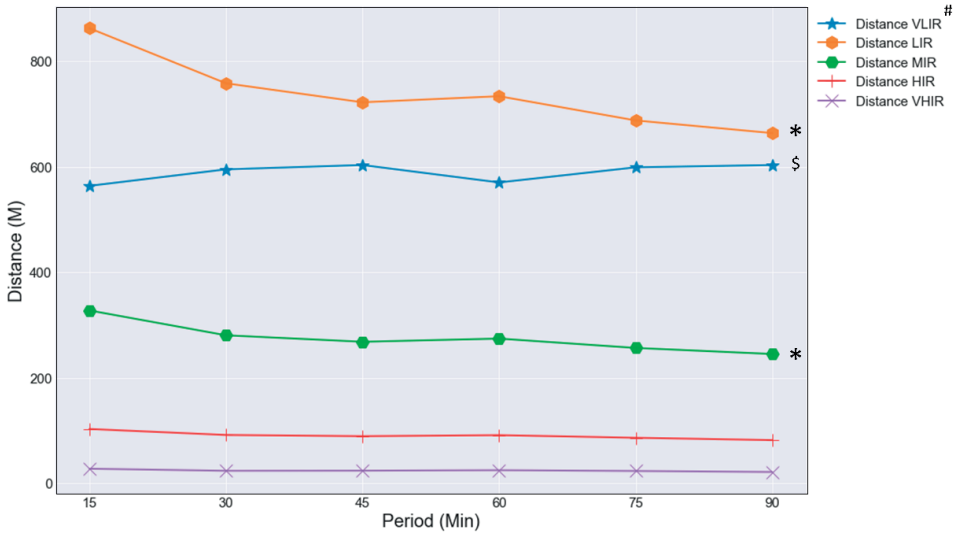


Figure 4. Distance covered in speed category entire match in 15-minute periods. # Abbreviations of the power categories VLIR, Very low Intensity Running; LIR, Low Intensity Running; MIR, Medium Intensity Running; HIR, High Intensity Running; VHIR, Very High Intensity Running.
* (P < 0.01) a significant decline between the first half (15-45Min) and the second half (60-90Min).
$ (P < 0.01) a significant increase between the 15-minute periods in the second half (60-90Min).

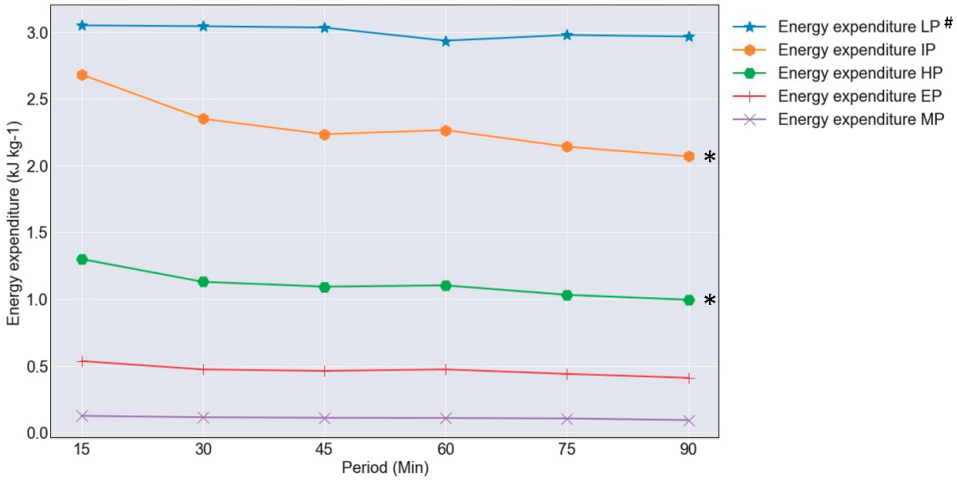


Figure 5. Energy expenditure in power category entire match players in 15-minute periods. # Abbreviations of the power categories: LP, Low Power; IP, Intermediate Power; HP, High Power; EP, Elevated Power; MP, Maximum Power.
* (P < 0.01) a significant decline between the first half (15-45Min) and the second half (60-90Min).

3.1.2 Entire match players versus substitutes

The average total distance covered by substitutes is higher than the average total distance covered by entire match players: 5123 ± 397m versus 4906 ± 225m ($P < 0.001$, $\varepsilon^2 = 0.12$) in the second half. In addition, there was a significant difference in distance covered between substitutes and the entire match between the 60-75 minutes ($P < 0.001$, $\varepsilon^2 = 0.19$) and 75-90 minutes of the match ($P < 0.001$, $\varepsilon^2 = 0.12$). (Figure 6)

The distance covered in speed category for VLIR ($P < 0.001$, $\varepsilon^2 = 0.32$), LIR ($P < 0.001$, , $\varepsilon^2 = 0.75$), and MIR ($P < 0.001$, $\varepsilon^2 = 0.23$) of the substitutes in the second half is higher for the entire match players in the second half. Furthermore, distance covered in speed category showed a decline for the entire match players in the speed categories MIR, HIR, and VHIR in the second half, while there was no such decline for substitutes.

The energy expenditure of the substitutes was higher in the second 15-minute period (7.11 ± 0.86kJ kg-1 vs. 6.69 ± 0.71) ($P = 0.007$, $\varepsilon^2 = 0.002$) and the last 15-minute period (7.31 ± 0.72 vs. 6.53 ± 0.72) ($P < 0.001$, $\varepsilon^2 = 0.02$) of the second half compared to the entire match players. In contrast to the substitutes, entire match players showed a decline in energy expenditure over the three 15-minute periods in the second half ($P < 0.001$, $\varepsilon^2 = 0.06$).

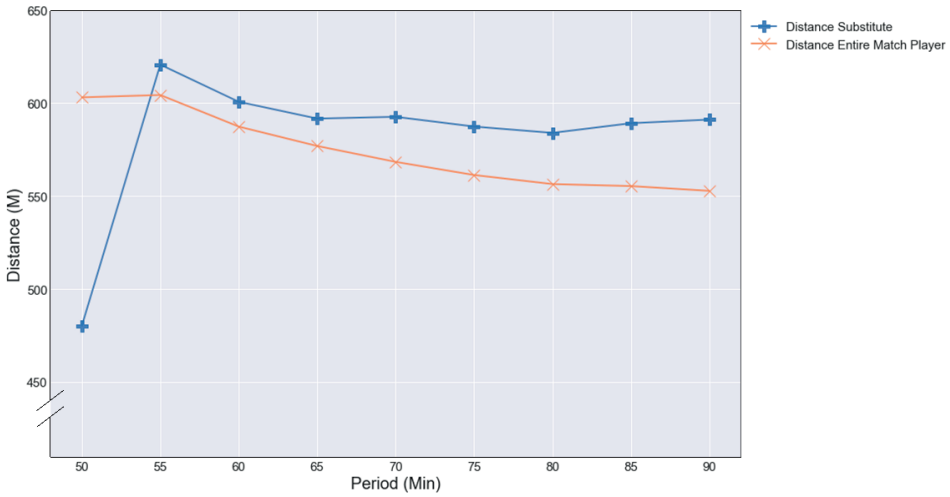


Figure 6. Average distance entire match players versus substitutes second half.

3.2 Machine learning

The three prediction models for the three different thresholds of 100%, 95%, and 90% of a player's average physical match performance showed differences in accuracy and f1 scores for both tree-based and baseline models. These differences were primarily due to the reduced number of underperformers in the 90% category. While the split between over- and underperformers is 50% for the 100% thresholds, the number of underperformers decreases to 1% for the 90% thresholds (Table 2). This naturally favours the correct prediction of performers and impedes the minority category (underperformers). Random Forest and Decision Tree outperformed Naïve Bayes in precision and recall for all three variables (distance covered, distance covered in speed category, energy expenditure in power category) and thresholds (Table 3). Overall, the Random Forrest approach showed the best performance for all variables. Comparing the three different variables, energy expenditure in power category showed the best score on precision in every threshold and therefore provided the best prediction models.

Overall, the precision of classifying underperforming players was increasing during the match. After 15 minutes applying either Random Forest or Decision Tree distance in speed category and energy expenditure in power category showed a precision of respectively 0.91, 0.88, and 0.92 for the thresholds 100%, 95%, and 90%. The baseline model Naïve Bayes was less precise than Decision Tree and Random Forest (Figure 7).

Table 2. Variable distribution of the performing and underperforming players.

Variables	Threshold 100%	Threshold 95%	Threshold 90%
Distance Covered			
Underperforming (n)	38490	60820	68347
Performing (n)	30590	8260	733
Distance In Speed Category Model			
Underperforming (n)	42014	64340	69520
Performing (n)	27866	5540	360
Energy Expenditure In Power Category			
Underperforming (n)	34416	7912	1604
Performing (n)	35463	61967	68275

Table 3. Machine learning metrics.

Variable: Distance Covered							
Threshold	Algorithm	Accuracy	AUC		Precision	Recall	F1-score
100%	Random Forest	0.90	0.94	Underperforming	0.97	0.92	0.94
				Performing	0.70	0.84	0.76
	Decision Tree	0.88	0.86	Underperforming	0.96	0.90	0.93
				Performing	0.64	0.82	0.72
	Naïve Bayes	0.57	0.58	Underperforming	0.83	0.59	0.69
				Performing	0.21	0.48	0.29
95%	Random Forest	0.75	0.79	Underperforming	0.71	0.67	0.69
				Performing	0.78	0.81	0.79
	Decision Tree	0.77	0.82	Underperforming	0.73	0.72	0.72
				Performing	0.80	0.81	0.81
	Naïve Bayes	0.73	0.75	Underperforming	0.68	0.67	0.67
				Performing	0.77	0.78	0.77
90%	Random Forest	0.93	0.95	Underperforming	0.55	0.87	0.67
				Performing	0.99	0.94	0.96
	Decision Tree	0.92	0.88	Underperforming	0.51	0.85	0.63
				Performing	0.99	0.93	0.96
	Naïve Bayes	0.74	0.74	Underperforming	0.15	0.52	0.24
				Performing	0.95	0.92	0.84
Variable: Distance In Speed Category Model							
Threshold	Algorithm	Accuracy	AUC		Precision	Recall	F1-score
100%	Random Forest	0.89	0.96	Underperforming	0.85	0.87	0.86
				Performing	0.91	0.90	0.91
	Decision Tree	0.74	0.81	Underperforming	0.65	0.73	0.68
				Performing	0.81	0.75	0.78
	Naïve Bayes	0.70	0.78	Underperforming	0.59	0.77	0.67
				Performing	0.82	0.65	0.72
95%	Random Forest	0.96	0.98	Underperforming	0.68	0.91	0.78
				Performing	0.99	0.96	0.98
	Decision Tree	0.94	0.92	Underperforming	0.59	0.89	0.71
				Performing	0.99	0.95	0.97
	Naïve Bayes	0.97	0.83	Performing	0.23	0.72	0.35
				Performing	0.97	0.80	0.87
90%	Random Forest	1.00	0.99	Underperforming	0.61	0.94	0.74
				Performing	1.00	1.00	1.00
	Decision Tree	0.99	0.97	Underperforming	0.41	0.94	0.57
				Performing	1.00	0.99	1.00
	Naïve Bayes	0.88	0.89	Underperforming	0.03	0.74	0.06
				Performing	1.00	0.88	0.93
Variable: Energy Expenditure In Power Category							
Threshold	Algorithm	Accuracy	AUC		Precision	Recall	F1-score
100%	Random Forest	0.89	0.96	Underperforming	0.88	0.89	0.89
				Performing	0.89	0.89	0.89

Table 3. Continued.

Variable: Energy Expenditure In Power Category							
Threshold	Algorithm	Accuracy	AUC		Precision	Recall	F1-score
	Decision Tree	0.82	0.92	Underperforming	0.82	0.81	0.82
				Performing	0.82	0.82	0.82
	Naïve Bayes	0.81	0.90	Underperforming	0.81	0.81	0.81
				Performing	0.80	0.80	0.80
95%	Random Forest	0.97	0.99	Underperforming	0.83	0.91	0.87
				Performing	0.99	0.98	0.98
	Decision Tree	0.96	0.98	Underperforming	0.74	0.85	0.79
				Performing	0.98	0.97	0.98
	Naïve Bayes	0.90	0.87	Underperforming	0.36	0.5	0.09
				Performing	0.90	0.99	0.94
90%	Random Forest	1.00	0.99	Underperforming	0.88	0.86	0.87
				Performing	1.00	1.00	1.00
	Decision Tree	0.96	0.98	Underperforming	0.74	0.85	0.79
				Performing	0.98	0.97	0.98
	Naïve Bayes	0.99	0.51	Underperforming	0.03	0.02	0.03
				Performing	0.99	0.99	0.99

AUC = Area Under Curve

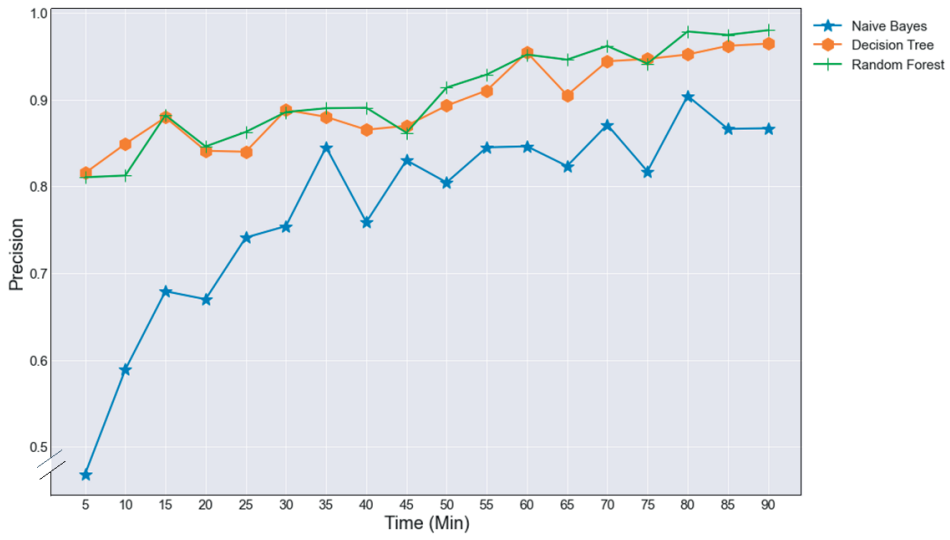


Figure 7. Random forest precision underperforming energy expenditure at 95% threshold in 5-minute periods.

4. DISCUSSION

The main goal of this study was to explore the possibility of predicting on physical performance of individual players and decision support for coaches to help them make an informed decision on player substitutions. Our study focused on a player's physical performance within the match, making the identification of underperforming players the critical point. In line with previous research, this study revealed that entire match players show a significant decline in physical performance during the match in distance covered, distance covered in speed category, and energy expenditure in power category [4], [7]. While earlier studies found a decline of a 10-15% reduction of the HIR and VHIR from the first to the second half [2], [3], our results did not show any decline in these high-intensity type-1 variables. Thereby, our findings are in consent with more recent studies [26], [27]. Furthermore, our results replicate the study of Liu et al.[26], who found that time spent in the very low intensity (VLIR) category is increasing while time in medium intensity categories is decreasing (LIR and MIR) and time in high-intensity categories are stable throughout a match. The same pattern can be seen for the energy expenditure in different power categories. Given these results, we can support our first hypothesis that type-1 and type-2 load variables can identify decreasing player performance throughout a match.

In order to answer our second research question, we found that substitutes perform better than entire match players on both type-1 and type-2 variables. Most of the substitutions occur at halftime and during the 60-90 minute period, which aligns with previous research [27]. In agreement with the literature, substitutes who had been introduced during the second half covered more distance and performed more high-intensity activities relative to entire match players over the same period [8]. In addition, second-half substitutes spent more energy in higher power categories [28]. As substitutes demonstrate higher values in physical performance variables than the entire match players, the substitution of underperformers may improve the team's performance and make the difference between winning and losing [5]. This study's machine learning models can identify and predict a players' physical performance in an early stage of the match. The Random Forest model outperformed both the Decision Tree and Naïve Bayes algorithm. For every threshold, the Random Forest model identified the underperformers and performers best. The precision of the variable energy expenditure in power category outperformed models based on the variables distance covered and distance covered in speed category. The outperformance of the variable energy expenditure in power category illustrates that the more advanced type-2 variable is most sensitive to recognizing a player's physical performance in an early stage of the match. The stronger the relation in reality between the variable and the outcome, the higher the precision of the machine learning model may be expected [29].

Following these arguments, the main finding of our study is that our machine learning models could reliably identify and predict the physical performance of a player after 15 minutes in the match. The early prediction of physical performance can support a decision support system as advocated by Robertson [12] and further illustrates the opportunities provided by machine learning in player monitoring during the match.

A limitation of the study is the exclusion of contextual factors like home or away, rank-position, position system, and score show a difference in the overall distance covered [17]. Although these contextual factors on their own influence the overall distance of the team. To generate a machine learning model on individual physical performance, every combination of the contextual factors needs to be sufficiently present in the data. Not every combination of an individual player, home or away, rank-position, position system, and score will be present in one season. A coach will need to use his or her insight and knowledge to judge the prediction of physical underperformance on its merits. The use of a machine learning approach also goes in hand with some limitations. To conduct a reliable model for an individual player, there must be enough entire match data available. We did not identify any literature in soccer to refer to the amount of data needing to be available. In the literature on fitness trackers, it is found that three days of repeated measures is necessary to represent adults' normal activity levels with an 80% confidence [30]. In parallel, three entire matches for a player may be sufficient to identify his average physical performance. A method to conduct a reliable model is to retrain models frequently and monitor precision to identify the optimum amount of data [31].

Another limitation is that physical underperformance is just one of several reasons for a coach to substitute a player. Substitutions can also be initiated by a player's injury, necessary tactical changes (e.g., because of being behind in a match), or tactical underperformance of a player [7]. In our study, the data was limited to the individual player's speed, acceleration, and distance measures. Next to contextual influences [17], other physiological markers of fatigue such as individual measures like heart rate, breathing, and body temperature were not included. Including contextual influences and physiological markers of fatigue in the machine learning model could enable a more informative system. Finally, the thresholds of physical underperformance were randomly chosen, and the 90% threshold is relatively rarely seen.

5. CONCLUSIONS

Our study confirmed that the identification of the physical performance could be based on type-1 and type-2 variables calculated from the position tracking systems. Also, substitutes perform better than entire match players on both type-1 and type-2

variables. The appliance of machine learning enables the prediction of a player's physical performance in an early stage in the match whereby the more sensitive type-2 variable outperforms the type-1 variables in the precision of the prediction.

5.1 Practical Implications

These findings show that it is possible to identify underperforming players in an early stage in the match. Applying machine learning in combination with monitoring the energy expenditure in power category during the match enables real-time support for the coach to decide on substitutions. For the nature of the game is the same for many leagues, monitoring expenditure in power category can be of use in many other environments than Dutch elite soccer. A precondition for the support system is to set up a dataset per player, which allows for tracking during the season and machine learning. Future research to refine the machine learning models may include the influence of contextual factors such as home-away, score, ranking, and player position.

Acknowledgments

We like to thank J van Norel, S.V.B. Vitesse, Arnhem, The Netherlands for sharing the competition data.

REFERENCES

[1] J. L. Wallace and K. I. Norton, "Evolution of World Cup soccer final games 1966-2010: Game structure, speed and play patterns," *J. Sci. Med. Sport*, vol. 17, no. 2, pp. 223–228, 2014, doi: 10.1016/j.jsams.2013.03.016.

[2] G. M. Wehbe, T. B. Hartwig, and C. S. Duncan, "MOVEMENT ANALYSIS OF AUSTRALIAN NATIONAL LEAGUE SOCCER PLAYERS USING GLOBAL POSITIONING SYSTEM TECHNOLOGY," *J. Strength Cond. Res.*, vol. 28, no. 3, pp. 834–842, 2014, doi: 10.1519/JSC.0b013e3182a35dd1.

[3] E. Rampinini, A. Bosio, I. Ferraresi, A. Petruolo, A. Morelli, and A. Sassi, "Match-related fatigue in soccer players," *Med. Sci. Sports Exerc.*, vol. 43, no. 11, pp. 2161–2170, 2011, doi: 10.1249/MSS.0b013e31821e9c5c.

[4] C. Carling, "Interpreting physical performance in professional soccer match-play: Should we be more pragmatic in our approach?," *Sport. Med.*, vol. 43, no. 8, pp. 655–663, 2013, doi: 10.1007/s40279-013-0055-8.

[5] T. Modric, S. Versic, D. Sekulic, and S. Liposek, "Analysis of the association between running performance and game performance indicators in professional soccer players," *Int. J. Environ. Res. Public Health*, vol. 16, no. 20, 2019, doi: 10.3390/ijerph16204032.

[6] M. Kempe, M. Vogelbein, and S. Nopp, "The cream of the crop: Analysing FIFA world cup 2014 and Germany's title run," *J. Hum. Sport Exerc.*, vol. 11, no. 1, pp. 42–52, 2016, doi: 10.14198/jhse.2016.111.04.

[7] S. P. Hills *et al.*, "Profiling the Responses of Soccer Substitutes: A Review of Current Literature," *Sport. Med.*, vol. 48, no. 10, pp. 2255–2269, 2018, doi: 10.1007/s40279-018-0962-9.

[8] P. S. Bradley, C. Lago-Peñas, and E. Rey, "Evaluation of the match performances of substitution players in elite soccer," *Int. J. Sports Physiol. Perform.*, vol. 9, no. 3, pp. 415–424, 2014, doi: 10.1123/IJSPP.2013-0304.

[9] FIFA, "Five substitutes option temporarily allowed for competition organisers," 2020. https://www.fifa.com/who-we-are/news/five-substitutes-option-temporarily-allowed-for-competition-organisers.

[10] E. J. Arriaza and M. D. Zuniga, "Soccer as a study case for analytic trends in collective sports training: A survey," *Int. J. Perform. Anal. Sport*, vol. 16, no. 1, pp. 171–190, 2016, doi: 10.1080/24748668.2016.11868879.

[11] P. Nosek, T. E. Brownlee, B. Drust, and M. Andrew, "Feedback of GPS training data within professional English soccer: a comparison of decision making and perceptions between coaches, players and performance staff," *Sci. Med. Footb.*, vol. 5, no. 1, pp. 35–47, 2021, doi: 10.1080/24733938.2020.1770320.

[12] P. S. Robertson, "Man & machine: Adaptive tools for the contemporary performance analyst," *J. Sports Sci.*, vol. 00, no. 00, pp. 1–9, 2020, doi: 10.1080/02640414.2020.1774143.

[13] L. Bate, A. Hutchinson, J. Underhill, and N. Maskrey, "How clinical decisions are made," *Br. J. Clin. Pharmacol.*, vol. 74, no. 4, pp. 614–620, 2012, doi: 10.1111/j.1365-2125.2012.04366.x.

[14] M. Buchheit and B. M. Simpson, "Player-Tracking Technology : Half-Full or Half-Empty Glass ?," *Int. J. Sports Physiol. Perform.*, vol. 12, no. S2, pp. 35–41, 2017.

[15] M. Mohr, P. Krustrup, and J. Bangsbo, "Match performance of high-standard soccer players with special reference to development of fatigue," *J. Sports Sci.*, vol. 21, no. 7, pp. 519–528, 2003, doi: 10.1080/0264041031000071182.

[16] P. E. di Prampero, A. Botter, and C. Osgnach, "The energy cost of sprint running and the role of metabolic power in setting top performances," *Eur. J. Appl. Physiol.*, vol. 115, no. 3, pp. 451–469, 2015, doi: 10.1007/s00421-014-3086-4.

[17] R. Aquino *et al.*, "Influence of Situational Variables, Team Formation, and Playing Position on Match Running Performance and Social Network Analysis in Brazilian Professional Soccer Players," *J. strength Cond. Res.*, vol. 34, no. 3, pp. 808–817, 2020, doi: 10.1519/JSC.0000000000002725.

[18] D. Linke, D. Link, and M. Lames, "Validation of electronic performance and tracking systems EPTS under field conditions," *PLoS One*, vol. 13, no. 7, pp. 1–19, 2018, doi: 10.1371/journal.pone.0199519.

[19] M. Mohr, P. Krustrup, and J. Bangsbo, "Fatigue in soccer: A brief review," *J. Sports Sci.*, vol. 23, no. 6, pp. 593–599, 2005, doi: 10.1080/02640410400021286.

[20] J. Bangsbo, F. M. Iaia, and P. Krustrup, "Metabolic response and fatigue in soccer.," *Int. J. Sports Physiol. Perform.*, vol. 2, no. 2, pp. 111–127, 2007, doi: 10.1123/ijspp.2.2.111.

[21] P. S. Bradley, W. Sheldon, B. Wooster, P. Olsen, P. Boanas, and P. Krustrup, "High-intensity running in English FA Premier League soccer matches," *J. Sports Sci.*, vol. 27, no. 2, pp. 159–168, 2009, doi: 10.1080/02640410802512775.

[22] C. Osgnach, S. Poser, R. Bernardini, R. Rinaldo, and P. E. Di Prampero, "Energy cost and metabolic power in elite soccer: A new match analysis approach," *Med. Sci. Sports Exerc.*, vol. 42, no. 1, pp. 170–178, 2010, doi: 10.1249/MSS.0b013e3181ae5cfd.

[23] M. Tomczak and E. Tomczak, "The need to report effect size estimates revisited. An overview of some recommended measures of effect size," *Trends Sport Sci.*, vol. 1, no. 21, pp. 19–25, 2014, [Online]. Available: http://www.wbc.poznan.pl/Content/325867/5_Trends_Vol21_2014_ no1_20.pdf.

[24] A. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, 2018, doi: 10.1613/jair.1.11192.

[25] T. B. Dijkhuis, F. J. Blaauw, M. W. van Ittersum, H. Velthuijsen, and M. Aiello, "Personalized physical activity coaching: A machine learning approach," *Sensors (Switzerland)*, vol. 18, no. 2, pp. 1–20, 2018, doi: 10.3390/s18020623.

[26] H. Liu, L. Wang, G. Huang, H. Zhang, and W. Mao, "Activity profiles of full-match and substitution players in the 2018 FIFA World Cup," *Eur. J. Sport Sci.*, vol. 0, no. 0, pp. 1–7, 2019, doi: 10.1080/17461391.2019.1659420.

[27] E. Rey, J. Lago-Ballesteros, and A. Padrón-Cabo, "Timing and tactical analysis of player substitutions in the UEFA champions league," *Int. J. Perform. Anal. Sport*, vol. 15, no. 3, pp. 840–850, 2015, doi: 10.1080/24748668.2015.11868835.

[28] D. B. Coelho *et al.*, "Effect of player substitutions on the intensity of second-half soccer match play," *Rev. Bras. Cineantropometria Desempenho Hum.*, vol. 14, no. 2, pp. 183–191, 2012, doi: http://dx.doi.org/10.5007/1980-0037.2012v14n2p183.

[29] M. A. Hall, "Correlation-based feature selection for machine learning," Waikato University, New Zealand, 1999.

[30] C. Tudor-Locke, L. Burkett, J. P. Reis, B. E. Ainsworth, C. A. Macera, and D. K. Wilson, "How many days of pedometer monitoring predict weekly physical activity in adults?," *Prev. Med. (Baltim).*, vol. 40, no. 3, pp.

293–298, 2005, doi: 10.1016/j.ypmed.2004.06.003.

[31] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock, "Methods and metrics for cold-start recommendations," in *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, 2002, pp. 253–260, doi: 10.1145/564418.564421.

CHAPTER 4

Transferring Targeted Maximum Likelihood Estimation for causal inference into sports science

Based on
"Transferring Targeted Maximum Likelihood Estimation for Causal Inference into Sports Science"

Talko B. Dijkhuis
Frank J. Blaauw

2022. Entropy 24(8):1060.
doi: 10.3390/e24081060.

ABSTRACT

Although causal inference has shown great value in estimating effect sizes in, for instance, physics, medical studies, and economics, it is rarely used in sports science. Targeted Maximum Likelihood Estimation (TMLE) is a modern method for performing causal inference. TMLE is forgiving in miss-specification of the causal model and improves the estimation of effect sizes using machine-learning methods. We demonstrate the advantage of TMLE in sport science by comparing the calculated effect size with generalized linear model (GLM). In this study, we introduce TMLE and provide a roadmap for making causal inference and apply the roadmap along with the methods mentioned above in a simulation study and case study investigating the influence of substitutions on the physical performance of the entire soccer team (i.e., the effect size of substitutions on the total physical performance). We construct a causal model, a miss-specified causal model, a simulation dataset, and an observed tracking dataset of individual players from 302 elite soccer matches. The simulation dataset results show that TMLE outperforms GLM in estimating effect size of the substitutions on the total physical performance. Furthermore, TMLE is most robust against model miss-specification in both the simulation and the tracking dataset. However, independent of the method used in the tracking dataset, it was found that substitutes increase the physical performance of the entire soccer team.

Keywords
Machine learning; statistics; methods; TMLE; causal inference.

1. INTRODUCTION

Empirical scientific research is intrinsically linked to statistical analysis and modelling. Statistical models are used to better understand phenomena and their underlying causal processes that are at play. Researchers rely on empirical data collected from these underlying causal systems that underpin these processes.

In the best case, this data is collected in a controlled environment using a Randomized Controlled Trial design (RCTs); a design that has been around for several centuries [1]. However, in many cases the world is messy, and especially in sports science an RCT during a match is often not possible and researchers rely on data obtained from observational studies. Controlling (all) variables is hard, if not impossible, or unethical. While the lack of RCTs seems to make causal inference difficult, methods exist that allow causal reasoning on observational datasets. Furthermore, alternative technologies exist that generally work better than the current status quo [2].

An elite soccer match is inherently only measurable by observing a complex set of latent causal relations, which complicates the determination of the isolated effects of an event on the outcome. Causal modelling of the influences in a match is intrinsically incomplete, and therefore applying a statistical method that is most robust to incorrectly specified models provides the best understanding of phenomena. A phenomenon of interest in soccer is the influence of substitutes. Substitutes are acknowledged to be important in soccer. In general, substitutions can be initiated by an injury of a player, necessary tactical changes (e.g., because of being behind in a match), or under-performance of a player [3]. Besides necessary substitutions (e.g., because of an injury), substitution may be the most powerful tool for coaches to influence a match. Substitutions can minimize or offset the effects of fatigue and give new stimuli to the match as elite substitutes introduced during the second half can cover more distance and perform more physically intensive actions relative to whole match players over the same period[4]. However, the observation that a substitute can cover a greater distance is a fraction of reality [4]. Despite an extensive body of research on substitutes, to the best of our knowledge, there is no single study that studies the causal inference of the influence of a substitute on the total physical performance of a soccer team. That is: does the total team physical performance increase due to the use of substitutes?

One particular field of causal inference that has received traction over the past years is the Targeted Learning approach [5]. The Targeted Learning methodology aims to reconcile traditional statistical inference with modern, state-of-the-art machine learning models. In this paper, we focus our interest on Targeted Maximum Likelihood Estimation (TMLE), a method that enables causal reasoning and modelling and that can improve

model performance and correctness. TMLE is a semi-parametric double-robust method that can withstand miss-specification of the causal model, improving the estimation of effect sizes using machine-learning methods. Double-robust implies that the estimation of the effect remains consistent if either the propensity score model[1] or the outcome model is miss-specified [6].

Although TMLE is not new, its use in the field of sports science is absent. Often GLMs are used to study the physical performance of teams [7]–[9]. A disadvantage of GLM is that it is not robust on miss-specification and is an oversimplified representation of the real world [10]. However, its simplicity is also one of GLMs' strengths. Assuming the model is well specified, it can give insight into the various essential coefficients for a measured outcome. Such statistical inference is generally impossible to achieve in complicated machine learning models [2]. Such machine learning models focus on prediction and learn this by minimizing a loss function, instead of focusing on statistical inference. TMLE aims to reconcile statistical inference and machine learning by introducing a two-step approach [2], [11], [12]. A machine learning algorithm is first trained on the dataset and then adapted to a particular question of interest in the so-called targeting step. With this step, non-parametric models, such as many machine learning models can be used while statistical inference is still possible [2], [13].

The aim of this paper is twofold. Firstly, we aim to provide a roadmap for making causal inference in sports science. Secondly, we aim to examine the applicability of the roadmap combined with a study of the performance TMLE in comparison with the traditional Generalized Linear Model (GLM) in identifying the effect size of a substitute in soccer. On the one hand we define a simulation study using simulation data on the influence of a substitute on the total soccer team distance as a measure for physical performance. To study the performance of TMLE in comparison with the traditional GLM, the identified substitution effect size of TMLE and GLM are compared using correct and miss-specified causal models. On the other hand, we apply observed match data to look at the effect size of a substitute on the total team performance in elite soccer using the roadmap combined with TMLE and GLM.

Thus, we provide the basis for bringing causal inference and TMLE into the toolbox of sports science research and improving the quality of causal inference in sport science.
The paper is structured as follows. In Section 2 we present the work that is related to the current study. In this we focus on scientific literature from the field of substitutes in

1 A propensity-score denotes the chance of a treatment given the confounders. If a certain stratum has a higher chance at receiving a treatment (e.g., being female increases the chances of receiving a treatment), a propensity-score can be used to control for this.

soccer, and from the field of targeted maximum likelihood estimation. In Section 3 we present the methods used in this paper. This section defines the causal roadmap and its application to the current problem. Section 4 presents the results of our study. We present both the results of our simulation study as well as our application of TMLE to substitutions in soccer. Finally, in Section 5 and Section 6 we discuss and conclude the work.

2. RELATED WORK

The related work on TMLE and causal modelling and the standard statistical methods to study substitution are the basis for our research on the applicability of causal inference in sport science.

2.1. Statistics and performance of substitutes in soccer

Research of performance, substitutes, and soccer, has previously only been done using traditional statistical methods [3], [4], [14]–[16]. For example, Bradley, Lago-Penas, and Rey [4]studied the match performances of substitute players using one-way independent measures Analysis of Variance (ANOVA). The performance of the substitutes was compared with the players completing the entire match. The meaningfulness of the differences between the substitutes and full match players was indicated by the Effect Size (ES). Effect size is, as defined by Kelley and Preacher, 2012 [17], “We define effect size as a quantitative reflection of the magnitude of some phenomenon that is used for the purpose of addressing a question of interest”. The authors show, amongst others, that substitutes cover a greater total distance (ES: 0.33–0.67).

Modric et al. [14] investigated the relation between Running Performance (RP) and Game Performance Indicators (GPI). The RP included the total distance covered, distance covered in five speed categories, and the GPI were determined by the position specific InStat index (InStat, Moscow, Russia). The InStat index is calculated based on a unique set of parameters for each playing position, with a higher numerical value indicating better performance. The exact calculations are only known by the manufacturer of the platform. The associations between RP and GPI were identified by calculating Pearson’s product moment correlation coefficient. Correlations were found between RP and GPI for different positions. For instance, the total running distance and high-intensity accelerations were correlated with the InStat index for Central Defenders ($r = 0.42$ and $r = 0.49$, respectively).

Hills et al. [3] profiled the match-day physical activities performed by substitutes, focusing separately on the pre- and post-pitch-entry periods. Linear mixed modelling

was conducted to differentiate outcome variables as functions of time. A variance components model with no predictors was established for each outcome measure before sequentially allowing intercepts and slopes to vary. A combination of random slopes and intercepts was employed based upon Bayesian information criterion assessments of model fit. One of the conclusions was: substitutes covered a greater ($p < 0.05$) total (+67 to +93 m) and high-speed (+14 to +33 m) distances during the first five minutes of match-play versus all subsequent epochs.

M. Lorenzo et al. [16] aimed, amongst others, to analyse the physical and technical performance of substitute players versus entire match players or players who were replaced. Linear mixed models analysed the differences between the performance of substitute, replaced, and entire match players. Bonferroni's post-hoc test and Cohens' d conducted the group comparison and the effect size. One of the results was substitute players showing higher total distance covered (ES: 0.99–1.06), number of sprints (ES: 0.60–0.64), and number of fast runs (ES: 0.83–0.91) relative to playing time than replaced and entire match players.

All studies mentioned above, and their applied methods have in common that they indicate an association between elements of a soccer match but leave out many factors that influence the association's actual effect size. A combination of the results of Modric et al. [14] and the remaining three [3], [4], [16] indicate that a substitute player has a better game performance. Even the combination leaves out the influence of the substitutions on the total performance. The methods used and the factors investigated grab only a tiny part of the overall complex system of a soccer match. As Morgulev et al. [18] indicate, it is hard to conclude causality in sports complex systems due to endogeneity problems even when a correlation is found. Endogeneity means either a variable correlated with both the independent variable in the model and with the error term or a left-out variable affecting the independent variable and separately affects the dependent variable. Sports complex systems are influenced by various left-out factors in the studied phenomenon, making it complex to find causal inference [18].

2.2. TMLE and causal modelling

Targeted learning is a unique methodology, which reconciles advanced machine learning algorithms and semi-parametric inferential statistics [2]. The data available for analysis in sports is proliferating [19] and presents a challenge to both inferential statistics and machine learning. The vast amount of data in sports from, for instance, a semi-automatic multiple-camera video technology in soccer, combined with the inherent complexity of the data-generating process complicates statistical inference and the underlying mathematical theory. Such as limiting the use of miss-specified models, acknowledging that the models do not contain and compensate for the truth, looking for

causal relationships in non-experimental data, the proper quantification of uncertainty, etcetera. The challenge is to prevent the specification of uninterpretable coefficients in miss-specified parametric models (e.g., GLMs) where different choices of such miss-specified models yield different answers [2], [20]. In contrast, the targeted learning method (e.g., TMLE) aims to construct confidence intervals for user-specified target parameters by targeting the estimates retrieved from data adaptive estimators (e.g., machine learning) while relying solely on accurate statistical assumptions. This approach can reduce differences in statistical analysis results as model choices are automated, allowing for consistent estimates regardless of the researcher conducting the study [21]. The Targeted Learning methodology focuses on the art of causal modelling [2]. Causal modelling is a technique used to provide a formal model for and express assumptions about data-generating processes [22]–[24]. Currently, the four main approaches used for causal modelling are (i) Graphical models, (ii) potential-outcome models, (iii) sufficient-component cause models, and (iv) structural equations models [22]. These approaches offer complementary perspectives and can be used together to enhance causal interpretations [25].

With our paper we aim to introduce a roadmap to use the TMLE methodology in the field of sports science. As such, we introduce causal inference as a new tool in the sports scientists' toolbox.

3. MATERIALS AND METHODS

We adhere to the causal roadmap as a procedure to structure scientific research [22], [26]. This roadmap takes the form of seven steps: (i) specifying the knowledge of the system to be studied using a causal model, (ii) specifying the data and their link to the causal model, (iii) specifying the target causality, (iv) assessing identifiability, (v) stating the statistical estimation problem, (vi) estimation, and (vii) interpretation. By following this roadmap, we create a clear distinction between the knowledge about the system under study and about the assumptions that need to be made to answer the research questions. TMLE is part of this procedure and is applied in the estimation step. The present work adheres to this general structure and is what we see as the blueprint for performing TMLE in sports science.

3.1. Specifying the knowledge of the system to be studied using a causal model

The first step in this roadmap is to define the knowledge about the system under study. Knowledge, in this case, is actual, fundamental knowledge about the system, and should not rely on assumptions on the underlying model. One way to define this system is by using a causal graph representation, which depicts the causal relations of the system

[23]. The causal graph for the influence of a substitute in soccer is shown in Figure 1.

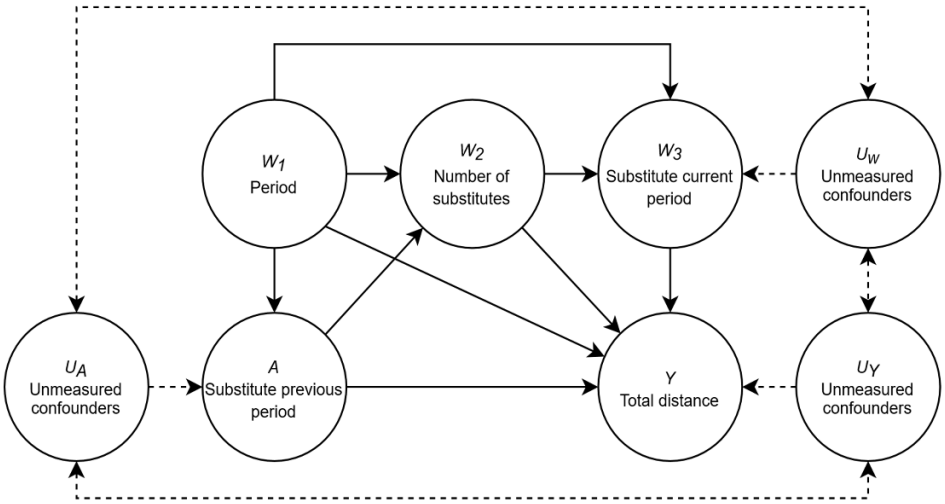


Figure 1. The causal model representation of the system being studied.

Y = the total distance of a team in five-minute periods; A = a substitute or not in the previous five-minute period; W_1 = the consecutive five-minute periods in the second half of the match; W_2 = the number of substitutes present; W_3 = number of substitutes in the current period; U = possible unknown confounders influencing A, W_3 and Y. The dashed lines indicate that this confounding effect is uncertain.

The causal graph shows the causal relations between variables in the system. For example, an arrow from A to B describes a causal effect of A on B, or in other words, A causes B. This figure also gives rise to some notation that will be used throughout the paper. The nodes on the top of the graph are the W variables, which indicate the measured confounders (i.e., factors) in the model, A indicates the intervention or treatment that has been performed, Y the outcome of the model, and U any unmeasured confounders that influence our results. With this notation we aim to stay close to the notation used in other literature (e.g.,[2], [27]).

Case study

We concretize the aforementioned variables as follows, $W = (W_1, W_2, W_3)$ are the three measured confounders in our model, in which ; W_1 is the consecutive five-minute periods in the second half, W_2 the number of substitutes present, and W_3 whether there was a

substitute in the current period. Our treatment variable, $A \sim B$ is a binary intervention which indicates whether a substitution happened in the previous five-minute period. $U_{W,A,Y} \sim Pu$ are the unmeasured confounders that potentially influence the variables in the model[2]. Pu is the unknown distribution from which $U_{W,A,Y}$ is instantiated. Finally, we have the outcome of our model, $Y \sim N$ (in which N denotes the normal distribution) a proxy for performance measured by the total distance covered by the team. A higher distance covered by the team indicates a higher performance.

The relationships between these variables are defined as follows; period W_1 influences the total distance of team Y, which is known to decline during the match [4]. As substitutions are highly dependent on the moment of the match, the period W_1has a relationship with the substitutes present W_2, current period substitutions W_3, and substitutions of the previous period A. The total distance of the team Y depends on the number of substitutes present given A and W_2, while substitutions cover more distance than all match players. When a substitute occurs within the current period W_3, it leads to a dead ball moment and reduces the overall distance Y. Substitutes in the current period and previous period are also influenced by unknown confounders like an injury or tactical decisions. The overall distance Y of a team does not solely depend on the period and substitutes, and other possible unknown confounders U in our model are not accounted for but potentially influence the total distance Y [28]. After this first step, we have a clear definition of the knowledge and the relationships between the different variables under study, allowing us to move to the data we have about this system.

3.2. Specifying the simulation data, the observed data, and its link to the causal model

In the second step we specify the observed and simulation data, and its link to the causal model. The causal model we defined in the first step presents what we know about the system, whereas the data describes what we have observed from it. The causal model describes various possible processes that yielded the data. This description of possible processes is strongly connected to the underlying statistical model of the data, that is, the set of all possible distributions from which the data originates. For this we define the data as $O \subset \mathcal{O} \sim P$, where $\mathcal{O}$ is the space of all possible generated data and P is the data generating distribution.

Case study

3.2.1. Simulation data

We implemented a data simulator to generate datasets according to the causal model in

2 Such as, playing home or away, rank of the teams, position system they play, current score, etc. These variables are by definition unknown and unmeasured. We do not know whether such variables exist and actually influence the model. However, they could be, which is why they are mentioned here.

Figure 1. The code of the data generating system is written in R version 4.0.2 and available online[3] . The observations originating from this simulator are defined as $\hat{O}_i = (W, A, Y) \sim P_s$, in which $W = (W_1, W_2, W_3)$ are the confounders and $A \in \{0,1\}$ is an indicator variable indicating whether a substitution happened in the previous period. P_s is the simulation probability distribution from which the simulation observations $\hat{O}$ were sampled[4]. The subscript *i* indicates a specific simulation observation $\hat{O}_i \in \hat{O}$.

3.2.2. Observed data

We retrospectively collected the in-match position tracking data from 302 competitive professional soccer matches between 18 teams during the Dutch premier league 'Eredivisie' 2018–2019 season. The players' time, position, speed, and acceleration were detected and recorded by the SportsVU optical tracking system (SportsVU, STATS LLC, Chicago, IL, USA). Linke et al. (2018) tested the SportsVU optical tracking system and rated the system as being adequately reliable[29].

For our analysis, two matches with erroneous and missing data were excluded. We only used the second half of the matches expecting the substitution being most effective. Additionally, the extra time at the end of the second half and goalkeepers were excluded from the dataset. The effect of substitution on the match was controlled by identifying both entire-match players and substitutes. Thus, entire-match players played the entire match, while the substitutes entered the match at a later stage.

The dataset was divided into periods of five minutes and contained a total of N = 5226 observations (O_n). As an illustration of the data, Figure 2 shows the increasing number 259 of substitutes during the second half. The influence of a substitution in a previous 260 period on the total distance of the team compared to no substitution in the previous 261 period is visualized in Figure 3.

Each observation $O_i \in O_n$ is considered mutually independent[5]. Each of these O_n is defined as $O_i = (W, A, Y) \sim P_0$, in which $W = (W_1, W_2, W_3)$ are the confounders, and $A \in \{0,1\}$ is an indicator variable indicating whether a substitution happened in the previous period, P_0 is the unknown real underlying probability distribution from which O_n was sampled, and Y is the total distance of the team in meters. In the remainder of the work, we will refer to P_n as the empirical distribution of the data. The observed

3 Available at https://github.com/dijkhuist/Entropy-TMLE-Substitutions.

4 The hat (ˆ) signifies that this is data from the simulator.

5 Note that the data we deal with in this case study possibly has a time dimension stronger than what we are currently showing in our causal model. In fact, Y at time t could potentially influence W_3, or even A and Y itself at time t + 1. As our aim with this paper is to introduce TMLE and causal inference in sports, we will not detail on the time dimensionality of the data. For more information on time series analysis in Targeted Learning, please see [30].

dataset is available online [6].

Note that in the remainder we work with a min-max normalized, bounded version of $Y \in [0,1]$. *While this is not relevant for the initial steps of the roadmap, the boundedness of* Y *will become important in the later steps (specifically the estimation step).*

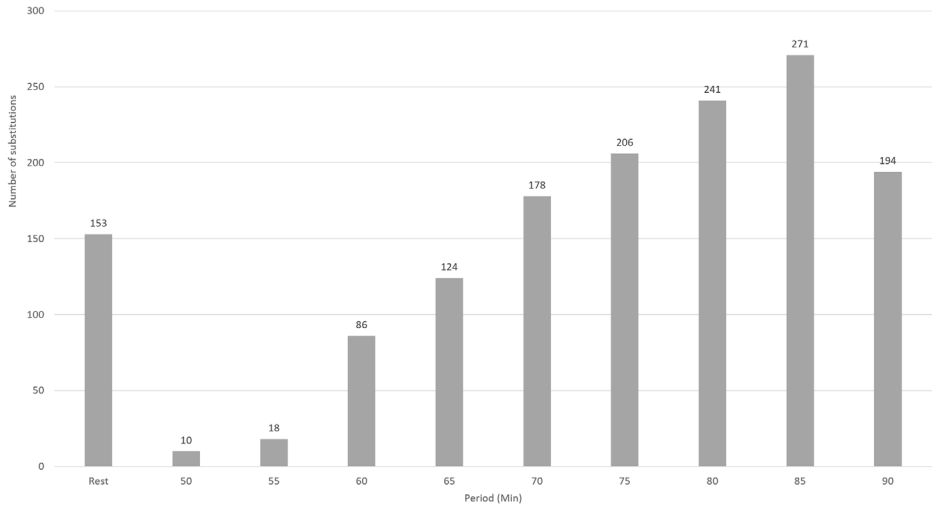


Figure 2. Number of substitutions in the second half per 5-minute period.

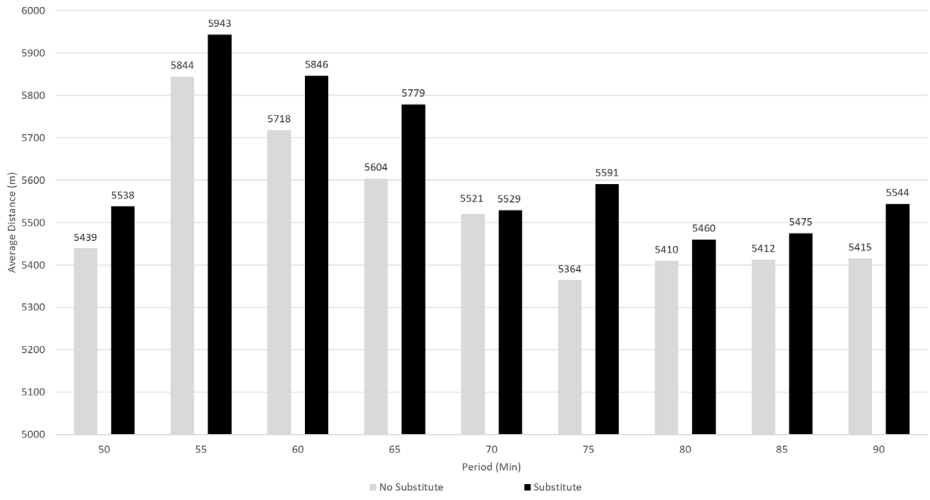


Figure 3. Difference in the total distance when a substitution took place in the previous period or not (A).

6 Available at https://github.com/dijkhuist/Entropy-TMLE-Substitutions/tree/main/Data.

3.3. Specifying the target quantity

The third step in the roadmap is the definition of the target, the causal quantity, or, more specifically, the definition of the causal question of interest. The target quantity can be seen as the main question we would like to answer about the underlying system. Examples of target quantities are: '*What is the average treatment effect of a medicine versus placebo?*' or '*How much does gender influence the outcome of a drug?*'. This approach is significantly different from general machine learning approaches, as these generally focus on optimizing a prediction for a multitude number of questions at hand. In contrast, the targeted learning approach only picks one specific question, drastically reducing the complexity of the problem [21] . To define this target quantity, we need to identify the target population with which we are working, the intervention we are doing on this target population, and the outcome we are interested in.

Case study

In our case study, we are interested in determining the effect of substitution (the intervention; A) on the total distance in meters (the outcome; Y) of the team (the target population). We can further specify our question using the notion of counterfactuals; an alternative scenario that has not occurred but that helps us to answer our question. In our case study, we want to see the effect of a substitution $A = 1$ versus not doing a substitution $A = 0$. In some cases, the actual observation we did might not have had a substitution at that time; thus, it represents a 'counterfactual world.' Using these counterfactuals, we can adequately define what we are interested in in our case: *we are interested in the difference in team distance between a substitution vs. no substitution simultaneously in time.*

3.4. Assessing identifiability

In the fourth step, we determine identifiability. It should be determined whether sufficient knowledge and data are available to answer the causal question or whether additional assumptions need to be made. The defined causal question can be modelled as an *average intervention effect*, or Average Treatment Effect (ATE), in which a substitution is seen as the intervention / treatment. In social studies, ATE is referred to as Effect Size [30], [31]. Formally, an ATE can generally be formulated using the G-computation formula [32],

$$\boldsymbol{\psi_0 = \Psi(P_0) = \mathbb{E}_W[\mathbb{E}(Y \mid A = 1, W) - \mathbb{E}(Y \mid A = 0, W)]}. \quad (1)$$

This G-computation formula determines the average effect of a treatment by determining the average difference between the outcomes for the treated and the non-treated. Note that we use the notation P_0 here to denote the true probability distribution from which O originates[7].

7 Note that we're not discussing the unmeasured confounders and the distribution thereof for the sake of clarity. Please see the Targeted Learning of van der Laan and Rose book [2] for more details.

Case study

For the target causality to be identifiable, we need to write our target parameter as a function of the actual distribution P_0. That is, identifiability would give us $\Psi(P_0) \equiv \Psi(P_n)$. In order to make this claim, we need to impose assumptions on the system. In our case study, we need two assumptions: (i) a positivity assumption and (ii) a no unmeasured confounders assumption (randomization assumption)[8].

The positivity assumption stated as $P(A = a \mid W) > 0 \mid \forall a \in A$ indicates having enough observations with treatments and controls for all strata of W. For each combination of $w \in W$, we assume that the probability of treatment is greater than zero. If this assumption does not hold, it is not possible to infer the outcomes for the missing strata. The assumption will hold both in the case of simulation data and the observed data[8].

The second assumption is the no unmeasured confounders. This assumption states that there is no unmeasured confounding between treatment A and outcome Y, that is $Y \perp\!\!\!\perp A \parallel W$. If we fail to make this assumption, it could be that there is an extraneous variable that influences both our treatment and our outcome variable, yielding the estimation of the causal effect of A on Y unreliable. In the simulation data there are no unmeasured confounders, as we control the causal model, the data, and the targeted quantity. This assumption is hard to validate for the observed data, as there are always unmeasured confounders in the real world. As can be seen in Figure 1, we know that there is the possibility that an underlying confounding effect exists, and we assume that in our case these effects do not exist / do not significantly impact the outcome of our model. If the dimension of W, measured confounders, is large enough, this assumption is likely to be valid. In this case study , for apparent reasons, this assumption is not satisfied.

3.5. Stating the statistical estimation problem

In the fifth step, we state the statistical estimation problem and determine whether all the goals are met to answer our causal question. To perform this estimation, we rely on several assumptions, which are both knowledge-based, and convenience-based [22]. Knowledge based assumptions are based on actual knowledge that we have about the causal model and the data. Convenience-based assumptions are assumptions that provide identifiability, if true.

Case study

In our case study (and in many cases), the knowledge-based assumptions are not enough to reach identifiability and reason about causality, and as such, we introduced two convenience assumptions; a positivity assumption and an unmeasured confounding

8 This assumption will not hold when any w ∈ W is continuous. If that would be the case, we need to discretize W until the assumption holds.

assumption (see Section 3.4). These assumptions are needed as we only have limited knowledge about the system we are dealing with. In general, such assumptions should be kept to a minimum (as few as possible, but enough to allow for statistical inference). In our case, the simulation dataset meets both the knowledge-based and the convenience-based assumptions, for we control all aspects of the simulation dataset. In contrast, the tracking dataset meets all assumptions except for the unmeasured confounding assumption.

3.6. Estimation

In the sixth step, the actual estimation done. So far, the roadmap has only helped define the problem we are solving and define the knowledge we have about the problem. With estimation, we aim to find a parameter ψ_n as an estimate of the true parameter ψ_0 of the true data generating distribution P_0. To provide some intuition, the observed data we collected, $O \sim P_0$ is an empirical realization of data retrieved from the true data generating distribution, P_0. Suppose P_0 is controlled by an infinite-dimensional parameter ψ_0 which controls the data P_0 generates. Since we do not know P_0, nor ψ_0, we aim to find the parameter ψ_n, which is as close as possible to ψ_0. We define a mapping function $\Psi: \mathcal{M} \rightarrow \psi$, in which $\mathcal{M}$ is the statistical model, defining all distributions $P_0 \in \mathcal{M}$ from this mapping follows that $\Psi(P_0) = \psi_0$ that is; the function Ψ yields the true parameter when provided the true distribution. Our goal is to find an estimator based on the empirical data, $\widehat{\Psi}(P_n) = \psi_n$ in which $\widehat{\Psi}: \mathcal{M}_{\text{non-parametric}} \rightarrow \psi$.

To illustrate the process of defining an estimator $\widehat{\Psi}(P_n)$ of $\Psi(P_0)$, our explanation will follow two stages. We will first start with a basic estimation procedure illustrated using a traditional generalized linear model (GLM) approach. Secondly, we show how an estimator of $\Psi(P_0)$ can be defined using Super Learning and TMLE. We can take this approach as we are dealing with a so-called *substitution estimator*, or *plug-in estimator*, allowing us to view the implementation of the estimator itself as an implementation detail [2].

3.6.1. GLM based estimation

The general estimation procedure relies on the definition of Q_0 the relevant part of P_0 needed for the target parameter. That is, $\Psi(P_0) \equiv \Psi(Q_0)$. In our definition of Ψ in Equation (1), $\Psi(P_0)$ only relies on $\bar{Q}_0(A, W) \equiv \mathbb{E}[Y \mid A, W]$ and on $Q_{0,W}$, the distribution of W[9]. As such, Q_0 is defined as the collection $Q_0 = (\bar{Q}_0, Q_{0,W})$. With these definitions, we now need to define algorithms that take in the empirical data, and for this we define the following steps:

9 We use the bar (-) to differentiate between Q_0 and the element $\bar{Q}_0$, which is consistent with other Targeted Learning literature

1. Estimate $\bar{Q}_0(A,W)$ (e.g., using machine learning or a parametric model). That is build an estimator for $\mathbb{E}[Y \mid A,W]$.
2. Generate predictions from the estimator for each observation, where we set A for each observation (i.e., create counterfactual worlds). That is, we $\bar{Q}_0(A=0,W)$ and $\bar{Q}_0(A=1,W)$ for each $O_i \in O$ (discarding the original values of A). With this we make predictions in the two counterfactual worlds 'what if everyone received a treatment?' versus 'what if no one received treatment?'.
3. Estimate ψ_n using the G-computation formula as defined in Equation (1)

Note that to estimate $Q_{0,W}$ we use the empirical distribution of W, and give each a weight of $\frac{1}{n}$.

In our initial example we assume a simplistic parametric linear model. Following the steps, we first estimate $\bar{Q}_0(A,W) \equiv \mathbb{E}[Y \mid A,W]$. Using a linear model, such as GLM, this can be estimated as

$$\bar{\boldsymbol{Q}}_{\boldsymbol{0,glm}}(\boldsymbol{A},\boldsymbol{W}) \equiv \mathbb{E}_{\boldsymbol{n}}[\boldsymbol{Y} \mid \boldsymbol{A},\boldsymbol{W}] = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 A} + \boldsymbol{\beta_2 W} \tag{2}$$

With the formula in Equation (4) we can estimate $\hat{Y}_1$ and $\hat{Y}_0$ We use the subscript 1 and 0 on $\hat{Y}$ to indicate that this value of $\hat{Y}$ was calculated by respectively setting A = 1 and A = 0. That is, $\hat{Y}_x$ is the evaluation of Equation (4) for all On, resulting in a list of tuples $\{\hat{Y}_1, \hat{Y}_0\}\ \forall O_i \in O$, which can be used to calculate the ATE as

$$\boldsymbol{\psi_n} = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \left[\mathbb{E}_{\boldsymbol{n}}[\boldsymbol{Y} \mid \boldsymbol{A}=\mathbf{1},\boldsymbol{W_i}] - \mathbb{E}_{\boldsymbol{n}}[\boldsymbol{Y} \mid \boldsymbol{A}=\mathbf{0},\boldsymbol{W_i}]\right] \tag{3}$$

$$= \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \widehat{\boldsymbol{Y}}_{\mathbf{1}} - \widehat{\boldsymbol{Y}}_{\mathbf{0}} \tag{4}$$

3.6.2. Super learning and TMLE based estimation

While the linear model provides an initial estimate, the underlying estimator follows a strictly parametric and linear nature, and thus poses various assumptions on the model that we currently cannot assume. To prevent these assumptions, the alternative is to use flexible machine learning techniques in a *super learner* approach and applying Targeted Maximum Likelihood estimation to perform the estimation of ψ_n.

Note that we describe some of the background and intuition behind Super Learner and TMLE. For more information and formal proofs we would like to refer to Van der Laan and Rose [2][10].

10 There are also several R packages available that automate the process discussed below. For this, see https://tlverse.org/

Machine learning and cross-validation

Machine learning focuses on training algorithm to perform an optimal prediction of an outcome Y given the input parameters $X, \mathbb{E}(Y \mid X)$. Training a machine learning model works by minimizing a so-called loss function over a series of cross-validation folds.

Cross-validation aims to estimate how well a trained model performs on unseen data by sequentially leaving out data from the training procedure by minimizing a loss function. Cross-validation splits up the data $Z = \{Z_1, \ldots, Z_n\}$ into training and validation sets. Training and validation sets can be modelled using a random variable $B_n \in \{0,1\}^n$. With V different cross-validation folds, B_n can take V different values, resulting in a $b_1, \ldots, b_v \in \{0,1\}^n$. Each b_v then corresponds to either of two sets; a training dataset$\{Z_i : \leq i \leq n, b_v(i) = 0\}$ and a validation set $\{Z_i : \leq i \leq n, b_v(i) = 0\}$. In this case, $b_v(i)$ corresponds to the i^{th} entry of vector b_v. In our case, we only use one of the splits as a test set, $\sum_{v=1}^{V} b_v = 1$. Thus, each observation falls once in the validation set, and is used $V - 1$ times in the training set.

Super Learning

Cross-validation forms the basis of machine learning and is equally important for super learning. Super learning is a specific instance of machine learning that applies an ensemble methodology to automatically select the best machine learning algorithm, or a convex combination of machine learning algorithms. The Super Learner selects the best estimator among all candidate estimators based on these cross-validation scores [5]. The methodology generally consists of two implementations: the discrete super learner and the continuous super learner. For each cross-validation fold, the discrete super learner starts with a set $L = \{l_1, \ldots, l_m\}$ learners. These learners can be anything used to perform the prediction $\mathbb{E}[Y \mid X]$, and could be as simple as a mean of the data, and as complex as a neural network or random forest. The Super Learner trains each $l_i \in L$ on each cross-validation fold, resulting in a set of estimators $\bar{L} = \{\bar{l}_{i,j} \ldots \bar{l}_{m,v}\}$ and an accompanying cross-validation risk (loss) for each cross-validation fold. Based on these cross-validation risks, the discrete super learner selects the algorithm with the lowest risk by averaging across the folds.

$$\underset{\bar{l}_m^r \in \bar{L}^r}{\boldsymbol{arg\,min}}\, \boldsymbol{SL_d}\left(\bar{l}_m^r\right) = \frac{1}{V}\sum_{j=1}^{V} \bar{L}_{m,j}^r \tag{5}$$

The continuous super learner applies a similar procedure, only instead of selecting the single best estimator, it aims to find weights $\alpha = \{\alpha_1, \ldots, \alpha_m\}$ where

$$\boldsymbol{\alpha} = \left\{\boldsymbol{\omega} \in \mathbb{R}_+^M : \sum_{m=1}^{M} \boldsymbol{\omega_m} = \mathbf{1}\right\} \tag{6}$$

for each learner. The Super Learner is then defined as the dot product

$$\boldsymbol{SL_c(L, \alpha) = L \cdot \alpha} \tag{7}$$

The weights in this case are calculated in such a way that they minimize the risk of the SL_c.

Targeted Maximum Likelihood Estimation

After the initial estimation step is completed, the next step is to perform the Targeted Maximum Likelihood Estimation (TMLE) step [2], [13]. The goal of TMLE is to reduce the bias of the estimation of the target parameter [33]. Figure 4 presents an abstract representation of TMLE and its goal. In this graph the circle depicts $\mathcal{M}$, the set of all possible probability distributions. As can be seen, $P_0 \in \mathcal{M}$, which maps to the target parameter $\Psi(P_0)$. Our aim is to use $P_n \in \mathcal{M}$ with the corresponding $\Psi(P_n^*)$ to create a targeted estimate closer to the true target parameter.

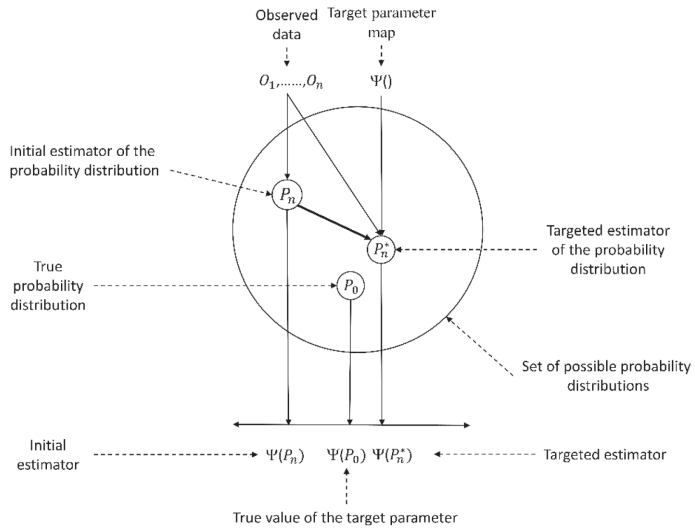


Figure 4. Graphical depiction of TMLE [2].

The definition of the ATE TMLE estimator ψ^* is given by

$$\boldsymbol{\psi^* = \frac{1}{n}\sum_{i=1}^{n} [\bar{Q}_n^*(1, W_i) - \bar{Q}_n^*(0, W_i)]} \tag{8}$$

Which is the targeted version of ψ (Equation (1)). We use the notation $\bar{Q}_n^0(A, W)$ to denote the initial estimate of $\mathbb{E}[Y \mid A, W]$ and, $\bar{Q}_n^*(A, W)$ to denote its targeted counterpart.

Targeting $\bar{Q}_n^0(A, W)$ involves the two new nuisance parameters: the treatment

mechanism $g_n(A|W)$ and the clever covariate $H_n(A_i, W_i)$. The treatment mechanism $g_n(A|W) \equiv P(A|W)$, can be estimated using, for example, super learning.

The clever covariate can balance the distributions of observed data of the samples under treatment versus the samples under control [11]. The clever covariate is defined for each individual as

$$\boldsymbol{H_n(A_i, W_i) = \left(\frac{I(A_i=1)}{g_n(A_i=1|W_i)} - \frac{I(A_i=0)}{g_n(A_i=0|W_i)}\right)} \quad (9)$$

This clever covariate does not need estimation, but is used for fluctuating the initial estimate of $\bar{Q}_n^0(A, W)$, by relying on information collected about the treatment and control groups (i.e., the ratio between treated vs. control) [11].

Based on these definitions, the steps that are needed in order to estimate the TMLE are as follows:

1. Estimate $\bar{Q}_n^0(A, W)$ (e.g., using machine learning or a parametric model)
2. Generate predictions from the estimator for each observation, where we set A for each observation. That is, we estimate $\bar{Q}_n^0(A = 0, W)$ and $\bar{Q}_n^0(A = 1, W)$ for each $O_i \in O$ (discarding the original values of A)
3. Estimate the treatment mechanism $g_n(A \mid W)$.
4. Create the clever covariate $H_n(A_i, W_i)$.
5. Update / fluctuate the initial estimate of $\bar{Q}_n^0(A, W)$ using the clever covariate.

The last step in this procedure describes updating the initial estimate. This is performed by applying a logistic regression on Y on H, using our initial estimate as offset. The logistic regression is used to ensure that TMLE is bounded, as introduced by min-max normalizing the outcome variable Y. The fluctuation can then be performed on a logistic scale [11].

$$\boldsymbol{\mathbf{logit}\left(\mathbb{E}(Y \mid A, W)\right) = \mathbf{logit}\left(\bar{Q}_n^0(A, W)\right) + \epsilon H_n(A, W)} \quad (10)$$

$$\boldsymbol{\bar{Q}_n^*(A, W) = \mathbf{expit}\left(\mathbf{logit}\left(\bar{Q}_n^0(A, W)\right) + \epsilon H_n(A, W)\right)} \quad (11)$$

Case study

For the current simulation study and the case study we did not implement these steps ourselves, but instead relied on an existing R packages that perform most of the calculations. We used the R 'tmle' package, version 1.5.0-1 for doing the Targeted Maximum Likelihood Estimation, and the 'superlearner' R package, version 2.0-26, for both the simulation study and the case study.

For simulation, we used the data simulation system conforming to the causal model. Because we *know* the exact configuration of this simulator, we can correctly, or purposely incorrectly, specify the data that our learning algorithms take into account. As such, we performed a series of experiments using GLM as defined in Section 3.6.1 and TMLE using super learning as defined in Section 3.6.2 applying standard learners and handpicked learners (TMLEH): glm, glm.interaction, step, step.interaction, glm.interaction, gam, randomForest, rpart. We used the continuous Super Learner in all experiments. We first calculated the actual expected ATE on the total distance of the soccer team (Y) given a substitution in the previous period (A) and used that as the ground truth of our simulator. After that, we estimated the ATE of a substitution in the previous period (A) on the total distance of the soccer team (Y) using the three algorithms mentioned above. First, we used a correctly specified model as input to show the optimal performance of each of the algorithms. After that, we used a miss-specified model leaving the substitution of the current period (W_3) out of the model to indicate how each of the algorithms could cope with this. The code of simulation is written in R 4.0.2 and available online[11] .

Next to the simulation study, we studied how TMLE can be applied to the observed dataset. For the application on the observed dataset, we calculated the ATE of a substitute in the previous period (using GLM as defined in Section 3.6.1 , TMLE and TMLEH using (continuous) super learning as defined in 3.6.2). First, we used a correctly specified model as input to answer the question on the influence of substitution in the previous period (A). After that, we used a miss-specified model leaving the substitute in the current period (W_3) out of the model to indicate how the algorithms would handle the absence of a confounder. The code of case study is written in R 4.0.2 and available online[12]

3.7. Interpretation

The last step of the roadmap is the estimation interpretation. Depending on the strength of the assumptions made in 3.5. The stronger the assumptions, the stronger the relationship between the phenom observed and the interpretation. To interpret the results of the data analysis, we can hierarchically depend on the strength of the assumptions on the use of statistical, counterfactual, feasible intervention, or randomized trial [22]. 'The use of a statistical model known to contain the true distribution of the observed data and of an estimator that minimizes bias and provides a valid measure of statistical uncertainty helps to ensure that analyses maintain a valid statistical interpretation. Under additional assumptions, this interpretation can be augmented.'[22].

Case study

In our case study, we made both knowledge-based and convenience-based assumptions on the simulation dataset and the observed dataset containing the true distribution and

11 Available at https://github.com/dijkhuist/Entropy-TMLE-Substitutions
12 Available at https://github.com/dijkhuist/Entropy-TMLE-Substitutions

allowing the analysis and interpretation to be statistical. Section 4 shows our results and the interpretation thereof.

4. RESULTS

Applying the simulation data on the defined causal model, both TMLE and TMLEH have less deviation of the true ATE of the influence of a substitute in the previous period on the total distance of the entire soccer team than GLM (Table 1 and Table 2). When the miss-specification of the causal model is applied (e.g., leaving out the substitution in the current period) , the increase of deviation of the true ATE is almost non-existent for TMLE and TMLEH, where GLM shows an increased deviation of the true ATE. Figure 5 illustrates the effect of the miss-specification, leaving out the substitute of the current period, on the resulting ATE of a substitute in a previous period. Applying the observed dataset, the influence of a substitution on the total distance of the soccer team differs per algorithm ATE: 0.0105 - 0.0149 (Table 3). Miss-specification of the causal model, leaving out the substitute of the current period, using the real dataset leads to less deviance in TMLE and TMLEH from the respective calculated ATE of the substitute in the previous period on the total distance of the soccer team than GLM (Table 3).

Table 1. Simulation of the correct causal model.

True ATE: 0.0646			
Measure	GLM	TMLE	TMLEH
ATE	0.1442	0.0647	0.0647
Confidence Interval 95%	0.1399-0.1485	0.0628-0.0665	0.0605-0.0688
Bias	0.0797	0.0001	0.0001
Bias %	123.50	0.22	0.17

GLM = Generalized Linear Model; TMLE = Targeted Maximum Likelihood Estimation; TMLEH = Targeted Maximum Likelihood Estimation using Handpicked algorithms; ATE = Average Treatment Effect (i.e., Effect Size) of a substitute in a previous period on the total distance of a soccer team.

Table 2. Simulation of miss-specified causal model.

True ATE: 0.0646			
Measure	GLM	TMLE	TMLEH
ATE	0.1491	0.0647	0.0646
Confidence Interval 95%	0.1399-0.1485	0.0628-0.0665	0.0613-0.0679
Bias	0.0846	0.0001	0.0000
Bias %	131.00	0.22	0.00

GLM = Generalized Linear Model; TMLE = Targeted Maximum Likelihood Estimation; TMLEH = Targeted Maximum Likelihood Estimation using Handpicked algorithms; ATE = Average Treatment Effect (i.e., Effect Size) of a substitute in a previous period on the total distance of a soccer team.

Table 3. Observed dataset causal model.

Measure	GLM	TMLE	TMLEH
Correct causal model			
ATE	0.0105	0.0149	0.0142
Confidence Interval 95%	-0.0007-0.0216	0.0007-0.0290	-0.0021-0.0303
Miss-specified causal model			
ATE	0.0193	0.0245	0.0247
Confidence Interval 95%	-0.0007-0.0216	0.0115-0.0374	0.0210-0.0381
Difference correct causal model and miss-specified causal model			
Difference correct causal model and miss-specified	0.0089	0.0096	0.0121
Difference correct causal model and miss-specified %	84.7	65.0	66.3

GLM = Generalized Linear Model; TMLE = Targeted Maximum Likelihood Estimation; TMLEH = Targeted Maximum Likelihood Estimation using Handpicked algorithms; ATE = Average Treatment Effect (i.e., Effect Size) of a substitute in a previous period on the total distance of a soccer team.

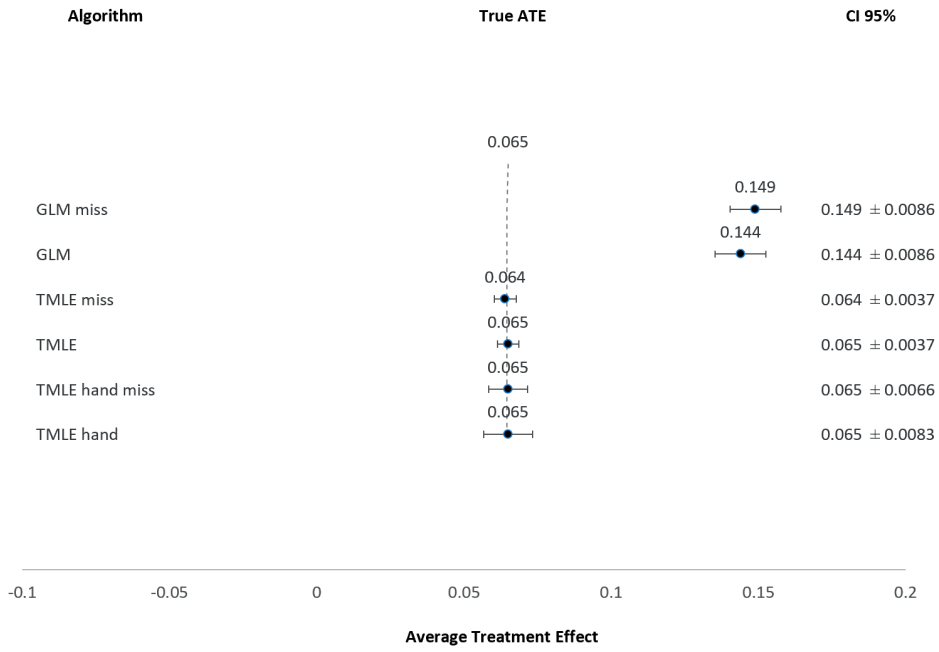


Figure 5. The Average Treatment Effect of the simulation of the causal model and the miss-specified causal model.
True ATE = True Average Treatment Effect (i.e. Effect Size) of a substitute in a previous period on the total distance of a soccer team. , CI 95% = Confidence Interval 95%; GLM miss = Generalized Linear Model with miss-specified causal model; GLM = Generalized Linear Model, TMLE miss = Targeted Maximum Likelihood Estimation with miss-specified causal model; TMLE = Targeted Maximum Likelihood Estimation; TMLE hand miss = Targeted Maximum Likelihood Estimation using Handpicked algorithms with miss-specified causal model; TMLE hand = Targeted Maximum Likelihood Estimation using Handpicked algorithms

5. DISCUSSION

We provided a roadmap as an approach for causal inference. The roadmap was applied to perform causal inference and examine on the one hand, the performance of TMLE, and on the other hand the accuracy in estimating the effect size between the traditional method GLM and the novel method TMLE. The comparison between GLM and TMLE was made by performing a simulation study on the effect of substitution on the total physical performance of a soccer team. We showed that GLM yields biased estimates of the effect size, whereas TMLE provides more accurate effect size estimations. These findings are consistent with earlier research [2], [11], [34].

Furthermore, we applied the causal roadmap using GLM and TMLE on observed elite soccer data. Our results indicate that a substitution in elite soccer increases the total team performance with 0.0105 to 0.01485 of the total distance covered. Other studies on performance, substitutes, and soccer also show that the performance of a substitute is higher when compared to an entire-match player [3], [4], [16] and that physical performance relates to overall game performance [14]. However, these studies leave out the influence of the substitutions and individual performance on the team performance. The causal roadmap provides a guide for causal inference. It helps to design statistical analyses, answering the causal question while making clear what assumptions are required to provide results with a causal interpretation[35]. Causal inference relates to statistical inference. Where causal inference means reasoning about causation, statistical inference means association reasoning with statistics. Statistical inference aims to assess parameters of a distribution from samples drawn from that distribution [27]. With the parameters, associations among variables and probabilities of future events can be inferred [27]. The associations and probabilities can be updated when new evidence or new data is available [27]. Causal inference aims to go one step further; the aim is to infer probabilities under static conditions and the dynamics of probabilities under changing conditions, for example, a substitution [27]. That is not to say that statistical inference cannot be used to establish causal relationships. Scientific explanations are an example of applying statistical inference, using, for instance, the Deductive-Nomological Model of Hempel and Oppenheim [36] applying laws to model statistical relevance designed to establish scientific explanations. Scientific explanations are causal explanations establishing a delicate relationship between statistical inference and causal inference. However, causal inference implies the dynamics of changing conditions where statistical inference does not. The combination of the causal roadmap and TMLE offers an opportunity to study the influence of a changing condition.

One limitation of the current study is our application of the causal roadmap. In the first step of this roadmap, it is important to state the knowledge one has about the

system under study. The aim of this paper is to introduce readers to TMLE and the causal roadmap. To reduce the complexity of the paper, we have reduced the complexity of the causal model by leaving out some possible time depending relations. We believe that this impact is low, but we would advise readers who are dealing with time-series data to look into TMLE methods that make use of time-series data.

TMLE is known as a double robust estimator, meaning that is it is consistent whenever the propensity score model is correctly specified or the outcome regression is correctly specified [6]. Although there are other double robust estimators methods like the Augmented Inverse Propensity Weighted (AIWP) Estimator, we limit ourselves to one method.

Van der Laan and Rose [2], compared different methods and found that Maximum likelihood estimation (MLE) based methods and estimating equations (IPTW and AIPTW) will underperform in comparison with TMLE. For we aimed to introduce causal inference and targeted learning in sport science, we choose to use the novel TMLE using machine learning and targeted learning.

In our experiments TMLE and TMLEH outperformed GLM for the observed data between the causal model and the miss-specified model. However, the difference in the effect size between the causal model and the miss-specified model was considerable for every method. The difference in effect size may be affected by the limited selection of the contextual factors. Since well-known contextual factors with an important influence on the physical performance, such as match location (home or away), score (win, draw or lose), rival level [7]–[9] were not available in our dataset and not taken into account. Therefore, our study does not fully meet the second assumption that there is no unmeasured confounding between treatment A and outcome Y, hence the use of the convenience assumption. In contrast, in the simulation study we have full control over the data generating distributions and their relations, and this study therefore allows us to fulfil the second assumption. Our goal with the simulation study is to show the applicability of the roadmap and TMLE to a practical problem, whilst having an objective means to compare the performance of TMLE to other methods. The double robustness of TMLE implies more resilience to endogeneity although the double robustness does not solve the endogeneity problem completely. In a study in pharmacoepidemiology, it is found that the more factors are taken into account, the better TMLE performs and becomes more independent of the treatment model specification [12]. When applying the complete set of factors, the outcomes were correct regardless of the treatment model specification [12]. In theory, when all factors are taken into account in the performance of a soccer team, TMLE will engage the true influence of a substitution.

6. CONCLUSION

Our study set out to provide a roadmap for causal inference and introduce the use of TMLE in sports science for other sports scientists. We applied the causal roadmap and showed that TMLE has a lower bias than GLM in a simulation setting both on the correct and the misspecified causal model. This result indicates that TMLE can be a more precise method than GLM in identifying and correctly estimating causal effects. Furthermore, when applying GLM and TMLE on the observed data on substitution, both methods found that the total physical performance improves when a substitution is made. However, the difference in the effect sizes between the correctly specified and the miss-specified model was considerable for TMLE and GLM. Furthermore, we showed that in these cases, TMLE was more precise than GLM.

7. PRACTICAL IMPLICATIONS

These findings show that the power of TMLE can help bring causal inference in sports science to the next level when more factors are taken into account. Future work will need to collect as much factor data as possible, enabling investigation of the influence of one factor in contrast with the traditional statistical methods where a selection of factors is made.

Funding
This research was partly funded by an SNN (Samenwerking Noord Nederland) MIT Grant under project code MITH20138.

Institutional Review Board Statement
The study was conducted according to the guidelines of the Declaration of Helsinki and approved by The Ethics Committee CTc UMCG of the University Medical Center Groningen, The Netherlands (protocol code: 201800430, 01/11/2018). 201800430

Informed Consent Statement
Informed consent was obtained from all subjects involved in the study.

Data Availability Statement
The data can be found on Github:https://github.com/dijkhuist/EntropyTMLE-Substitutions/tree/main/Data

Acknowledgments
The authors thank Prof. Dr. K.A.P.M. Lemmink, Prof. Dr. M. Aiello, Prof. Dr. H. Velthuijsen,

and Dr. M. Kempe for their valuable suggestions on improving the manuscripts' clarity.

Conflicts of Interest

The authors declare no conflict of interest. The funding provider had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

REFERENCES

[1] M. L. Meldrum, “A brief history of the randomized controlled trial: From oranges and lemons to the gold standard,” *Hematol. Oncol. Clin. North Am.*, vol. 14, no. 4, pp. 745–760, 2000, doi: 10.1016/s0889-8588(05)70309-9.

[2] M. J. van der Laan and S. Rose, *Targeted Learning*, vol. 20. New York, NY: Springer-Verlag New York, 2011.

[3] S. P. Hills *et al.*, “A match-day analysis of the movement profiles of substitutes from a professional soccer club before and after pitch-entry,” *PLoS One*, vol. 14, no. 1, pp. 1–15, 2019, doi: 10.1371/journal.pone.0211563.

[4] P. S. Bradley, C. Lago-Peñas, and E. Rey, "Evaluation of the match performances of substitution players in elite soccer," *Int. J. Sports Physiol. Perform.*, vol. 9, no. 3, pp. 415–424, 2014, doi: 10.1123/IJSPP.2013-0304.

[5] M. J. Van Der Laan, E. C. Polley, and A. E. Hubbard, “Super Learner,” *Stat. Appl. Genet. Mol. Biol.*, vol. 6, no. 1, 2007.

[6] A. N. Glynn and K. M. Quinn, “An introduction to the augmented inverse propensity weighted estimator,” *Polit. Anal.*, vol. 18, no. 1, pp. 36–56, 2009, doi: 10.1093/pan/mpp036.

[7] C. Lago, L. Casais, E. Dominguez, and J. Sampaio, “The effects of situational variables on distance covered at various speeds in elite soccer,” *Eur. J. Sport Sci.*, vol. 10, no. 2, pp. 103–109, 2010, doi: 10.1080/17461390903273994.

[8] J. Castellano, A. Blanco-Villaseñor, and D. Álvarez, "Contextual variables and time-motion analysis in soccer," *Int. J. Sports Med.*, vol. 32, no. 6, pp. 415–421, 2011, doi: 10.1055/s-0031-1271771.

[9] V. I. Kalapotharakos, A. Gkaros, and E. Vassliades, “Influence of contextual factors on match running performance in elite soccer team,” *J. Phys. Educ. Sport*, vol. 20, no. 6, pp. 3267–3272, 2020, doi: 10.7752/jpes.2020.s6443.

[10] A. S. Benjamin *et al.*, “Modern machine learning outperforms GLMs at predicting spikes,” *bioRxiv*, pp. 1–13, 2017, doi: 10.1101/111450.

[11] N. Kreif, S. Gruber, R. Radice, R. Grieve, and J. S. Sekhon, “Evaluating treatment effectiveness under model misspecification: A comparison of targeted maximum likelihood estimation with bias-corrected matching,” *Stat. Methods Med. Res.*, vol. 25, no. 5, pp. 2315–2336, 2016, doi: 10.1177/0962280214521341.

[12] M. Pang, T. Schuster, K. B. Filion, M. Eberg, and R. W. Platt, “Targeted maximum likelihood estimation for pharmacoepidemiologic research,” *Epidemiology*, vol. 27, no. 4, pp. 570–577, 2016, doi: 10.1097/EDE.0000000000000487.

[13] M. J. van der Laan and D. Rubin, “Targeted maximum likelihood learning,” *Int. J. Biostat.*, vol. 2, no. 1, 2006, doi: 10.2202/1557-4679.1043.

[14] T. Modric, S. Versic, D. Sekulic, and S. Liposek, “Analysis of the association between running performance and game performance indicators in professional soccer players,” *Int. J. Environ. Res. Public Health*, vol. 16, no. 20, 2019, doi: 10.3390/ijerph16204032.

[15] M. Kempe, M. Vogelbein, and S. Nopp, “The cream of the crop: Analysing FIFA world cup 2014 and Germany’s title run,” *J. Hum. Sport Exerc.*, vol. 11, no. 1, pp. 42–52, 2016, doi: 10.14198/jhse.2016.111.04.

[16] M. Lorenzo-Martínez, A. Padrón-Cabo, E. Rey, and D. Memmert, "Analysis of Physical and Technical

Performance of Substitute Players in Professional Soccer," *Res. Q. Exerc. Sport*, vol. 00, no. 00, pp. 1–8, 2020, doi: 10.1080/02701367.2020.1755414.

[17] K. Kelley and K. J. Preacher, "On effect size," *Psychol. Methods*, vol. 17, no. 2, pp. 137–152, 2012, doi: 10.1037/a0028086.

[18] E. Morgulev, O. H. Azar, and R. Lidor, "Sports analytics and the big-data era," *Int. J. Data Sci. Anal.*, vol. 5, no. 4, pp. 213–222, 2018, doi: 10.1007/s41060-017-0093-7.

[19] R. Rein and D. Memmert, "Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science," *Springerplus*, vol. 5, no. 1, 2016, doi: 10.1186/s40064-016-3108-2.

[20] A. Chambaz, I. Drouet, and J.-C. Thalabard, "Causality, a Trialogue," *J. Causal Inference*, vol. 2, no. 2, p. 41, Jan. 2014, doi: 10.1515/jci-2013-0024.

[21] F. J. Blaauw, "The non-existent average individual: Automated personalization in psychopathology research by leveraging the capabilities of data science," p. 294, 2018.

[22] M. L. Petersen and M. J. Van Der Laan, "Causal models and learning from data: Integrating causal modeling and statistical estimation," *Epidemiology*, vol. 25, no. 3, pp. 418–426, 2014, doi: 10.1097/EDE.0000000000000078.

[23] O. D. Duncan, *Introduction to Structural Equation Models*. New York, NY: Academic Press, 1975.

[24] J. Pearl and D. Mackenzie, *The Book of Why*. New York: Basic Books, 2018.

[25] S. Greenland and B. Brumback, "An overview of relations among causal modelling methods," *Int. J. Epidemiol.*, vol. 31, no. 5, pp. 1030–1037, 2002, doi: 10.1093/ije/31.5.1030.

[26] J. Ahern and A. E. Hubbard, *A Roadmap for Estimating and Interpreting Population Intervention Parameters*, 2nd ed. San Francisco: Jossey-Bass, 2017.

[27] J. Pearl, "Causal inference in statistics: An overview," *Stat. Surv.*, vol. 3, no. September, pp. 96–146, 2009, doi: 10.1214/09-SS057.

[28] R. Aquino *et al.*, "Influence of Situational Variables, Team Formation, and Playing Position on Match Running Performance and Social Network Analysis in Brazilian Professional Soccer Players," *J. strength Cond. Res.*, vol. 34, no. 3, pp. 808–817, 2020, doi: 10.1519/JSC.0000000000002725.

[29] D. Linke, D. Link, and M. Lames, "Validation of electronic performance and tracking systems EPTS under field conditions," *PLoS One*, vol. 13, no. 7, pp. 1–19, 2018, doi: 10.1371/journal.pone.0199519.

[30] M. J. van der Laan and S. Rose, *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. 2018.

[31] "Treatment effects, Effect sizes, and Point estimates." https://www.meta-analysis.com/pages/effects.php (accessed Apr. 25, 2022).

[32] J. M. Robins, "A New Approach To Causal Inference in Mortality Studies With a Sustained Exposure Period - Application To Control of the Healthy Worker Survivor Effect'.," *Math. Model.*, vol. 7, pp. 1393 – 1512, 1986, doi: 10.1016/0270-0255(86)9008-6.

[33] S. Gruber and M. J. Van Der Laan, "A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome," *Int. J. Biostat.*, vol. 6, no. 1, 2010, doi: 10.2202/1557-4679.1260.

[34] M. A. Luque-Fernandez, M. Schomaker, B. Rachet, and M. E. Schnitzer, "Targeted maximum likelihood estimation for a binary treatment: A tutorial," *Stat. Med.*, vol. 37, no. 16, pp. 2530–2546, 2018, doi: 10.1002/sim.7628.

[35] M. L. Petersen, “Applying a Causal Road Map in Settings with Time-dependent Confounding,” *Empidemiology*, vol. 25, no. 6, pp. 898–901, 2014, doi: 10.1117/12.2549369.Hyperspectral.

[36] C. G. Hempel and P. Oppenheim, “Studies in the Logic of Explanation,” *Philos. Sci.*, vol. 15, no. 2, pp. 135–175, 1948, doi: 10.1086/286983.

CHAPTER 5A

Increase in the Acute:Chronic Workload Ratio relates to injury risk in competitive runners

Based on
"Increase in the Acute:Chronic Workload Ratio Relates to Injury Risk in Competitive Runners"

Talko B. Dijkhuis
Ruby T.A. Otter
Marco Aiello
Hugo Velthuijsen
Koen A.P.M. Lemmink

2020. International Journal of Sports Medicine 41(11):736–43.
doi: 10.1055/a-1171-2331.

ABSTRACT

Injuries of runners reduce the ability to train and hinder competing. Literature shows that the relation between potential risk factors and injuries are not definitive, limited, and inconsistent. In team sports, workload derivatives were identified as risk factors. However, there is an absence of literature in running on workload derivatives. This study used the workload derivatives acute workload, chronic workload, and acute:chronic workload ratios to investigate the relation between workload and injury risk in running. Twenty-three competitive runners kept a daily training log for 24 months. The runners reported training duration, training intensity and injuries. One week (acute) and 4-week (chronic) workloads were calculated as the average of training duration multiplied by training intensity. The acute:chronic workload ratio was determined dividing the acute and chronic workloads. Results show that a fortnightly low increase of the acute:chronic workload ratio (0.10–0.78) led to an increased risk of sustaining an injury ($p < 0.001$). Besides, a low increase of the acute:chronic workload ratio (0.05–0.62) between the second week and third week before an injury showed an association with increased injury risk ($p=0.013$). These findings demonstrate that the acute:chronic workload ratio relates to injury risk.

Keywords
Injury and prevention, prediction, rating of perceived exertion.

1. INTRODUCTION

The yearly occurrence of time-loss injuries in middle-distance runners (64%), long-distance runners (32%), and marathon runners (52%) is high[1]. Most of the injuries are associated with overuse [1], [2]. Injuries lead to a reduced training effort and the inability to compete, which can be detrimental to the career of competitive runners. Therefore, prevention of injuries is important. The Translating Research into Injury Prevention Practice (TRIPP) framework is built on the fact that the professionals only adopt the results of injury research when it helps preventing injuries [3]. The TRIPP framework defines six consecutive steps for research in building the evidence base for the prevention of injuries [3]. The first step within the TRIPP framework is to undertake injury surveillance. The second step is to identify risk and protective factors and injury mechanisms. The third step is to develop preventive measures. The fourth step is creating ideal conditions for scientific evaluation of the preventive measures. The fifth step is the description of the intervention context and development of implementation strategies and the sixth and final step is to implement the intervention in context and evaluate the effectiveness.

Despite an extensive body of research on identifying risk factors, to the best of our knowledge, there is the no single study that reveals modifiable risk factors in running enabling the third step of TRIPP: development of preventive measures. In the literature, there is consensus on two nonmodifiable risk factors in runners: (i) a history of running injuries and (ii) an irregular and/or absent menstruation for female runners [4], [5]. For many proposed modifiable risk factors in running, like distance, duration, frequency, pace, interval, weight, and footwear, there is an absence of clear support for an association with injury risk [4], [5]. Although workload and changes in workload are mentioned as modifiable risk factor in runners, and adjustment of the workload may prevent overuse injuries, the results on the relationship between workload as a single nonrelative factor and injuries in running are ambiguous, limited, and even inconsistent [4], [6], [7].

In contrast to the studies on running, a clear relationship between workload and injuries was identified in competitive team sports, such as Australian football [8]–[10], rugby [11], cricket [12], and soccer [13]. These studies found an association between an increase in the relative workload and the risk on sustaining an injury in the same or subsequent week. The relative workload was calculated as a rolling average (RA) of the acute workload in relation to the RA of the chronic workload (acute:chronic ratio). In contrast to the acute:chronic ratio, the acute and chronic workloads in isolation (i.e., not as ratios) was not consistently associated with increased injury risk [11].

Although in literature different time periods are designated as acute and chronic workload, for the acute:chronic ratio most commonly one week of workload (acute

workload) compared with a four week workload (chronic workload) is reported [14]. There's a discussion whether RA or exponentially weighted moving averages (EWMA) are more suitable to use in the acute:chronic ratio [15], [16]. Itis found in elite Australian Football that EWMA in higher ratio's (>2.0) may be a more sensitive indicator [14], [15]. Although both RA and EWMA correctly identify increased injury risk [14], [15]. In the afore mentioned studies in team sports[8]–[13] the calculation of acute and chronic workloads were mathematically coupled, i.e., the acute workload is contained in the chronic workload, and are spuriously correlated [17]. A solution is to uncouple i.e. the acute workload is not included in the chronic workload [17]. However in practice both coupled and uncoupled lead to the same results [18]. Many studies take measures of external and internal workload, into account in the calculation of the acute:chronic ratio [11], [12], [19]. While external workload defined as the work completed independently of internal characteristics [20] (i.e. duration, distance, number of throws, speed) is significant in comprehending the physical effort of the athlete, the internal workload, or the relative physiological and psychological stress is essential in determining the workload [21]. Foster et al. (2001) proposed a monitoring tool for training load based on rating of perceived exertion (RPE) [22]. This method, known as session-RPE method (sRPE), takes into account both the intensity and the duration [22]. The combination of intensity and duration is sRPE is a valid stand-alone tool for both training and competition to calculate the workload [22], [23]. Although applying sRPE in combination with the acute:chronic workload ratio (ACWR) may be promising for identification of the impact of workload on injury risk [24], there is an absence of studies that relates sRPE based ACWR with injury risk in running. Previous studies in running on workload and injury risk defined workload as a single nonrelative factor, like duration, distance or frequency [6], [7], [11]. The aim of the present study is to investigate the sRPE based acute workload, chronic workload, ACWR, and week-to-week and fortnightly ACWR difference as modifiable risk factors, in relation to injury risk of competitive runners.

2. MATERIALS AND METHODS

2.1. Participants

A group of 23 competitive runners (16 male, 7 female) of the same training group and the same coach participated in the study during a period of 24 months. The runners competed in race distances of 800 meters to marathon on regional (5 runners), national (15 runners), and international (3 runners) level. Table 1 shows the runners' baseline characteristics. Written informed consent was obtained from all individual runners participating in the study. The ethics committee of University Medical Center Groningen, the Netherlands (METc 2011/186), approved the research protocols.

Table 1. Baseline characteristics of the runners.

	Male	Female	Total
Number	16	7	23
Age (years; mean ± SD)	22.5 ± 6.3	21.4 ± 4.4	22.2 ± 5.7
Height (cm; mean ± SD)	185 ± 5	172 ± 7	181 ± 8
Body weight (kg; mean ± SD)	68.6 ± 6.0	58.3 ± 4.0	65.4 ± 7.2
Percentage body fat* (%; mean ± SD)	8.5 ± 2.3	17.6 ± 4.2	11.3 ± 5.2
VO_2max**(ml/kg/min; mean ± SD)	66.7 ± 5.9	62.7 ± 7.4	65.5 ± 6.5

SD = Standard Deviation; cm = centimetre; kg = kilogram; ml = millilitre; min = minutes; VO_2max = Maximal measured Oxygen Uptake
*The percentage body fat was estimated using the Tanita BC 418.
**The VO_2max was measured with a maximal incremental treadmill test including breath-by-breath gas analysis using the Cortex Metalyzer 3 B.

2.2. Definition of injury

An injury was defined as any musculoskeletal problem of the lower extremity or back that led to an inability to execute training or competition as planned for at least one week [25]. Only injuries sustained as a result of training or competition were considered. Recovery from an injury was defined as the ability to complete the normal training schedule. At the start of the study, the runners filled out a validated questionnaire on injury history based on Fuller et al. [26]. During the study period of 24 months, the runners kept a daily log on sustained injuries. The coach added information about the observed injuries to this log.

2.3. Quantifying workloads

The coach developed a training and competition schedule for each runner and recorded the ability of the individual runner to execute the planned schedule. Each individual runner filled out their daily training and competition schedule for duration and intensity of all training sessions and running competition events. The training sessions consisted of various types of training, for example endurance training, technique training, and strength training. The duration of the training and competition sessions was reported in minutes. In addition, the intensity was determined by the rating of perceived exertion per session (sRPE), which was reported by the runners approximately 30 minutes after each session on the Borg Scale ranging from 6 to 20 [27]. The workload of each training session and competition event was calculated by multiplying the sRPE scores with the duration and was expressed in arbitrary units (AU).

2.4. Data analysis

Data of one runner was removed from the data set for not adequately recording duration and intensity. The remaining data on workload were divided into weekly blocks from Monday to Sunday. The weekly blocks represent the acute workload. The chronic

workload was calculated as the four week rolling average of the acute workload [9], [11], [12], [19], [28]. The ACWR was determined dividing the acute workload by the chronic workload (the coupled approach), indicating the relative size of acute workload compared to the chronic workload [9], [11], [12], [19], [28]. An ACWR below one represents an acute workload that is lower than the chronic workload. Conversely, an ACWR value above one represents an acute workload, which is higher than the chronic workload. The first four weeks of the study, the weeks in which runners were injured, as well as the four weeks after recovery from the injury were removed from the analysis of the ACWR and the chronic workload [12]. It is only after four weeks of normal workload that the chronic workload represents a non-biased value with respect to the injury occurrence [12]. Removing the weeks in which runners were injured created a separation between the load calculation window and the injury risk window [29], [30]. Subsequently the injury lag period was generalized to a risk window of a seven day period. Figure 1 shows a visualization of a runners' ACWR for the 24 months, a sustained injury, and the corresponding recovery period to illustrate the influence of a very low chronic workload on the ACWR.

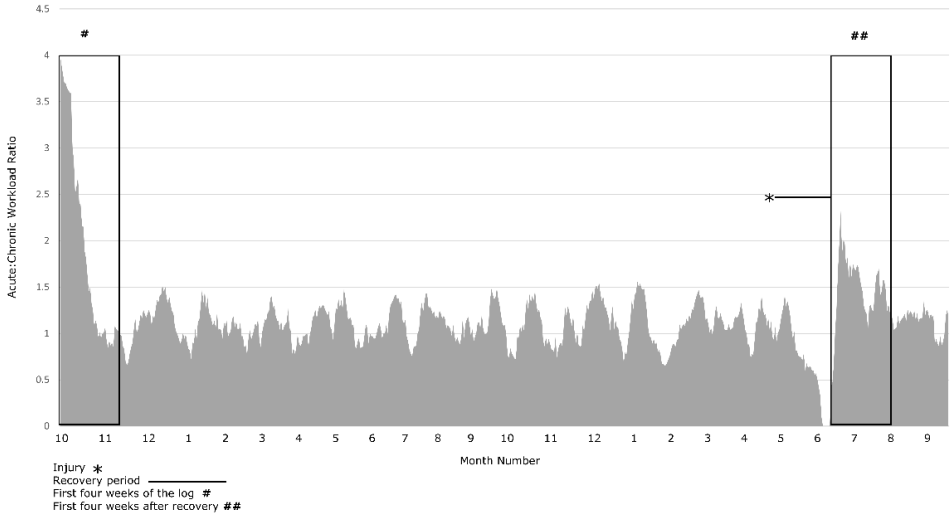


Figure 1. Visualisation of a runners 24 months ACWR containing an injury and corresponding recovery period, with a biased ACWR due to a biased chronic workload ratio in the first four weeks of the log and four weeks after a recovery period.

The normality of the distribution of the acute workload, the chronic workload, the ACWR and differences between the ACWR were tested. For all statistical analysis, we used IBM SPSS 2.4, unless indicated otherwise.

The z-score for the acute workload, the chronic workload and the weekly ACWR of the individual runners were calculated to indicate whether the observed value was above or below the average for the individual. The acute workload, chronic workload and the ACWR were classified accordingly [11], [12]. The classifications consisted of the following week categories: (i) Very low, (ii) Low (iii) Moderate low, (iv) Moderate high, (v) High, (vi) Very high. The thresholds of categories based on the corresponding the z-scores are presented in Table 2.

Table 2. Workload classifications and boundaries.

Workload category	z-Score	Acute Workload AU	Chronic Workload AU	Weekly ACWR ratio
Very low	*≤ -2.00*	*≤ 3810*	*≤ 6297*	*≤ 0.24*
Low	*-1.99 – -1.00*	*3811 – 8170*	*6298 – 9158*	*0.25 – 0.68*
Moderate Low	*-0.99 – -0.01*	*8171 – 10880*	*9159 – 10832*	*0.69 – 1.10*
Moderate High	*0.00 – 0.99*	*10881 – 14998*	*10833 – 13485*	*1.11 – 1.53*
High	*1.00 – 1.99*	*14999 – 18052*	*13486 – 19675*	*1.54 – 1.96*
Very High	*≥ 2.00*	*≥ 18053*	*≥ 19676*	*≥ 1.96*
Workload difference category	**z-Score**	**Weekly ACWR ratio difference**	**Fortnightly ACWR ratio difference**	
High decrease	*≤-2.00*	*≤ -0.57*	*≤ -0.53*	
Moderate decrease	*-1.99 – -1.00*	*-0.56 – -0.24*	*-0.56 – -0.27*	
Low decrease	*-0.99 – -0.01*	*-0.25 – 0.05*	*-0.27 – 0.10*	
Low increase	*0.00 – 0.99*	*0.05 – 0.62*	*0.10 – 0.78*	
Moderate increase	*1.00 – 1.99*	*0.63 – 1.14*	*0.79 – 1.29*	
High increase	*≥ 2.00*	*≥ 1.15*	*≥ 1.30*	

AU=Arbitrary Units
ACWR = Acute:Chronic Workload Ratio

The week-to-week ACWR difference is the difference of the ACWR between two subsequent weeks. The fortnightly ACWR difference is the difference between the average of the ACWR of two subsequent weeks compared with average of the ACWR of the following two subsequent weeks. The week-to-week and fortnightly ACWR difference were categorized in the following week-to-week and fortnightly AWCR difference categories:(i)High decrease,(ii) Moderate decrease, (iii) Low decrease, (iv) Low increase, (v) Moderate increase, (vi) High increase [11, 12]. The thresholds of the weekly and fortnightly difference categories were based on the distribution of the z-scores. The weekly and fortnightly difference categories and the corresponding thresholds are presented in Table 2.

2.5. Determining association

The association between workload and injury risk was determined for the workload categories related to the four-week blocks preceding the injury. The risk of sustaining

an injury was calculated using a binary logistic regression model that modelled acute workload week categories, chronic workload week categories, the ACWR week categories, the week-to-week and fortnightly ACWR difference categories as independent variables and injury/no injury as dependent variable. The 'Low' week category and the 'Low decrease' ACWR difference category were the reference categories. The data were statistically analysed using R version 3.4.4 (R Foundation for Statistical Computing, Vienna, Austria) and the caret library, version 6.0.79

2.6. Determining relative risk and prediction

A two-by-two table was used to determine basics for the metrics [31]. The two-by-two table consists of four categories: (i) True Positive (TP; i.e., the support for the identified associated categories in relation with the injury incidence), (ii) True Negative (TN; i.e., the support for the non-associated categories and the non-injury incidence), (iii) False Positive (FP; i.e., the support for the identified associated categories which did not result in an injury), and (iv) False Negative (FN; i. e., the support for the non-associated categories which resulted into an injury). The metrics for injury occurrence were the relative risk (RR), the standard error (SE) of log RR, the 95 % confidence interval (CI 95%), and the p-value of the relative risk. The RR, its SE, the CI 95% and p-value were calculated accordingly [31], [32]. The predictive power of the significant workload variables and the affiliated categories were calculated by the sensitivity and specificity [33]. The relative risk was calculated as

$$\text{RR} = \frac{TP/(TP+FP)}{FN/(FN+TP)},$$

for which the SE of the log of the RR can be calculated as

$$\text{SE}\{\ln(\text{RR})\} = \sqrt{\frac{1}{TP} + \frac{1}{FN} - \frac{1}{TP+FP} - \frac{1}{TN+TP}}.$$

When a category caused a division by zero in calculation of the RR or the SE, 0.5 was added to all four categories of the two-by-two table [3]. We calculated the 95% CI as $ln\,ln\,(RR) \pm 1.96 * SE\{ln\,ln\,RR\}$. We determined the p-value with the calculated z-value, z-value = $\frac{lnln\,(RR)}{SE\{lnln\,(RR)\}}$. Finally, we calculated the sensitivity and specificity. The sensitivity was calculated as the proportion of correctly identified injuries, as sensitivity = $\frac{TP}{TP+FN}$. The specificity was calculated as the proportion of correctly identified non-injuries, as specificity = $\frac{TN}{TN+FP}$. The calculations were performed using Microsoft Excel2016. We confirm the study meets the ethical standards of the International Journal of Sports Medicine [34].

3. RESULTS

3.1. Workload

The 22 runners conducted 13046 training sessions with a total number of 20139 training hours. The average weekly training hours were 8.9±4.6 and the average duration of a training session was 77.6±39.3 minutes. The session RPE was 12.3±3.1 on the Borg scale, the workload per session was 1031±661 AU, the daily workload was 1241±815 AU. The acute workload per week was 6801±3675 AU, the chronic workload per week was 6750±3185 and the overall corresponding ACWR was 0.99±0.47. The descriptive statistics for the 22 runners' workload variables split between the weeks preceding the injury and the weeks not preceding an injury are presented in Table 3. Excluding the first four weeks of the study, the weeks in which runners were injured, and the four weeks after recovery from the injury, reduced the number of training weeks by 25.9%, i.e., from 2066 to 1530 weeks of the data set. The frequency distributions of the variables are to be found in Table 4.

Table 3. Descriptive statistics for all runners' workload variables.

Workload					.		
	Weeks preceding an injury					Weeks without an injury	Difference between average pre-injury and non-injury
	Week-4	Week-3	Week-2	Week-1	Average	Average	p-value
Chronic	6401 ± 2301	7378 ± 2685	7050 ± 2382	6850 ± 2360	6920 ± 2433	6772 ± 3195	0.502
Acute	7238 ± 3291	7014 ± 3228	8099 ± 3986	7163 ± 2778	7379 ± 3321	6791 ± 3695	0.484
Acute:chronic	0.99 ± 0.32	1.05 ± 0.35	1.12 ± 0.35	1.15 ± 0.19	1.08 ± 0.30	1.11 ± 0.44	0.601

All data are mean ± Standard Deviation

Table 4. Frequency workload classifications and boundaries.

Workload category	z-Score	Acute Workload AU	Chronic workload AU	Weekly ACWR ratio
Very low	*≤ -2.00*	0	43	59
Low	-1.99--1.00	193	171	77
Moderate low	-0.99--0.01	589	556	621
Moderate high	0.00 – 0.99	533	541	734
High	1.00 – 1.99	173	181	19
Very high	*≥ 2.00*	42	38	20

Workload difference category	z-Score	Week-to-week ACWR difference	Fortnightly ACWR difference
High decrease	*≤ -2.00*	22	40
Moderate decrease	-1.99--1.00	55	78
Low decrease	-0.99--0.01	623	578
Low increase	0.00 – 0.99	761	714
Moderate increase	1.00 – 1.99	61	98
High increase	*≥ 2.00*	8	22

AU = Arbitrary Units, ACWR = Acute:Chronic Workload Ratio

3.2. Injuries

During the 24 months, 21 runners sustained one or more injuries (Table 5). Initially, 57 injuries were identified with an average injury rate of 3.6/1000 h. Four injuries skewed the mean recovery time, accounting for 1002 recovery days out of 3247 recovery days.

Table 5. Overview of the injuries.

Injuries	Male	Female	Total
Runners with no injuries (Frequency)	1	0	1
Runners with one injury (Frequency)	2	2	4
Runners with two injuries (Frequency)	5	2	7
Runners with three injuries (Frequency)	3	0	3
Runners with four injuries (Frequency)	4	2	6
Runners with five injuries (Frequency)	0	0	0
Runners with six injuries (Frequency)	1	0	1
Injury location (back/hip/knee/calf-Achilles/ankle-foot) (Frequency)	3/5/9/17/7	2/1/0/7/7	5/6/9/24/14
Time to recovery (days; median(range))	48(7-201)	77(9-306)	56 (7-306)

3.3. Association

There were no associations ($P<0.05$) between the acute workload, the chronic workload or the weekly ACWR and the injury risk. However, two ACWR difference categories showed significant associations with the injury risk: (i) the fortnightly ACWR difference

category ‘Low increase’ (p<0.001) and (ii) the week-to-week ACWR difference category ‘Low increase’ between week three and two before an injury (p=0.013) (Table 6).

Table 6. Binary logistic regression on difference categories of the Acute:Chronic Workload Ratio before the injury.

Difference categories (reference difference category Low decrease)				
	p-value			
	Week -3-4	**Week -2-3**	**Week -1-2**	**Fortnightly: Week -12-34**
High decrease	0.992	0.987	0.451	0.988
Moderate decrease	0.791	0.682	0.897	0.257
Low increase	0.125	0.013*	0.877	0.001*
Moderate increase	0.791	0.494	0.897	0.174
High increase	0.990	0.991	0.989	0.991

*Significant difference (p < 0.05) between the periods preceding an injury and the periods not preceding an injury.

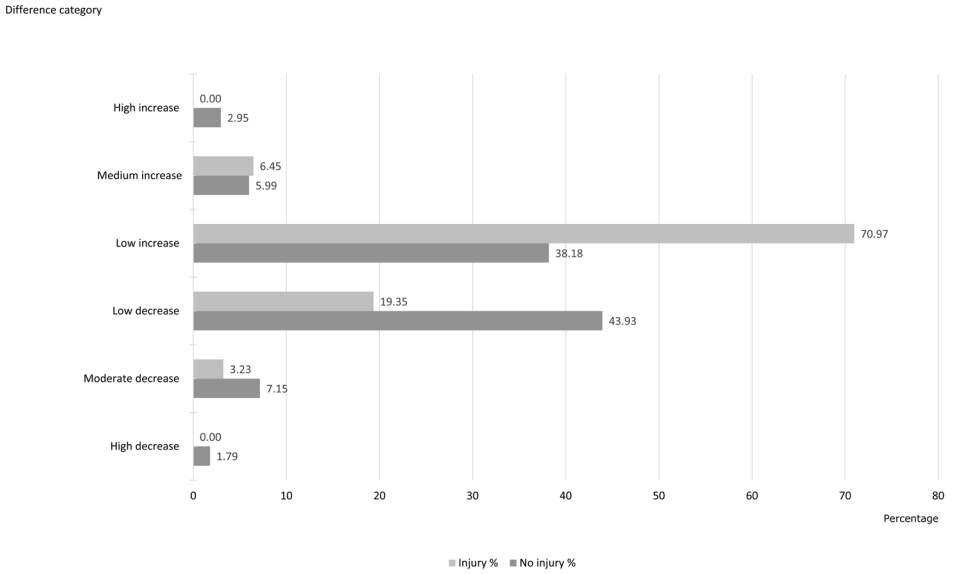


Figure 2. Distribution of the difference categories of the fortnightly ACWR, comparing the period before the injury with the remaining weeks without resulting in an injury.

3.4. Relative Risk and prediction

Fortnightly and between week two and three the ACWR difference category 'Low increase' was associated with the risk on sustaining an injury. The runners sustaining an injury with the fortnightly ACWR difference category 'Low increase' had a RR of injury of 4.49 (CI 95%: 2.02 – 9.96, $p < 0.000$). The relative risk for sustaining an injury with the weekly ACWR difference category 'Low increase' between week two and three was 2.74 (CI 95%:1.30 – 5.76, $p = 0.012$). In terms of percentage, the ACWR difference category 'Low increase' is overrepresented in the four-week period before the injury in comparison with the periods not preceding an injury. Figure 2 illustrates the distribution of the fortnightly AWCR difference categories comparing the period before the injury with the period without injuries. The predictability of an injury expressed by the specificity and sensitivity is limited. The fortnightly AWCR difference category 'Low increase' has specificity of 0.62 and a sensitivity of 0.74. Where the AWCR difference category 'Low increase' between week two and three has a specificity of 0.57 and a sensitivity of 0.68.

4. DISCUSSION

The current study expressed the workload in running as the combination of duration and RPE and investigated the association between the acute workload, the chronic workload, the ACWR, the change in the ACWR and the injury risk in competitive runners. We did not find an association between the acute workload, the chronic workload, the ACWR and injury risk. However, a 4.5-fold increase in injury risk was associated with low increase (0.10–0.78) of the fortnightly AWCR difference. Also, a 2.7-fold increase in injury risk was demonstrated for a low increase (0.05–0.62) of the week-to-week ACWR difference between week three and two before an injury. These findings suggest that there is an association between increased ACWR and the risk of sustaining an injury.

The injury incidence of 3.6/1000 h was comparable to previous studies on competitive runners that found incidences from 2.5–7.4 injuries per 1000 h for long-distance runners [35], [36] and 5.6–5.8 for sprinters and middle-distance runners [36]. Conform literature most injuries in the current study were reported in the calf-Achilles region [23], [27].

A main difference between the literature on running and injury risk and this study is the definition of workload. Previous studies in running defined workload as a single nonrelative factor, like duration, distance, or frequency [4, 6, 7], whereas we applied a combination of duration and RPE, the sRPE [23]. The sRPE was expressed in acute and chronic workload. The current research did not show an association between acute or chronic workload and injury risk. Based on the literature on running one cannot draw a conclusion on the relationship between a single nonrelative workload factor and

injuries risk [4], [6], [7]. The reason for not identifying a relationship in both literature and our study might be found in the employed method of using a nonrelative factor. This emphasizes the importance of including relative measures to a runner's individual training progression. Therefore, this study also used a relative measure (ACWR). The ACWR as a single factor was not associated with injury risk although other studies showed that spikes in the acute workload, is associated with an increase in injury risk in Australian football [9], rugby [11] and cricket [12] . Contrary to these studies, there were relatively few spikes in the current training data set. In other words, the competitive runners in this study were not regularly exposed to a high increase of acute workloads. Absence of spikes in our dataset does not rule out that there is an association, but we were not able to study this phenomenon when using the acute workload and the related ACWR.

In contrast with previous studies in competitive running using average load to identify injury risk, the present study was the first to take the change in the relative workload into account. The study demonstrated an association between an increase in the fortnightly and week-to-week ACWR difference and injury risk. This is consistent with the studies in Australian football, rugby, cricket, and soccer [9]–[13], [19], [37].

A notable finding is the delay of two weeks between the increase of the fortnightly and week-to-week ACWR difference and the injury manifestation. A similar observation was made in cricket and Australian football [9], [12]. Those studies showed an increase in injuries the subsequent week after a high increase of workload. A possible explanation for the difference in delay is the occurrence of spikes in the week before the injury in their study whereas in our study a more cumulative overloading took place.

Although an increased risk of sustaining an injury was found, the predictive value of the increase of the fortnightly and week-to-week ACWR difference is low. The fortnightly AWCR difference category 'Low increase' had a specificity of 0.62 and a sensitivity of 0.74. The week-to-week AWCR difference category 'Low increase' between week three and two before an injury had a specificity of 0.57 and a sensitivity of 0.68.The low specificity and low sensitivity illustrates that the 'Low increase' of the fortnightly and week-to-week AWCR difference, though an important signal for an increase in injury risk, is insufficient as a single predictor of an injury. This is consistent with Carey et al. [29] and Fanchini et al. [37] where objective, subjective and relative measures proved to have poor ability to predict an injury. Another limitation of the study is the calculation of the ACWR. The ACWR is only an unbiased measure after 28 days of completing a normal training schedule. Therefore, the first four weeks of data at the start of the running season, the first four weeks of data after recovery, and the data of the rehabilitation period could not be used for monitoring ACWR. The removal of the first four weeks

of data of the running season prevented studying possible influences of the start of a season.

Another limitation is the removal of the first four weeks after recovery along with the removal of the rehabilitation period. This removal eliminated the possibility to study the influence of the possible workload difference between rehabilitation training and regular training. Although when the ACWR is looked at in an elite training setting, the assessment of the ACWR during recovery can be an indicator whether an athlete is prepared well enough to enter a normal training schedule [38].

From our study, we conclude that the ACWR is a useful measure to identify an increased injury risk in competitive running. The ACWR could be taken into account when designing training schedules, observing the ability to execute the planned training schedule, and monitoring the ACWR recorded by the runner. Timely identification of an increase of the ACWR may enable timely preventive measures decreasing the injury risk in runners.

Funding

This work was supported by SIA RAAK-PRO under Grant[PRO-2-018] and [TOP.UP01.008]

REFERENCES

[1] B. Kluitenberg, M. van Middelkoop, R. Diercks, and H. van der Worp, “What are the Differences in Injury Proportions Between Different Populations of Runners? A Systematic Review and Meta-Analysis,” *Sport. Med.*, vol. 45, no. 8, pp. 1143–1161, Aug. 2015, doi: 10.1007/s40279-015-0331-x.

[2] A. Hreljac, “Impact and Overuse Injuries in Runners,” *Med. Sci. Sport. Exerc*, vol. 36, no. 5, pp. 845–849, 2004, doi: 10.1249/01.MSS.0000126803.66636.DD.

[3] C. Finch, “A new framework for research leading to sports injury prevention,” *Journal of Science and Medicine in Sport*, vol. 9, no. 1–2. Elsevier, pp. 3–9, May 01, 2006, doi: 10.1016/j.jsams.2006.02.009.

[4] A. Hulme, R. O. Nielsen, T. Timpka, E. Verhagen, and C. Finch, “Risk and Protective Factors for Middle- and Long-Distance Running-Related Injury,” *Sport. Med.*, vol. 47, no. 5, pp. 869–886, 2017, doi: 10.1007/s40279-016-0636-4.

[5] B. T. Saragiotto, T. P. Yamato, L. C. Hespanhol Junior, M. J. Rainbow, I. S. Davis, and A. D. Lopes, “What are the Main Risk Factors for Running-Related Injuries?,” *Sport. Med.*, vol. 44, no. 8, pp. 1153–1163, Aug. 2014, doi: 10.1007/s40279-014-0194-6.

[6] C. Damsted, S. Glad, R. O. Nielsen, H. Sørensen, and L. Malisoux, "Is there evidence for an association between changes in training load and running-related injuries? A systematic review.," *Int. J. Sports Phys. Ther.*, vol. 13, no. 6, pp. 931–942, Dec. 2018, Accessed: Apr. 11, 2019. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/30534459.

[7] R. Ø. Nielsen, I. Buist, H. Sørensen, M. Lind, and S. Rasmussen, "Training Errors and Running Related Injuries: a Systematic Review," *Int. J. Sports Phys. Ther.*, vol. 7, no. 1, pp. 58–75, 2012.

[8] A. Esmaeili, W. G. Hopkins, A. M. Stewart, G. P. Elias, B. H. Lazarus, and R. J. Aughey, “The individual and combined effects of multiple factors on the risk of soft tissue non-contact injuries in elite team sport athletes,” *Front. Physiol.*, vol. 9, no. SEP, pp. 1–16, 2018, doi: 10.3389/fphys.2018.01280.

[9] N. B. Murray, T. J. Gabbett, A. D. Townshend, B. T. Hulin, and C. P. Mclellan, “Individual and combined effects of acute and chronic running loads on injury risk in elite Australian footballers,” *Scand. J. Med. Sci. Sport.*, no. 2007, pp. 1–9, 2016, doi: 10.1111/sms.12719.

[10] J. D. Ruddy, C. W. Pollard, R. G. Timmins, M. D. Williams, A. J. Shield, and D. A. Opar, “Running exposure is associated with the risk of hamstring strain injury in elite Australian footballers,” *Br. J. Sports Med.*, p. bjsports-2016-096777, 2016, doi: 10.1136/bjsports-2016-096777.

[11] B. T. Hulin, T. J. Gabbett, D. W. Lawson, P. Caputi, and J. a Sampson, “The acute:chronic workload ratio predicts injury: high chronic workload may decrease injury risk in elite rugby league players,” *Br. J. Sports Med.*, vol. 50, no. 4, pp. 231–236, 2016, doi: 10.1136/bjsports-2015-094817.

[12] B. Hulin *et al.*, “Spikes in acute workload are associated with increased injury risk in elite cricket fast bowlers,” *Artic. Br. J. Sport. Med.*, 2013, doi: 10.1136/bjsports-2013-092524.

[13] A. Jaspers, J. P. Kuyvenhoven, F. Staes, W. G. P. Frencken, W. F. Helsen, and M. S. Brink, “Examination of the external and internal load indicators’ association with overuse injuries in professional soccer players,” *J. Sci. Med. Sport*, vol. 21, no. 6, pp. 579–585, 2018, doi: 10.1016/j.jsams.2017.10.005.

[14] A. Griffin, I. C. Kenny, T. M. Comyns, and M. Lyons, “The Association Between the Acute:Chronic Workload Ratio and Injury and its Application in Team Sports: A Systematic Review,” *Sport. Med.*, vol. 50, no. 3, pp.

561–580, 2019, doi: 10.1007/s40279-019-01218-2.

[15] N. B. Murray, T. J. Gabbett, A. D. Townshend, and P. Blanch, “Calculating acute: Chronic workload ratios using exponentially weighted moving averages provides a more sensitive indicator of injury likelihood than rolling averages,” *Br. J. Sports Med.*, vol. 51, no. 9, pp. 749–754, 2017, doi: 10.1136/bjsports-2016-097152.

[16] P. Menaspà, "Are rolling averages a good way to assess training load for injury prevention?," *Br. J. Sports Med.*, vol. 51, no. 7, pp. 618–619, 2017, doi: 10.1136/bjsports-2016-096131.

[17] L. Lolli *et al.*, “Mathematical coupling causes spurious correlation within the conventional acute-to-chronic workload ratio calculations,” *British journal of sports medicine*, vol. 53, no. 15. pp. 921–922, 2019, doi: 10.1136/bjsports-2017-098110.

[18] T. J. Gabbett, B. Hulin, P. Blanch, P. Chapman, and D. Bailey, “To Couple or not to Couple? for Acute:Chronic Workload Ratios and Injury Risk, Does it Really Matter?,” *Int. J. Sports Med.*, vol. 40, no. 9, pp. 597–600, 2019, doi: 10.1055/a-0955-5589.

[19] S. Malone, A. Owen, M. Newton, B. Mendes, K. D. Collins, and T. J. Gabbett, “The acute:chonic workload ratio in relation to injury risk in professional soccer,” *J. Sci. Med. Sport*, vol. 20, no. 6, pp. 561–565, 2017, doi: 10.1016/j.jsams.2016.10.014.

[20] L. K. Wallace, K. M. Slattery, and A. J. Coutts, “The ecological validity and application of the session-rpe method for quantifying training loads in swimming,” *J. Strength Cond. Res.*, vol. 23, no. 1, pp. 33–38, Jan. 2009, doi: 10.1519/JSC.0b013e3181874512.

[21] S. L. Halson, “Monitoring Training Load to Understand Fatigue in Athletes,” *Sports Medicine*, vol. 44, no. Suppl 2. Springer, pp. 139–147, Nov. 2014, doi: 10.1007/s40279-014-0253-z.

[22] C. Foster *et al.*, “A New Approach to Monitoring Exercise Training,” *J. Strength Cond. Res.*, vol. 15, no. 1, pp. 109–115, 2001, doi: 10.1519/00124278-200102000-00019.

[23] M. Haddad, G. Stylianides, L. Djaoui, A. Dellal, and K. Chamari, “Session-RPE method for training load monitoring: Validity, ecological usefulness, and influencing factors,” *Frontiers in Neuroscience*, vol. 11, no. NOV. Frontiers Media SA, p. 612, 2017, doi: 10.3389/fnins.2017.00612.

[24] R. Johnston, R. Cahalan, M. O’Keeffe, K. O’Sullivan, and T. Comyns, “The associations between training load and baseline characteristics on musculoskeletal injury and pain in endurance sport populations: A systematic review,” *J. Sci. Med. Sport*, vol. 21, no. 9, pp. 910–918, Sep. 2018, doi: 10.1016/J.JSAMS.2018.03.001.

[25] S. W. Bredeweg, S. Zijlstra, and I. Buist, “The GRONORUN 2 study: effectiveness of a preconditioning program on preventing running related injuries in novice runners. The design of a randomized controlled trial,” *BMC Musculoskelet Disord*, vol. 11, p. 196, 2010, doi: 10.1186/1471-2474-11-196.

[26] C. W. Fuller *et al.*, “Consensus statement on injury definitions and data collection procedures in studies of football (soccer) injuries.,” *Br. J. Sports Med.*, vol. 40, no. 3, pp. 193–201, Mar. 2006, doi: 10.1136/bjsm.2005.025270.

[27] A. Borg, “Psychophysical Bases of Perceived Exertion.,” *Med. Sci. Sport. Exerc.*, vol. 14, pp. 377–381, 1982.

[28] T. J. Gabbett, B. T. Hulin, P. Blanch, and R. Whiteley, “High training workloads alone do not cause sports injuries: how you get there is the real issue,” *Br. J. Sports Med.*, vol. 50, no. 8, pp. 1–2, 2016, doi: 10.1136/

bjsports-2015-095567.

[29] D. L. Carey, P. Blanch, K. L. Ong, K. M. Crossley, J. Crow, and M. E. Morris, “Training loads and injury risk in Australian football - Differing acute: Chronic workload ratios influence match injury risk,” *Br. J. Sports Med.*, vol. 51, no. 16, pp. 1215–1220, 2017, doi: 10.1136/bjsports-2016-096309.

[30] J. A. Sampson *et al.*, “Injury risk-workload associations in NCAA American college football,” *J. Sci. Med. Sport*, vol. 21, no. 12, pp. 1215–1220, 2018, doi: 10.1016/j.jsams.2018.05.019.

[31] D. G. Altman, *Practical statistics for medical research*. London: Chapman and Hall, 1991.

[32] D. J. Sheskin, *Handbook of parametric and nonparametric statistical procedures.*, 3rd ed. Boca Raton: Chapman and Hall/CRC, 2004.

[33] D. G. Altman and J. M. Bland, “Diagnostic tests. 1: Sensitivity and specificity.,” *BMJ*, vol. 308, no. 6943, p. 1552, Jun. 1994, doi: 10.1136/BMJ.308.6943.1552.

[34] D. J. Harriss, A. Macsween, and G. Atkinson, “Ethical Standards in Sport and Exercise Science Research: 2020 Update,” *Int. J. Sports Med.*, vol. 40, no. 13, pp. 813–817, 2019, doi: 10.1055/a-1015-3123.

[35] B. W. Jakobsen, K. Króner, S. A. Schmidt, and A. Kjeldsen, "Prevention of injuries in long-distance runners," *Knee Surgery, Sport. Traumatol. Arthrosc.*, vol. 2, no. 4, pp. 245–249, 1994, doi: 10.1007/BF01845597.

[36] J. Lysholm and J. Wiklander, “Injuries in runners,” *Am. J. Sports Med.*, vol. 15, no. 2, pp. 168–171, Mar. 1987, doi: 10.1177/036354658701500213.

[37] M. Fanchini, E. Rampinini, M. Riggio, A. J. Coutts, C. Pecci, and A. McCall, “Despite association, the acute:chronic work load ratio does not predict non-contact injury in elite footballers,” *Sci. Med. Footb.*, vol. 00, no. 00, pp. 1–7, 2018, doi: 10.1080/24733938.2018.1429014.

[38] P. Blanch and T. J. Gabbett, “Has the athlete trained enough to return to play safely? The acute:chronic workload ratio permits clinicians to quantify a player’s risk of subsequent injury,” *Br. J. Sports Med.*, vol. 50, no. 8, pp. 471–475, 2016, doi: 10.1136/bjsports-2015-095445.

CHAPTER 5B

Prediction of running injuries from Acute:Chronic Workload Ratio: a machine learning approach

Adapted from
"Prediction of running injuries from Training load: a machine learning approach"

Talko B. Dijkhuis
Ruby T.A. Otter
Hugo Velthuijsen
Koen A.P.M. Lemmink

eTELEMED 2017, The Ninth international Conference on eHealth, Telemedicine and Social Medicine, 2017, pp 109-110
Best Paper Award

ABSTRACT

The prediction of running injuries is problematic. Applying machine learning techniques may be a solution. We aimed to develop a machine learning model to predict injuries in competitive runners. Twenty-three competitive runners kept a daily training log for two years. One-week (acute) and 4-week (chronic) workloads were calculated as the average training duration multiplied by the perceived exertion. The acute:chronic workload ratio (ACWR) was calculated by dividing the acute and chronic ratios. The prediction of sustaining an injury was based on the ACWR and the machine learning algorithms Bayes and Random Forest. Results show that the area under curve is low (0.43-0.60). Just as the precision of predicting an injury (0.03-0.2). Therefore, the precision of the machine learning algorithm predicting injuries must be higher to prevent running injuries actively.

Keywords

Human performance; predictive analysis; load; injuries; monitoring; endurance athlete

INTRODUCTION

We used the context and data of [1]. In [1], using statistics was the fundament of modelling and knowledge about the risk of sustaining an injury in running. In the following, we investigate the possibility of applying machine learning to predict sustaining an injury. We showed the influence of the Acute:Chronic Workload Ratio (ACWR) change on the risk of sustaining an injury. The knowledge of the relationship between workload and the effect on injuries might also be improved by using machine learning techniques. However, machine learning techniques are pointed at enabling prediction instead of fully understanding the system [2]. Although, the input of the knowledge about the influence of change in relative workload on sustaining an injury indicates the variables in the machine learning model. The identified load variables will be used to develop an machine learning model for predicting injuries. Statistically proven relations improve the quality of the machine learning models[3], [4]. The aim is to predict the risk of sustaining an injury using the ACWR of competitive runners using machine learning. To our knowledge, no study in running has investigated the combination of ACWR and machine learning techniques to predict injuries supporting the trainer and runner on intervention in training.

MATERIALS AND METHODS

The participants, the definition of an injury, quantifying workloads, the definition of ACWR, and data analysis, are the same as in [1].

Machine Learning

A machine learning model was constructed to predict the occurrence of an injury. The differences in load between runners were eliminated by using the ACWR. Therefore, we used the ACWR calculated [1] as the machine learning model features. To construct the ACWR as features, the ACWR was split between the four weeks preceding the injury and the four weeks not preceding an injury, labelled as preceding an injury or not. Also, the fortnightly difference between the ACWR preceding an injury or not was used. The constructed features are represented in Table 1.

Table 1. ACWR-based features

ACWR week-4	The ACWR, four weeks before either an injury or not
ACWR week-3	The ACWR, three weeks before either an injury or not
ACWR week-2	The ACWR, two weeks before either an injury or not
ACWR week-1	The ACWR, one week before either an injury or not
ACWR week-1-2	The average ACWR over two weeks before either an injury or not
ACWR week-3-4	The average ACWR over two weeks, two weeks before either an injury or not

We constructed two datasets, one with the ACWR-based features one, two, and three weeks before sustaining an injury or not and another with the complete set of ACWR features. First, the ACWR datasets were split into 80% training and 20% test sets. Next, the training set was resampled to have an equal division of injury and no injury using the SMOTE method [5]. Finally, machine learning models were generated using SMOTE training sets, and the original test sets were applied to identify the quality of the model. Since there is no linear relationship in the relative load, the tree-based algorithm Random Forest algorithm was applied. The machine learning models were constructed using parameter tuning, randomized search, and cross-validation [6]. A Naïve Bayes classifier was used as the baseline model to support the validity of the Random Forest algorithm. Because it is customary to evaluate the machine learning approach, Random Forest should outperform the simple Naïve Bayes baseline classifier. However, random Forest is known for its slightly unstable behaviour. Therefore the generation of the model was repeated 50 times. Subsequently, the model with the highest precision was used to test performance. Precision is the percentage of correctly identified injuries. The following overall performance measures were calculated for each model: accuracy, precision, recall, Area Under Curve (AUC), and F1-score. Finally, precision, recall, and the F1- score were calculated on correctly identified injuries and no injuries.

The scikit-learn package 0.24.0 in Python 3.8.8 was used to construct and judge the machine learning models. The source code, access to the data, and corresponding Jupiter notebook of the machine learning procedure are available as open-source software on Github. (https://github.com/dijkhuist/ Running-Injuries-Machine-Learning, accessed on 16-04-2022).

RESULTS

Of the four models combined with the two data sets, the Random Forest model with all features performed best with an f1-score of 0.92. Also, the Random Forest model with all features performed best on the prediction of no injury, with an f1-score of 0.95. However, the precision was 0.03, and the recall was 0.11. Random Forest outperformed

Naïve Bayes in accuracy, precision, and f1 score but not in the recall or AUC. Table 2 presents the performance of Random Forest and Naïve Bayes.

Figure 1. shows an example of the confusion matrix of the test set in combination with all features and Random Forest.

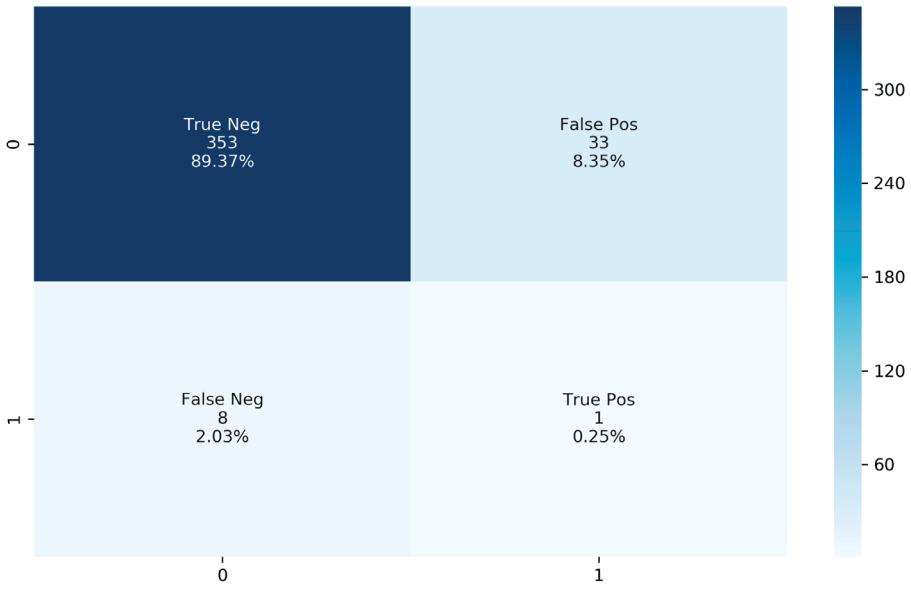


Figure 1. An example of the confusion matrix of the Random Forest model using all features

Table 2. Performance of Random Forest and Naïve Bayes

Algorithm	Features	Accuracy	Precision	Recall	F1 algorithm	AUC	Injury prediction	Precision	Recall	F1 Injury
Random Forest	All*	0.89	0.03	0.22	0.88	0.52	No injury	0.98	0.91	0.95
							Injury	0.03	0.11	0.05
Bayes	All	0.37	0.02	1.00	0.53	0.60	No injury	1.00	0.37	0.54
							Injury	0.2	1.00	0.04
Random Forest	Limited#	0.87	0.04	0.125	0.91	0.43	No injury	0.97	0.86	0.92
							Injury	0.04	0.12	0.06
Naïve Bayes	Limited	0.41	0.03	0.83	0.56	0.60	No injury	0.99	0.40	0.57
							Injury	0.03	0.83	0.6

* All features are (i)the Acute:Chronic Workload Ratio, four weeks, three weeks, two weeks, one week before either an injury or not, (ii) the average Acute:Chronic Workload Ratio over two weeks before either an injury or not, (iii) the average Acute:Chronic Workload Ratio over two weeks, two weeks before either an injury or not.
#The limited number of features are: the Acute:Chronic Workload Ratio, four weeks, three weeks, two weeks, before either an injury or not
UAC = Area Under Curve

DISCUSSION

The main objective was to explore the possibility of predicting the occurrence of an injury with machine learning in competitive runners and provide support to trainers and runners to intervene on time in the training load. The Random Forest model outperformed the Naïve Bayes algorithm. However, none of the models will accurately predict sustaining an injury. Although the overall accuracy and f1-score of the Random Forest models are reasonably high (between 0.86 and 0.92), the recall and precision are very low (between 0.00 and 0.11). A very low precision means that it is impossible to identify injuries correctly. Although the AUC of the Naïve Bayes models was still above 0.5, the precision and recall are too low to be of practical use. A limitation of the study was the limited amount of injuries compared to the available training weeks. Only 31 injuries were identified as applicable in a data set over two years of training of 22 competitive runners. Meaning there was only a sparse small dataset. Lövdal et al. did construct machine learning models predicting injuries on an extended dataset of the same team, using seven years of data of 64 runners and extensively more features [7]. Lövdal et al. suggest a practical appliance of machine learning to predict injuries is possible[7]. However, the specificity of the prediction was between 0.741 and 0.746 (how good is the model at avoiding falsely identified injuries), and sensitivity(= recall) was between 0.504 and 0.584 (how good is the model in identifying the injuries) [7], which has limited use in practice. We found that the ACWR ratio relates to sustaining an injury [1]; however, the relation seems too weak to train a machine learning model. Although generally, Random Forest is a well-performing algorithm, as we showed in [8], some algorithms are supposed to perform better on small data sets, such as the Artificial Neural Network algorithm [9].

Using machine learning algorithms that are more equipped explicitly for small, sparse datasets for further research might be interesting. However, probably more data is needed to train the machine learning algorithm effectively.

CONCLUSION

The prediction of sustaining an injury using the ACWR and machine learning is inaccurate. The realized machine learning models offer no support for trainers or runners in practice, and injury cannot be predicted precisely enough.

Acknowledgments

The authors would like to thank Henk van der Worp for identifying the injuries in the runners' data and Marco Aiello for suggestions on improving the original paper.

REFERENCES

[1] T. B. Dijkhuis, R. Otter, M. Aiello, H. Velthuijsen, and K. Lemmink, “Increase in the Acute:Chronic Workload Ratio relates to Injury Risk in Competitive Runners,” *Int. J. Sports Med.*, vol. 41, no. 11, pp. 736–743, 2020, doi: 10.1055/a-1171-2331.

[2] D. Bzdok, N. Altman, and M. Krzywinski, “Points of Significance: Statistics versus machine learning,” *Nature Methods*, vol. 15, no. 4. Nature Publishing Group, pp. 233–234, Apr. 03, 2018, doi: 10.1038/nmeth.4642.

[3] I. Jebli, F. Z. Belouadha, M. I. Kabbaj, and A. Tilioua, “Prediction of solar energy guided by pearson correlation using machine learning,” *Energy*, vol. 224, p. 120109, 2021, doi: 10.1016/j.energy.2021.120109.

[4] J.-H. Choi, “Investigation of the correlation of building energy use intensity estimated by six building performance simulation tools,” *Energy Build.*, vol. 147, pp. 14–26, 2017, doi: 10.1016/j.enbuild.2017.04.078.

[5] A. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, 2018, doi: 10.1613/jair.1.11192.

[6] T. B. Dijkhuis, F. J. Blaauw, M. W. van Ittersum, H. Velthuijsen, and M. Aiello, “Personalized physical activity coaching: A machine learning approach,” *Sensors (Switzerland)*, vol. 18, no. 2, 2018, doi: 10.3390/s18020623.

[7] S. S. Lövdal, R. J. R. Den Hartigh, and G. Azzopardi, "Injury Prediction in Competitive Runners With Machine Learning," *Int. J. Sports Physiol. Perform.*, vol. 16, no. 10, pp. 1522–1531, 2021, doi: 10.1123/ijspp.2020-0518.

[8] T. B. Dijkhuis *et al.*, “Personalized Physical Activity Coaching: A Machine Learning Approach,” *Sensors*, vol. 18, no. 2, p. 623, Feb. 2018, doi: 10.3390/s18020623.

[9] C. F. Caiafa, Z. Sun, T. Tanaka, P. Marti-Puig, and J. Solé-Casals, "Machine learning methods with noisy, incomplete or small datasets," *Appl. Sci.*, vol. 11, no. 9, May 2021, doi: 10.3390/APP11094132.

CHAPTER 6

General discussion

GENERAL DISCUSSION

This thesis aimed to reduce the data analytics gap represented by the four identified problems while examining the associated potential solutions to enable meaningful insights and predictions related to physical activity and physical performance. These insights and predictions can be used to make more informed decisions regarding physical activity and physical performance interventions.

The chapters in this thesis explored potential solutions to the identified problems. We briefly discuss the problems and their corresponding potential solutions to revisit the introduction. One problem in data analytics for physical activity and physical performance is the limited use of personalised data for meaningful insights and predictions. The potential solution is using personalised data from wearable sensor devices like Fitbit or Garmin or optical tracking systems such as SportsVU. Another problem is the vast amount and complexity of data, making it challenging to build accurate prediction models. The suggested solution is to use more sensitive performance measures in combination with various machine learning algorithms. The third problem involves overly simplified models and assumptions that make unrealistic assumptions about the underlying reality, limiting the value of insights. The possible solution is a causal roadmap combined with a causal model. The fourth problem is the absence of confounding variables, such as contextual variables or individual characteristics, that could influence physical performance. The two suggested methodological solutions use statistical methods that account for the absence of confounding variables and a two-way approach that applies traditional statistical analysis and machine learning to the same incomplete dataset to identify the best fit. All the suggested solutions could reduce the data analytics gap in physical activity and physical performance data and provide meaningful insights and predictions.

KEY FINDINGS AND DISCUSSION

The use of personalised data enables personal meaningful insights and predictions

As found in Chapter 2, based on personal Fitbit step data, we could predict whether a person would reach her/his daily number of steps. Similarly, in Chapter 3, using personalised SportsVU data enabled predicting individual soccer players' performance throughout a soccer match.

However, it is unclear upfront which level of granularity in the data is needed to enable meaningful insights and accurate predictions. In some cases, a higher level of granularity

introduces variability into the data, which reduces the ability to gain meaningful insights and the accuracy of predictions. Transforming the data to a lower granularity level helps filter out variability and produce meaningful insights and more accurate predictions. For instance, in Chapter 2, the too high-level granularity of the minute step data of the Fitbit prevented to predict whether individuals would reach their daily number of steps. Therefore, the minute step data had to be transformed into steps per hour. Likewise, in Chapter 3, the 10Hz individual position data contained a too high-level granularity and had to be transformed into 5-minute periods of performance measures to enable predicting individual soccer players' performance throughout the match. By grouping the raw individual activity data out of the monitoring systems into appropriate time intervals, we can achieve a suitable level of data granularity for making personalised predictions. Nevertheless, one needs to experiment to optimise the level of granularity.

When there are insufficient occurrences of the outcome of interest in the data, it can be challenging to train machine learning models effectively. This is because machine learning algorithms rely on patterns and relationships within the data; when the outcome of interest is rare or underrepresented (i.e. imbalanced dataset), the algorithm may not identify meaningful patterns [1]. For example, as we found in Chapter 2, when a participant did not wear his or her Fitbit regularly, it was hard to determine a pattern or a daily average of steps. As a result, we had to remove these participants from the dataset. Alternatively, specialised machine learning algorithms have been developed to handle imbalanced datasets, which take into account the imbalanced nature of the data and adjust their predictions accordingly. These machine learning algorithms can be combined with data balancing techniques such as oversampling or undersampling the dataset to improve predictions. [2]. For instance, in Chapter 3, we effectively used Random Forest, a machine learning technique less biased toward the majority class to address the imbalance in the dataset. Furthermore, to balance the dataset we used the SMOTE algorithm, which generated synthetic data points for the underrepresented class to balance the dataset, making the dataset more suitable for training machine learning models.

The use of more sensitive performance measures influences positively the quality of predictions.

Chapter 3 demonstrated that the precision of machine learning algorithms in predicting physical performance increases with the sensitivity of physical performance measures. Sensitivity is defined as the responsiveness of the physical performance measures to changes in physical activity [3]. For example, the most sensitive measure, 'energy in power category', leads to a 20% more accurate machine learning model than applying the least sensitive physical performance measure, 'distance covered'. Consequently, our analysis indicates that the strength of the prediction models is related to the choice

of physical performance measures. The literature [4], [5] has acknowledged that the accuracy of an algorithm can differ across datasets due to various factors, including feature selection, such as physical performance measures. Therefore, it is crucial to choose the physical performance measures carefully.

The use of various machine learning algorithms influences the quality of predictions positively.

Chapter 2 involved training eight machine learning algorithms and comparing their performance to a baseline algorithm. The authors determined the optimal algorithm for a given dataset by assessing the performance of each algorithm, with the Random Forest algorithm demonstrating superior predictive ability for individual thresholds, while the ADABoost algorithm showed the highest precision for group-level predictions, which aligned with previous research findings [4], [5]. Furthermore, as discussed in Chapters 2 and 3, we found that the choice of machine learning algorithm significantly impacted the precision of individual predictions, highlighting that determining the most suitable algorithm for a given dataset is a context-dependent art rather than a science.

Identifying the best combination of the dataset and algorithm is time-consuming. Identifying the best combination might be accelerated by utilising ensemble methods, combining multiple algorithms to produce one cohesive model [5]. For example, Chapter 4 employed an advanced ensemble method Super Learning. Super Learning is supposed to outperform other single algorithms by automatically selecting the best algorithm or combination of algorithms [6] to assess the impact of a substitute player on a soccer team's physical performance.

However, even when an accurate machine learning model is constructed, machine learning models can become outdated, for machine learning models are confronted with a constantly changing environment and data [7]. For instance, the individual player performance can change over time due to factors such as injuries, fatigue, or changes in form. This can affect the team's performance and lead to a drift in the machine model's predictions. Therefore, it is crucial to continuously monitor the accuracy and precision of machine learning models and retrain the machine learning models when their predictions no longer correspond with the changing circumstances and data [7].

Following a causal roadmap in combination with a causal model prevents oversimplified assumptions.

By following a causal roadmap and creating a causal model in Chapter 4, depicting the causes and effects of a substitution in soccer, we could identify the factors that

drive the team's physical performance and the potential consequences of substitution. Moreover, applying the Directed Acyclic Graph (DAG) [8] as a causal model helped identify and explicate unobserved confounding variables that may influence the results. This approach ensured that the assumptions were based on a thorough understanding of substitution and the team's physical performance. For example, in the case of a soccer match, by explicating that the physical performance of the team is influenced by contextual variables such as match location (home or away), score (win, draw or lose), or competitive level of the opponent [9]–[11], while not present in the dataset, prevents the reliance on oversimplified assumptions. However, the second assumption of the applied causal roadmap dictates that there should be no unmeasured confounding between the variable of interest's change and the outcome response. For instance, in Chapter 4, the second assumption can be expressed as no unmeasured confounding influences the outcome of substitution on the team performance. Therefore, explicating unmeasured confounding variables using the DAG while ensuring an explicit definition of reality contradicts this assumption. It is challenging to include all confounding variables, as numerous variables can affect the variable of interest's change and the outcome response. Therefore, it is important to acknowledge the limitations of following the causal roadmap in combination with DAG. A solution might be found in the application of a less strict but formal and explicit data analytics method such as 'Knowledge Discovery in Databases' [12]–[14] (KDD) in sports and daily life. KDD provides a framework for formalising data analytics, data handling, machine learning applications, and statistical modelling [12]. As such, explicating the steps taken during data analytics offers guidance and insights that avoid the unrealistic assumption of the causal roadmap that there is no unmeasured confounding between the variable of interest's change and the response of the outcome in sports and daily life.

Using statistical methods that account for the absence of confounding variables improves the quality of the statistical insights.

In Chapter 4, we combined TMLE with the ensemble machine learning method Super Learner to determine the influence of substitution on the soccer team's physical performance. The findings revealed that TMLE effectively reduces the negative impact of missing or unmeasured variables on the calculated accuracy of the effect of substitution. Also, the applied ensemble machine learning method Super Learner is known for its accurate predictions on data containing missing variables [6], [15]. Additionally, the combination of TMLE and the Super Learner could be effectively employed in other fields where unmeasured confounding variables are a concern.

Although the combination of TMLE with the Super Learner reduces the negative impact of missing variables in statistical insights, as highlighted by references [16], [17], the

issue of incomplete datasets in personalised predictions is a common challenge that can significantly impact the accuracy and reliability of the predictions.

> *Gaining insights into the relations of physical activity, physical performance, and the resulting outcomes is sometimes the maximum that can be achieved.*

Chapter 5A used an incomplete dataset on running and overuse injuries to perform statistical analysis. The analysis revealed that increased load is associated with increased injury risk. However, the low specificity and sensitivity of the relationship between load and increased injury risk made it an insufficient predictor of injury risk. The low specificity results in possible falsely identified runners who are not at risk of sustaining an injury. Conversely, the low sensitivity may fail to identify runners at risk of injury. Although it is a meaningful insight that increased load is associated with increased risk, this also highlights the limitations of statistics when the necessary variables are not present in the dataset. Also, the practical use of machine learning predictions can be limited when an incomplete dataset is used. For example, in Chapter 5B, machine learning did not contribute to correctly predicting an overuse injury. We used machine learning on the load dataset of competitive runners of Chapter 5A. Our findings revealed that the machine learning model's precision and recall of the prediction of an overuse injury due to an increase in the load were lower than desired. The machine learning model's low precision implies possibly falsely identifying runners as being at risk of injury. In addition, the model's low recall suggests a failure to identify a significant number of runners at risk of injury. Our results align with those of Lövdal et al., who applied machine learning to a partially similar dataset [18].

For statistical or machine learning models to be practical, high specificity, sensitivity, precision, and recall are essential. Although the insight that load can impact injury risk is valuable, predicting an overuse injury with only load as the variable, without considering other important confounding variables, is unlikely to be useful [19]. Regularly intentionally increasing the load is a crucial part of training to improve physical performance [20]–[22]. The pattern of increased load is frequently observed before an injury occurs, but it is even more prevalent in the dataset due to being a standard training pattern. Thus, it is crucial to incorporate confounding internal and external variables in the dataset to improve the reliability of the injury risk prediction. For instance, in predicting injury risk in running, adding internal factors, such as personality dimensions like attribution style or cognitive style [23], and contextual factors, such as social context or a negative life event [24], may contribute to the applicability of machine learning.

There are multiple reasons why datasets may be incomplete, such as the absence of confounding variables or incomplete observations, resulting in data gaps that can have a negative impact on predictive models. This issue is broader than the missing variables in Chapters 5A and 5B. For example, in Chapter 2, additional information was desired to include variables that may have influenced a person's daily walking routine, such as work schedule, days off, or meetings during lunch breaks. Similarly, in Chapter 3, for example, the ranking or score of a football match was unknown, as were any system changes, which reduced the informative value of the prediction.

STRENGTHS AND LIMITATIONS, AND RECOMMENDATIONS

The research presented in this thesis possesses several strengths and limitations.
A major strength of our research was the use of existing large, personalised datasets of physical activity and physical performance. Large datasets can help to identify patterns and trends in physical activity and physical performance data that may be difficult to detect in smaller samples. In addition, utilising datasets comprising information on individual participants has facilitated the ability to generate detailed insights and predictions at the individual level, demonstrating the feasibility of more informed intervention on an individual level.

Furthermore, another strength is the implemented various machine learning algorithms and advanced statistical techniques, including Targeted Maximum Likelihood Estimation allowing for identifying complex relationships and patterns that may be difficult or impossible to discern using traditional statistical techniques. This innovative approach has not been previously implemented in the field of sport science, and the results of this study demonstrate that integrating these techniques expands the methodological toolbox available to researchers in sport science. Furthermore, the various machine learning algorithms allow for the development of more accurate and reliable models for predicting physical performance in individuals.

An additional strength is that it marks the first application of both a causal roadmap and a causal model in the field of sport science. The applied causal roadmap and causal model substantially advance understanding and analysing complex phenomena like in elite sports. By incorporating the principles of causality, using the causal roadmap and a causal model allows for a more thorough examination of the underlying mechanisms and factors that influence athletic performance. Overall, this approach holds promise for advancing our understanding and developing more rigorous and informative research.

Next to the strengths of this thesis, there are limitations. One of the limitations is the low number of variables in the datasets, limiting the accuracy and usability of prediction or identified relationships. Additionally, as mentioned in the discussions of Chapters 2, 3, 4, and 5, the studies use only the physical activity data captured by wearables, monitoring systems, or athletes' logs, not solving the endogeneity problem. Therefore, it may be necessary to design and construct monitoring systems according to the selected physical performance measures *and* add internal and contextual factors that may improve the applicability of machine learning in practice. For instance, as aforementioned, in predicting injury risk in running, adding internal factors, such as personality dimensions such as attribution style or cognitive style [23], and contextual factors, such as social context or a negative life event [24], may contribute to the applicability of machine learning. Alternatively, in soccer, adding contextual factors, such as match location (home or away), score (win, draw or lose), and rival level, may improve the accuracy of machine learning models [9]–[11].

Another limitation of this thesis is the labour-intensive process of data preparation. The physical activity data used in this thesis was initially raw data from wearables, monitoring systems and athletes' logs. However, the raw data must be prepared to construct physical performance measures. The data was prepared by hand for every study, erroneous data were excluded, and missing data were either excluded or imputed. The data preparation took much time before statistical analysis or machine learning could be applied. To enable timely prediction and more informed decisions, labour-intensive data preparation processes must be eliminated to apply machine learning in a live situation. To ensure a minimum of data preparation and live prediction, predictive monitoring systems must be developed with the selection of the physical performance measures in the back of one's mind. The predictive monitoring systems must also transform tracking data in a pipeline directly to physical performance measures that enable machine learning and prediction in the short term or even in real-time. The predictive monitoring systems preferably have to conform to the guidelines of trustworthy AI, as stated by the European Union [25]. The systems must be lawful (respecting all applicable laws and regulations), ethical (respecting ethical principles and values), and robust (from a technical perspective) while considering its social environment.

Taking advantage of data analytics and machine learning requires transparency and trustworthiness. By ensuring that machine learning and prediction are built on solid ethical principles and practices and open to scrutiny and validation, we can increase the confidence level in the results and enable their wider adoption. While we hope this thesis has contributed to confidence in data analytics and machine learning in daily life and elite sports, realising the adoption, transparency, and trustworthiness of data analytics and machine learning needs to be examined.

GENERAL CONCLUSION

This thesis showed that personalised data, machine learning, advanced statistics, and a causal roadmap in combination with a causal model could help to reduce the data analytics gap between physical activity and physical performance data and the ability to extract meaningful insights and predictions. While reducing the data analytics gap, we showed the potential of data analytics to gain meaningful insights and predictions on physical activity, physical performance, or injuries, enabling more informed interventions in physical activity. These results provide a foundation for future research to reduce the data analytics gap even more.

PRACTICAL IMPLICATIONS AND OUTLOOK

The practical implications of using data in coaching and training are significant. One approach could be to develop rich datasets that combine individual and contextual data, using systems that collect data frequently and consistently. To achieve this, collaborating with data experts who can help interpret this data would be helpful. Another approach could be to put together a multidisciplinary team of data experts who work together with coaches and trainers. By fostering a structural collaboration between science and practice, teams can develop effective strategies and predictive monitoring systems using data to improve coaching and training outcomes. With the correct data and data analytics possibly combined with automated data processing, coaches and trainers can gain valuable insights and predictions into their athletes' performance, training methods and recovery, enabling timely interventions leading to better overall results. Moreover, establishing a structural collaboration between science and practice can improve virtual coaching strategies and virtual coaching systems by generating data from the continuous monitoring of individuals in their daily lives. The process of continuous monitoring could make it possible to generate personalised valuable insights, predictions and recommendations. As a result, individuals can make real-time adjustments based on these insights to optimize their performance and overall well-being.

REFERENCES

[1] H. He and E. A. Garcia, “Learning from imbalanced data,” *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1007/978-3-030-04663-7_4.

[2] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, “Handling imbalanced datasets : A review,” *Science (80-.).*, vol. 30, no. 1, pp. 25–36, 2006.

[3] M. Buchheit and B. M. Simpson, “Player-Tracking Technology : Half-Full or Half-Empty Glass ?,” *Int. J. Sports Physiol. Perform.*, vol. 12, no. S2, pp. 35–41, 2017.

[4] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, “Comparing different supervised machine learning algorithms for disease prediction,” *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, pp. 1–16, 2019, doi: 10.1186/s12911-019-1004-8.

[5] I. Ibrahim and A. Abdulazeez, “The Role of Machine Learning Algorithms for Diagnosing Diseases,” *J. Appl. Sci. Technol. Trends*, vol. 2, no. 01, pp. 10–19, 2021, doi: 10.38094/jastt20179.

[6] M. J. van der Laan, E. C. Polley, and A. E. Hubbard, “Super Learner,” *Stat. Appl. Genet. Mol. Biol.*, vol. 6, no. 1, 2007.

[7] S. Makinen, H. Skogstrom, E. Laaksonen, and T. Mikkonen, “Who needs MLOps: What data scientists seek to accomplish and how can MLOps help?,” *Proc. - 2021 IEEE/ACM 1st Work. AI Eng. - Softw. Eng. AI, WAIN 2021*, pp. 109–112, 2021, doi: 10.1109/WAIN52551.2021.00024.

[8] J. Pearl and D. Mackenzie, *The Book of Why*. New York: Basic Books, 2018.

[9] C. Lago, L. Casais, E. Dominguez, and J. Sampaio, “The effects of situational variables on distance covered at various speeds in elite soccer,” *Eur. J. Sport Sci.*, vol. 10, no. 2, pp. 103–109, 2010, doi: 10.1080/17461390903273994.

[10] J. Castellano, A. Blanco-Villaseñor, and D. Álvarez, "Contextual variables and time-motion analysis in soccer," *Int. J. Sports Med.*, vol. 32, no. 6, pp. 415–421, 2011, doi: 10.1055/s-0031-1271771.

[11] V. I. Kalapotharakos, A. Gkaros, and E. Vassliades, “Influence of contextual factors on match running performance in elite soccer team,” *J. Phys. Educ. Sport*, vol. 20, no. 6, pp. 3267–3272, 2020, doi: 10.7752/jpes.2020.s6443.

[12] U. Fayyad, G. Piatetsky-shapiro, and P. Smyth, “From Data Mining to Knowledge Discovery in Databases,” *AI Mag.*, vol. 17, no. 3, pp. 37–54, 1996, doi: 10.1007/978-3-319-18032-8_50.

[13] A. Sims *et al.*, “Data-Centric Automated Data Mining,” *Big Data Res.*, vol. 2, no. 2, pp. 1–36, Jun. 2016, doi: 10.1519/JSC.0000000000000499.

[14] C. A. Palacios, J. A. Reyes-Suárez, L. A. Bearzotti, V. Leiva, and C. Marchant, "Knowledge Discovery for Higher Education Student Retention Based on Data Mining: Machine Learning Algorithms and Case Study in Chile," 2021, doi: 10.3390/e23040485.

[15] M. J. van der Laan and S. Rose, *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. 2018.

[16] J. G. Claudino, D.-O. Capanema, T.-V. De-Souza, J. C. Serrão, A.-C. Machado Pereira, and G.-P. Nassis, "Current Approaches to the Use of Artificial Intelligence for Injury Risk Assessment and Performance Prediction in Team Sports: a Systematic Review," *Sport. Med. - Open*, vol. 5, no. 1, 2019.

[17] E. Morgulev, O. H. Azar, and R. Lidor, “Sports analytics and the big-data era,” *Int. J. Data Sci. Anal.*, vol. 5,

no. 4, pp. 213–222, 2018, doi: 10.1007/s41060-017-0093-7.

[18] S. S. Lövdal, R. J. R. Den Hartigh, and G. Azzopardi, "Injury Prediction in Competitive Runners With Machine Learning," *Int. J. Sports Physiol. Perform.*, vol. 16, no. 10, pp. 1522–1531, 2021, doi: 10.1123/ijspp.2020-0518.

[19] D. van Poppel *et al.*, "Risk factors for overuse injuries in short- and long-distance running: A systematic review," *J. Sport Heal. Sci.*, vol. 10, no. 1, pp. 14–28, Jan. 2021, doi: 10.1016/J.JSHS.2020.06.006.

[20] S. L. Halson, "Monitoring Training Load to Understand Fatigue in Athletes," *Sports Medicine*, vol. 44, no. Suppl 2. Springer, pp. 139–147, Nov. 2014, doi: 10.1007/s40279-014-0253-z.

[21] S. L. Halson and A. E. Jeukendrup, "Does overtraining exist? An analysis of overreaching and overtraining research.," *Sport. Med.*, vol. 34, no. 14, pp. 967–981, 2004, [Online]. Available: http://search.ebscohost.com/login.aspx?direct=true&db=ccm&AN=106594041&site=ehost-live.

[22] R. Meeusen *et al.*, "Prevention, diagnosis, and treatment of the overtraining syndrome: Joint consensus statement of the european college of sport science and the American College of Sports Medicine," *Med. Sci. Sports Exerc.*, vol. 45, no. 1, pp. 186–205, 2013, doi: 10.1249/MSS.0b013e318279a10a.

[23] P. Shrivastava, R. Venugopal, and Y. Singh, "A Study of Personality Dimensions in Sports Performance," *J. Exerc. Sci. Physiother.*, vol. 6, no. 1, pp. 39–42, 2010.

[24] R. T. A. Otter, M. S. Brink, R. L. Diercks, and K. A. P. M. Lemmink, "A Negative Life Event Impairs Psychosocial Stress, Recovery and Running Economy of Runners," *Int. J. Sports Med.*, vol. 37, no. 3, pp. 224–229, Dec. 2016, doi: 10.1055/s-0035-1555932.

[25] High-Level Expert Group on Artificial Intelligence, "Ethics Guidelines for Trustworthy AI," pp. 2–36, 2019, [Online]. Available: https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top.

APPENDICES &

Summary
Samenvatting
Achtergrond en academisch werk
Dankwoord
Research Institute SHARE

SUMMARY

As a human being, whether in daily life or elite sports, it is vital to have the ability to perform physically. To maintain or improve our physical performance, humans must perform physical activity regularly or extensively. The effect of physical activity depends on various conditions, such as the amount of physical activity, the individual's physical capability, and the individual's physical capacity.

Physical activity is important for ordinary people to stay healthy. To sustain health or improve the ability to perform physically, it is not necessary to exercise fanatically; daily physical activities such as cycling, walking, household or gardening also contribute. However, insufficient physical activity can lead to a loss of physical capacity, illness, and a lower life expectancy.

In sports, the term workload denotes the combination of frequency, duration, and (perceived) intensity of physical activity an athlete undertakes during training. While having natural talent contributes to success in sports, becoming a top athlete requires many years of training. During these years of training, a delicate balance must be maintained between capacity, workload and recovery. An imbalance between capacity, workload and recovery can result in underload or overload, leading to performance degradation or injuries.

In recent years, the ability to monitor workload and human physical activity has increased tremendously due to the emergence of wearables and the data wearables collect. In addition, the technological advances in monitoring systems in elite sports, such as football, basketball and hockey, have dramatically increased the ability to monitor individual athletes.

Collecting data using the technological advancements mentioned above could enable machine learning and statistics gaining insights and predicting physical activity and physical performance. Intervention in physical activity is necessary when adverse consequences related to physical activity or physical performance are expected. The goal of intervention is twofold: to maintain or improve physical activity and physical performance while preventing physical performance decline or injury.

Although a lot of data is available, extracting useful personal insights and predictions can be challenging. Data analytics can help extract insights and predictions from collected data. Data analytics involves examining, transforming, and interpreting data to gain meaningful insights, and predictions. Data analytics implies the use of various statistical and machine learning techniques. However, data analytics is not widely

used in behavioural, human movement, or sports science. Additionally, using Artificial Intelligence and machine learning based on wearable and monitoring systems data in sports is still in the early stages. While data analytics offers opportunities, problems in extracting meaningful, personalised insights and realising predictions based on physical activity and physical performance data exist. Four problems are identified in data analytics of physical activity and physical performance data, creating a data analytics gap. To address these problems, we propose potential solutions for each identified problem.

The first problem is the limited use of individualised prediction based on personalised. In order to provide individualised insights and predictions, a boundary condition is that the data contains sufficient personal information. The potential solution is to use personalised data from wearables like Fitbit or Garmin or optical tracking systems such as SportsVU for individualised insights and predictions. **The second problem** is the vast amount and complexity of data, making it challenging to create accurate prediction models. The suggested solution is to use more sensitive physical performance measures in combination with various machine learning algorithms. These more sensitive measures can better detect changes in the measured system. **The third problem** involves the use of simplified models of reality and hypotheses that make unrealistic assumptions about reality, limiting the value of insights. The possible solution is to apply a causal roadmap combined with a causal model. The causal roadmap strictly prescribes which steps must be taken in analysis and clarifies the underlying assumptions. A causal model provides insight into how reality is modelled, and which variables have been left out. **The fourth problem** is the absence of variables in the data that can influence the measured physical activities and performance. For example, in soccer, contextual variables such as a home or away game or the weather. There are two potential methodological solutions. The first solution uses statistical methods that take into account the absence of influencing variables, and the second tests whether traditional statistical analysis or machine learning applied to the same incomplete dataset provides better insights and predictions.

The aim of this thesis is to reduce the gap in data analytics by addressing the four identified problems. By exploring the associated potential solutions, it may become possible to provide meaningful insights and predictions regarding physical activity and performance. These insights and predictions can be used to make more informed decisions regarding physical activity and physical performance interventions.

The chapters in the thesis explored the potential solutions to the identified problems.

In Chapter 2, we used personalised data captured by wearable devices to predict employees' daily physical performance automatically. Specifically, we used Fitbits to track the Hanze University of Applied Science employees' daily step counts and employed

advanced machine learning algorithms to predict whether they would achieve their step goals. By automatically analysing physical activity and physical performance, timely detection of anomalies in behaviour and identifying effective coaching strategies may become possible. The results showed that tree-based algorithms best predict whether an employee will achieve his or her step goal.

In Chapter 3, we studied the predictability of physical performance in elite soccer matches using various physical performance measures and machine learning techniques. We gathered data from 302 matches in a single season using the SportsVU optical tracking system, which recorded the positions of each player throughout the matches. Based on this data, we measured physical performance using three increasingly sensitive performance measures, i.e. distance covered, distance in speed zones, and energy expenditure in power zones. These physical performance measures were used in different machine learning models to identify and predict the physical performance of individual players throughout an elite soccer match. We found that the more sensitive the performance measures were, the better the physical performance of the individual player could be predicted.

In Chapter 4, we studied the impact of substitutions on a soccer team's physical performance using a causal roadmap and causal model. The causal roadmap strictly prescribed which steps should be taken in analysis and clarified the underlying assumptions. In which a causal model provided insight into how reality is modelled, and which variables have been left out. Our causal model included variables such as the number and timing of substitutions and the total distance covered. The causal roadmap and causal model helped us identify assumptions and potential sources of bias and confounding that could affect causal effect estimates.

In Chapter 4, we also provided an in-depth analysis of statistical methods. We tested the accuracy of estimating the impact of substitutes on a football team's physical performance using replacement data and data from the SportsVU optical tracking system. We compared the accuracy of two methods: Targeted Maximum Likelihood Estimation (TMLE) and a generalised linear model based on the complete dataset. We also tested the accuracy of these methods when a critical variable was removed from the dataset. The more robust TMLE method offered more accurate insights than the generalised linear model, especially in the absence of a critical variable in the dataset.

In Chapter 5, we examined the impact of workload on injuries in runners by applying both statistical analysis and machine learning. Our dataset consisted of individual data from test, training sessions, and injury logs. We used physical load measures like training duration and rate of perceived exertion to construct an acute:chronic workload ratio. Our

objective was to test whether traditional statistical analysis or machine learning applied to the inherently incomplete dataset provides better insights and predictions. We found a statistical relationship between workload and injury risk. However, both statistics and machine learning are limited in delivering actionable predictions to prevent injuries.

Conclusion

This thesis showed that personalised data, machine learning, sensitive performance indicators, advanced statistics, and a causal roadmap in combination with a causal model could help to reduce the data analytics gap. As a result, this creates the possibility of extracting meaningful insights and predictions from physical activity and physical performance data. While reducing the data analytics gap, we showed the potential of data analytics to gain meaningful insights and predictions on physical activity, physical performance, or injuries, enabling more informed interventions in physical activity. These results provide a foundation for future research to reduce the data analytics gap even more and potentially the realisation of automated monitoring, prediction, and coaching systems.

SAMENVATTING

Voor een mens, of het nu in het dagelijks leven of in de topsport is, is het van belang om het vermogen te hebben om fysiek te presteren. Om fysieke prestaties te behouden of te verbeteren, moeten mensen regelmatig of uitvoerig fysiek actief zijn. Het effect van fysieke activiteit hangt af van de omstandigheden, zoals de hoeveelheid fysieke activiteit, de fysieke mogelijkheden van het individu en de fysieke capaciteit van het individu.

Fysieke activiteit is belangrijk voor gewone mensen om gezond te blijven. Om de gezondheid op peil te houden of fysiek beter te presteren, is het niet nodig om fanatiek te sporten. Dagelijkse fysieke activiteiten zoals fietsen, wandelen, huishouden of tuinieren kunnen ook bijdragen. Echter onvoldoende fysieke activiteit in het dagelijks leven kan leiden tot een verlies van fysieke capaciteit, ziekte en een lagere levensverwachting.

In de sport, geeft de term werkbelasting de combinatie aan van frequentie, duur, en (ervaren) intensiteit van de fysieke activiteit die een athleet onderneem tijdens training. Talent draagt bij aan het succesvol kunnen zijn in sport, maar het vergt vele jaren van training om als topsporter succesvol te zijn. Tijdens de trainingsjaren moet een delicaat evenwicht worden bewaard tussen capaciteit, werkbelasting en herstel. Een onbalans tussen capaciteit, werkbelasting en herstel, kan leiden tot onderbelasting of overbelasting, en kan achteruitgang van de prestaties of blessures tot gevolg hebben.

Door de opkomst van wearables en de gegevens die de wearables verzamelen, zijn de mogelijkheden om werkbelasting en menselijke prestaties te volgen enorm toegenomen. Ook In topsporten, zoals voetbal, basketbal en hockey, hebben de technologische ontwikkelingen van monitoringsystemen de mogelijkheden om individuele atleten te volgen enorm vergroot.

Het verzamelen van gegevens met behulp van de hierboven genoemde technologische vooruitgang maakt het mogelijk, om met behulp van statistiek en machine learning, inzicht te verkrijgen in en het voorspellen van fysieke activiteit en fysieke prestaties. Interventie is nodig wanneer nadelige resultaten van fysieke activiteit of fysieke prestaties worden verwacht. Aan de ene kant om fysieke prestaties te behouden en betere prestaties mogelijk te maken. Aan de andere kant om prestatievermindering of blessures te voorkomen.

Hoewel er veel gegevens beschikbaar zijn, kan het moeilijk zijn om nuttige persoonlijke inzichten en voorspellingen te extraheren. Data analytics kan helpen bij het extraheren van inzichten en voorspellingen uit verzamelde gegevens. Data analytics omvat het onderzoeken, transformeren en interpreteren van gegevens om zinvolle inzichten en

voorspellingen te verkrijgen. Data analytics impliceert het gebruik van verschillende statistische en machine learning technieken. Data analytics wordt echter niet veel gebruikt in gedrags-, bewegings- en sportwetenschap. Daarnaast staat het gebruik van Artifical Intelligence en machine learning op basis van wearable- en monitoringsysteemdata in de sport nog in de kinderschoenen. Hoewel data analytics kansen biedt, zijn er problemen bij het extraheren van zinvolle, persoonlijke inzichten en het realiseren van voorspellingen op basis van fysieke activiteits- en prestatiegegevens. Vier geïdentificeerde problemen creëren een data-analytics kloof. Voor elk van vier geïdentificeerde problemen stellen we mogelijke oplossingen voor.

Een eerste probleem is het beperkte gebruik van gepersonaliseerde gegevens en geïndividualiseerde voorspellingen en inzichten. Om geïndividualiseerde inzichten en voorspellingen te kunnen geven, is het een randvoorwaarde dat de gegevens voldoende persoonlijke informatie bevatten. Een mogelijke oplossing is het gebruik van gepersonaliseerder gegevens uit wearables zoals Fitbit of Garmin of optische volgsystemen zoals SportsVU om geindividualiseerde inzichten en voorspellingen te kunnen geven. **Een tweede probleem** is de enorme hoeveelheid en complexiteit van gegevens, waardoor het een uitdaging is om nauwkeurige voorspellingsmodellen te creëren. Een mogelijke oplossing is om meer sensitieve fysieke prestatieindicatoren te gebruiken en in combinatie met verschillende machine learning-algoritmen de beste voorspelmodellen te selecteren. Een meer sensitieve prestatieindicator is beter in staat veranderingen in de gegevens van het gemeten systeem te detecteren. **Het derde probleem** betreft het gebruik van vereenvoudigde modellen van de werkelijkheid en hypotheses die onrealistische aannames doen over werkelijkheid, waardoor de waarde van inzichten wordt beperkt. De mogelijke oplossing is het toepassen van een causale roadmap gecombineerd met een causaal model. De causale roadmap schrijft strikt voor welke stappen er in een analyse genomen moeten worden en expliciteert de onderliggende aannames. Waarbij een causaal model inzichtelijk maakt hoe de werkelijkheid gemodeleerd is en welke variabelen buiten beschouwing zijn gelaten. **Het vierde probleem** is het ontbreken van variabelen in de gegevens die de gemeten fysieke activiteiten en prestaties kunnen beïnvloeden. Een voorbeeld daarvan zijn contextuele variabelen in voetbal zoals een uit- of thuiswedstrijd of het weer. Er zijn twee potentiële methodologische oplossingen. De eerste oplossing maakt gebruik van statistische methoden die om kunnen gaan met ontbrekende variabelen. De tweede vergelijkt traditionele statistische analyse met machine learning om te kijken welke methode betere inzichten en voorspellingen levert als er variabelen ontbreken.

Het doel van dit proefschrift is om de kloof in data analytics te reduceren door de vier geïdentificeerde problemen aan te pakken. Door de bijbehorende potentiële oplossingen te onderzoeken kan het mogelijk worden om zinvolle inzichten en voorspellingen te

geven met betrekking tot fysieke activiteit en prestaties. Deze inzichten en voorspellingen kunnen worden gebruikt om beter geïnformeerde beslissingen te nemen met betrekking tot fysieke activiteit en fysieke prestatie interventies.

De hoofdstukken in het proefschrift verkenden mogelijke oplossingen voor de geïdentificeerde problemen.

In hoofdstuk 2 gebruikten we persoonlijke gegevens die zijn vastgelegd door wearables om de dagelijkse fysieke prestaties van de individuele werknemers automatisch te voorspellen. We gebruikten Fitbits om de dagelijkse stappen van de Hanzehogeschool werknemers bij te houden en pasten geavanceerde machine learning algoritmen toe om te voorspellen of de stapdoelstellingen zouden worden bereikt. Door het automatisch analyseren van fysieke activiteit en fysieke prestaties, kan tijdige detectie van afwijkingen in gedrag en het identificeren van effectieve coachingstrategieën mogelijk worden. De resultaten lieten zien dat tree-algoritmen het beste voorspellen of de individuele medewerker zijn of haar doel zal bereiken.

In hoofdstuk 3 bestudeerden we de voorspelbaarheid van fysieke prestaties in topvoetbalwedstrijden met behulp van verschillende fysieke prestatieindicatoren en machine learning technieken. We verzamelden gegevens van 302 wedstrijden in één seizoen met behulp van het SportsVU optische volgsysteem. SportVU registreerde de posities van elke speler gedurende de wedstrijden. Op basis van deze gegevens hebben we de fysieke prestaties gemeten met behulp van drie in toenemende mate sensitieve prestatieindicatoren: afgelegde afstand, afstand in snelheidszones en energieverbruik in vermogenszones. Deze prestatieindicatoren werden gebruikt in verschillende machine learning modellen om de fysieke prestaties van individuele spelers tijdens een topvoetbalwedstrijd te identificeren en te voorspellen. We ontdekten dat hoe sensitiever de prestatieindicatoren zijn, hoe beter de fysieke prestaties van de individuele speler kunnen worden voorspeld.

In hoofdstuk 4 bestudeerden we de impact van wisselspelers op de fysieke prestaties van een voetbalteam met behulp van een causale roadmap en causaal model. De causale roadmap schreef strikt voor welke stappen er in een analyse genomen moesten worden en expliciteerde de onderliggende aannames. Waarbij een causaal model inzichtelijk maakte hoe de werkelijkheid gemodeleerd is en welke variabelen buiten beschouwing zijn gelaten. Ons causaal model bevatte variabelen zoals het aantal en de timing van vervangingen en de totale afgelegde afstand waarbij contextuele variabelen zoals uit- of thuiswedstrijd, systeem, ranking of stand van de wedstrijd expliciet uitgesloten waren. De causale roadmap en het causale model hielpen ons bij het gestructureerd analyseren en het voorkomen van impliciet modeleren van de werkelijkheid.

In hoofdstuk 4 analyseerden we verschillende statistische methoden. We hebben de nauwkeurigheid getest van het schatten van de impact van invallers op de fysieke prestaties van een voetbalteam met behulp van positie- en vervangingsdata en data van het optische volgsysteem SportsVU. We vergeleken de nauwkeurigheid van twee methoden: Targeted Maximum Likelihood Estimation (TMLE) en een gegeneraliseerd lineair model op basis van de volledige dataset. We hebben ook de nauwkeurigheid van deze methoden getest wanneer een cruciale variabele uit de dataset werd verwijderd. De robuustere TMLE-methode bood nauwkeurigere inzichten dan het gegeneraliseerde lineaire model, vooral bij afwezigheid van een cruciale variabele in de dataset. Door middel van statistische analyse en machine learning onderzochten we in **hoofdstuk 5** de impact van trainingsbelasting op blessures bij hardlopers. Onze dataset bestond uit gegevens van testen, trainingssessies en blessurelogboeken. We gebruikten fysieke werkbelastingindicatoren zoals trainingsuren en ervaren mate van inspanning, om een acute: chronische werkbelasting ratio te creëren. Ons doel was om te testen of traditionele statistische analyse of machine learning toegepast op de inherent incomplete dataset betere inzichten of voorspellingen oplevert. We vonden een statistisch significant verband tussen trainingsbelasting en blessurerisico. Echter, zowel statistische analyse als machine learning zijn beperkt in het geven van bruikbare voorspellingen voor het voorkomen van blessures.

Conclusie

Op basis van de bevindingen van dit proefschrift is de conclusie dat het combineren van gepersonaliseerde gegevens, sensitieve prestatieindicatoren, machine learning, geavanceerde statistieken en een causale roadmap in combinatie met een causaal model, kunnen helpen om de data analytics kloof tussen fysieke activiteit en fysieke prestatiegegevens en het vermogen om er zinvolle inzichten en voorspellingen te extraheren te verkleinen. Terwijl we de kloof in data analytics verkleinden, toonden we het potentieel van het toepassen van data analytics aan. Door het toepassen van data analytics en het verkleinen van de data analytics kloof wordt het mogelijk beter geïnformeerde interventies in fysieke activiteit te plegen.

Deze resultaten bieden een basis voor toekomstig onderzoek om de data-analysekloof nog meer te verkleinen en mogelijk geautomatiseerde monitoring, voorspelling en coachings-systemen te realiseren .

ACHTERGROND EN ACADEMISCH WERK

Talko Dijkhuis is op 28 februari 1969 geboren in Nijeveen, Nederland. Hij studeerde van 1988 tot 1994 Bedrijfskunde aan de Rijksuniversiteit Groningen, waarvan een half jaar in Stockholm aan de 'Stockholms Universitet'. Na een jaar de dienstplicht te hebben vervuld als radiotelegrafist bij het 43e AfdVa Bravo peloton in Havelte, heeft hij van 1994 tot 1995 gewerkt voor Vending@Work. Voor Vending@Work heeft hij een vestiging opgezet in Bochum, Duitsland. In 1995 is hij begonnen in de ICT bij Vertis. Gedurende de 10 jaar bij Vertis heeft hij verschillende functies vervuld: business analist, business intelligence consultant, Oracle DBA, projectleider van maatwerk en Oracle EBS applicatieontwikkeling, manager software ontwikkelstraat en Oracle EBS consultant. In februari 2006 is Talko in dienst gekomen van de Hanzehogeschool en heeft hij verschillende rollen vervuld zoals docent business intelligence, informatie architect, hogeschooldocent Business IT & Management, coördinator Honours, onderzoeker New Business & ICT en hoofddocent van de leergang Data Science.

In september 2016 begon Talko aan zijn promotieonderzoek. Dit promotieonderzoek gaat over de rol van nieuwe data-analyse en voorspeltechnieken zoals het toepassen van machine learning bij het monitoren van fysieke activiteiten in het dagelijks leven en sport. Het doel van het onderzoek is het vroegtijdig voorspellen van een fysieke prestatie zodat tijdig bijgestuurd kan worden. Dit onderzoek is uitgevoerd binnen het lectoraat New Business & ICT en voor het Marian van Os Center of Expertise Ondernemen van de Hanzehogeschool. Het onderzoek is in de eerste twee jaar voornamelijk door de onderzoeksgroep Distributed Systems, onderdeel van de Factulty of Science & Engineering van de Rijsksuniversiteit Groningen, begeleid met als tweede instituut Human Movement Sciences van de Rijksuniversiteit Groningen. Na twee jaar is de begeleiding komen te liggen bij Human Movement Sciences. Een deel van het onderzoek heeft plaatsgevonden bij de Shanghai Polytechnic University, Shanghai, China in 2018. Het verblijf aan de Shanghai Polytechnic University is gecombineerd met een gastdocentschap bij de Waikato University, Hamilton, New Zealand, 2018. Het promotieonderzoek is in 2024 afgerond.

Tijdens de onderzoeksperiode zijn, naast de werkzaamheden voor de Hanzehogeschool en promotieonderzoek, ook de courses, 'Publishing in English', 'Ethics of Research and Scientific Integrity for Researchers' en 'Research Data Management awareness workshop' gevolgd en succesvol afgerond.

Vanaf september 2023 is Talko de ankerman voor onderzoekslijn Data Gedreven Zorg, de hogeschooldocent voor Business IT & Management, Deeltijd HBO-ICT en de hoofddocent Data Science van de IT-Academy Noord Nederland.

List of publications

Dijkhuis, T. B., Blaauw, F. J., Van Ittersum, M. W., Velthuijsen, H., & Aiello, M. (2018).
Personalized physical activity coaching: a machine learning approach. Sensors, 18(2), 623.

Dijkhuis, T. B., Otter, R., Aiello, M., Velthuijsen, H., & Lemmink, K. (2020).
Increase in the acute: chronic workload ratio relates to injury risk in competitive runners.
International journal of sports medicine, 41(11), 736-743.

Dijkhuis, T. B., Kempe, M., & Lemmink, K. A. (2021).
Early prediction of physical performance in elite soccer matches—A machine learning approach to support substitutions.
Entropy, Data Analytics in Sport Sciences: Changing the Game, 23(8), 952.

Dijkhuis, T.B., Blaauw, F.J. (2022).
Transferring targeted maximum likelihood estimation for causal inference to sport science.
Entropy, Data Analytics in Sport Sciences: Changing the Game. 24(8), 1060

Dijkhuis, T.B., Franke, H.H., Paans, W., Dieperink, W., Groenboom, R.M. (2023)
Nutritional status of the ventilated COVID-19 patient in the Intensive Care Unit.
(Submitted)

Conference papers in combination with presentations

Blok, J., **Dijkhuis, T.**, & Dol, A. (2017).
Toward a generic personalized virtual coach for self-management: a proposal for an architecture.
In 9th International Conference on eHealth, Telemedicine, and Social Medicine 2017.

Dijkhuis, T.B., Otter, R., Velthuijsen, H., & Lemmink, K. (2017).
Prediction of running injuries from training load: a machine learning approach. In 9th International Conference on eHealth, Telemedicine, and Social Medicine 2017.

Best paper award

Dijkhuis, T.B., Blok, J., & Velthuijsen, H. (2018).
Virtual coach: predict physical activity using a machine learning approach.
In 10th International Conference on eHealth, Telemedicine, and Social Medicine 2018.

Van Noppen – Kleist, K., Mulder, W, **Dijkhuis, T.B.**, Dam, M.R. (2019)
Virtual coach: towards personalized mental support

The Eleventh International Conference on eHealth, Telemedicine, and Social Medicine 2019.

Dijkhuis, T.B., Blok, J., Velthuijsen, H., Lemmink, K.A.P.M. (2019)
Taking the difference between leisure time and workdays into account to improve virtual coaching, The Eleventh International Conference on eHealth, Telemedicine, and Social Medicine 2019.

Chair on conferences

Chair and coordinator of a special track: DataVCoach: data driven personalized virtual coaching, The Eleventh International Conference on eHealth, Telemedicine, and Social Medicine 2019.

Conference presentations

Dijkhuis, T.B., Dam, M.R. Enabling Automated Coaching, Conference Health By Tech, Groningen, 2019

Otter, R., **Dijkhuis, T.B.**, Van der Worp, H., Velthuijsen, H., Aiello, M., Lemmink, K.A.P.M, Brink, M.
Patroonherkenning in trainingsbelasting om blessures van getrainde hardlopers te voorspellen, DSO: Samen kennis maken, Zwolle, 2017

Kempe, M., **Dijkhuis, T.B.**
Coach, I don´t feel it today- A Machine Learning approach for early in-game performance prediction, World Congress on Science and Soccer, Ciombra, 2022

Presentations

Dijkhuis, T.B., Injury Risk in Running and Machine Learning, Waikato University, Hamilton, New Zealand, 2018

Dijkhuis, T.B., Blok, J., Data Science, IT Academy Noord, Groningen, 2017

DANKWOORD

Het is gelukt, het proefschrift is er! Er zijn een heleboel mensen die hebben meegewerkt aan het bereiken van deze mijlpaal. Al deze mensen wil ik graag bedanken. Ik ga mijn uiterste best doen om iedereen een plek te geven in dit dankwoord. Mocht ik iemand zijn vergeten bij naam te noemen dan spijt mij dat, alsnog bedankt voor je bijdrage!

Begeleiding

Allereerst wil ik mijn (co)promotoren bedanken voor de prettige en goede begeleiding.
Prof. Dr. K. Lemmink, Koen, dank voor je geduld en het effectieve begeleiden bij het schrijven van artikelen en het proefschrift. Je inzichten, aanwijzingen en commentaren en vragen hebben mij zeer geholpen om dit traject tot een mooi einde te brengen. Bij de artikelen was je erg sterk in het eenduidig maken en structureren. Ook je belangstelling als het privé even wat zwaarder was, heeft me geholpen om toch de draad weer op te pakken als ik het niet zag zitten. Ondanks je drukke werkzaamheden heb je gestructureerd effectieve feedback gegeven en altijd vanuit een positieve grondhouding. Ik heb dit enorm gewaardeerd.

Prof. Dr. M. Aiello, Marco, in de eerste twee jaar was jij mijn eerste promotor en hebt m'n eerste schreden goed begeleid met veel ruimte voor eigen invulling. Je was altijd snel bereikbaar en altijd bereid tot een overleg. Ook als je je om 00.00 mailt krijg je binnen een half uur antwoord. In het bijzonder herinner ik mij nog een overleg tussen jou Frank en mij waarbij jij in Australië zat, Frank in Amerika en ik in Nederland. Door allemaal wat te schuiven konden we toch een tijdstip voor overleg vinden. Niet alleen mijn eerste schreden, ook nadat je tweede promotor werd, heb je met raad en daad terzijde gestaan en goede feedback gegeven. Dank daarvoor!

Dr. H. Velthuijsen, Hugo, dankzij jou ben ik dit avontuur begonnen. Je had een welgemeend goed advies een aantal jaren geleden. 'Stop met vechten tegen de bierkaai en speel mee op niveau.' Tijdens mijn burn-out heb je me onder je hoede genomen en zodra je een kans zag bij de Hanze om mij een gesponsorde promotieplek te kunnen geven heb je dat gedaan. Ondertussen zijn we alweer 8 jaar verder. In de tussentijd hebben we vele (ook informele) gesprekken gevoerd en heb je mij geholpen stappen te zetten. Daarnaast heb je al m'n stukken gelezen en voorzien van commentaar op de lijn van het verhaal en verbetering van het Engels. Hopelijk blijven we elkaar vaak treffen!
Ik wil de beoordelingscommissie met Prof dr. J. Kok, Prof dr. J. van Gemert-Pijnen en Prof dr. M. Biehl hartelijk danken voor het beoordelen van het proefschrift en het deelnemen aan de promotiecommissie. De overige leden van de oppositie wil ik eveneens hartelijk danken voor hun deelname aan de promotiecommissie.

Paranimfen

Paranimfen Frank en Wico, jullie wil ik bedanken dat jullie mijn paranimfen willen zijn, en vooral voor de ondersteuning en de samenwerking in de afgelopen jaren, zeker bij de laatste loodjes. Frank, je warme welkom bij Distributed Systems maakte dat ik graag op de Bernouilleborg kwam. Toen je gepromoveerd was en de kamerindeling wijzigde, werd mijn motivatie toch wat minder om in de Bernouilleborg aan de slag te gaan. Samen hebben we twee artikelen geschreven. Het was een voorrecht om met jou te werken, je bent superaardig, intelligent en betrouwbaar en ik heb daarnaast ook veel van je geleerd over het puntjes op de i zetten. Een derde gezamenlijk artikel waren we aan begonnen maar onze counterpart in Parijs maakt het allemaal wel heel ingewikkeld. Uiteindelijk hebben we maar eieren voor ons geld gekozen. Wico, je bent altijd een bruisende bron van energie waar ik niet alleen inhoudelijk mee kan sparren maar ook over andere zaken van gedachte kan wisselen. Van AI tot ZXSpectrum, van bier tot ouders. We hebben samen nog een conferencepaper ingediend, helaas kon je toen niet mee naar de conference maar ik ben ervan overtuigd dat we dat in de toekomst nog eens een keer gaan doen. Daarnaast maakte je altijd tijd vrij om te spiegelen of mee te denken over, bijvoorbeeld, de introductie en het achterhalen van de onderliggende structuur in alle hoofdstukken. We gaan ongetwijfeld nog onder het genot van een versnapering de toekomst en andere zaken bespreken.

Medeauteurs

Alle medeauteurs anders dan mijn (co)promotoren of paranimfen, wil ik danken voor het samenwerken. We hebben samen mooie artikelen gemaakt. Soms was een artikel in een keer geplaatst, sommige artikelen vergden wat meer rondes. Miriam, we met het eerste artikel van dit proefschrift geschreven op basis van de data die jij verzameld hebt, je had goede input op het coachings-deel. Matthias, het voetbalartikel was op basis van jouw idee en het schrijven in de Coronatijd was het niet altijd makkelijk. Ik zie je in de online meetings nog zitten in de kamer zonder raam met je kano als achtergrond. Je hebt me geholpen de voetballerij te doorgronden en door te zetten. Ruby, het was worstelen met de hardloop data en de blessures, je was oneindig optimistisch, ook als we elkaar niet begrepen. Na vele omzwervingen met de data en bewerkingen, eerst wel machine learning toen niet meer en een paar conference papers is er uiteindelijk toch een goed journal paper tot stand gekomen, dank!

Lectoraat

Bij de Hanzehogeschool ben ik onderdeel van het lectoraat Digitale Transformatie en de leeropdrachten New Business & ICT en Digital Health. Bij New Business & ICT was eerst Hugo lector en promotor, later Klaas† lector en nu Rix als lector en Hilbrand als lector van Digital Health. Rix en Hilbrand geven mij persoonlijk en inspirerend de kans om met de onderzoekslijnen Data Gedreven Zorg en Point of Care technologie een waardevolle

bijdrage aan de maatschappij te leveren. Ook alle andere kenniskringleden dank voor jullie belangstelling en ondersteuning in de afgelopen jaren! Aranka, we hebben uiteindelijk de eindstreep gehaald; Austin, dank voor het doornemen van mijn artikelen en je aanwijzingen; Harald, de virtual fitness coach vormt al jaren een rode draad in ons bestaan; Herman, je hebt me veel geleerd over hartritme variabiliteit en ook dank voor gezelligheid bij het congres in Athene; Jan Baljé, op het gebied van promoveren hebben we niet veel samengewerkt maar des temeer op het ontwikkelen van een minor AI en nu het vormgeven van het lectoraat en het aanvragen van subsidies, samen af en toe het rondje lopen om stoom af te blazen en nieuwe ideeën op te doen waardeer ik zeer; Anniek; Jantine; Johan; Rick, en Roland.

Collega's van HBO-ICT

Er zijn vele teamleiders gekomen en gegaan bij HBO-ICT, een drietal wil ik hier bedanken voor het ondersteunen en helpen bij het promoveren. Anke, jij was de eerste die mij motiveerde om te gaan promoveren. Dat de eerste poging uiteindelijk strandde, lag aan de persoonlijke omstandigheden waar jij mij in ondersteunde en alle begrip had dat ik ergens anders mijn energie voor nodig had. Marianne, jij hebt mij veel ruimte gegeven om het promoveren goed te kunnen doen en in Corona tijd ook regelmatig een wandeling gemaakt om de motivatie te behouden. Sake, je hebt regelmatig geïnformeerd en gestimuleerd om mijn promotie af te ronden. Ook gaf je mij de ruimte om gedurende het jaar af en toe een week onder te duiken om aan de promotie te werken, dank hiervoor. Verder natuurlijk alle collega's van HBO-ICT en in het bijzonder van het expertteam BITM en Curriculumcommissie dank voor het dulden van mijn gedeeltelijke afwezigheid voor het onderwijs. We gaan verder met de weg omhoog die we een aantal jaren geleden hebben ingezet!

School of ICT

Hangend aan nostalgie komen we nog jaarlijks bij elkaar om de goede oude tijd te laten herleven. Ada, Bart, Denise (onze afspraken gaan vaak niet door maar we blijven afspraken maken ;-)), Dineke, Jos Bos (eens per maand halen we samen met Bart ook nog weleens herinneringen op, dat gaan we blijven doen!), Jos Bredek (ook weer eens samen een hapje doen?), Henk T. (jammer dat onze International Programs zijn afgelopen, de bierproeverij was legendarisch), Jacob, Jan Baljé, Nienke en Sharon laten we vooral bij elkaar blijven komen. Al 15 jaar gaan wij, Froukje, samen in het kader van internationalisering op pad, soms met z'n tweeën soms met meer. Tijdens deze trips hebben we lief, leed, werk en gezelligheid gedeeld. Dank voor je gezelligheid en steun, we blijven elkaar zien.

Vrienden

Bart, goede vriend, sinds 14 jaar hebben we veel lief en leed met elkaar gedeeld, onder andere verloren gegane relaties, huwelijken, teloorgang van dierbaren, kennisdelen, soms een boek bespreken. Maar vooral activiteiten op het gebied van socialiseren met bier, hardlopen en fietsen. Hopelijk blijven we nog lang heel op de fiets en gaan we jaarlijks door met onze Eroïca hobby.

Manfred, buurman, goede vriend en co-ouder, je hebt vele verhalen aangehoord over mijn promoveren en je hebt veel advies gegeven. Maar waar we echt vriendschap hebben opgebouwd is de tijd dat de kinderen jong waren en dat wij de kinderen op hetzelfde moment aan het verzorgen waren. Vaak hebben we samen gegeten en elkaar ondersteund in de dagelijkse zorg. Daarnaast hebben we menig avond aan de keukentafel doorgebracht om het leven te bespreken. Dat gaan we blijven doen, dank voor alle steun en luisterend oor!

Marc, goede vriend, het leven gaat niet altijd over rozen, maar uiteindelijk komen we wel weer op het rechte pad. Vele avonden hebben we samen gegeten en de stand van zaken doorgenomen. Vaak kunnen we aan elkaar spiegelen en over en weer adviseren. Het is druk geweest het afgelopen jaar gelukkig hebben we elkaar nog regelmatig kunnen treffen, gegeten moet er toch altijd. Laten we dat vooral blijven doen!

Petra & Arjan, Petra, je bent een dierbare vriendin, onze vriendschap gaat terug tot aan de eerste jaren van de studie in Groningen. Altijd ben je een trouwe vriendin geweest. Al gaan we nu niet meer naar hardrockconcerten van bijvoorbeeld Monsters of Rock of Aerosmith, we hebben genoeg basis om het zonder concerten te doen. Heel regelmatig zijn er logeerpartijen bij jou en Arjan in Wijhe met warme belangstelling en ondersteuning voor elkaar. Bij mij is het door de kattenallergie van Arjan wat lastiger ;-), maar fietsen bij Wijhe met Arjan is ook beter uitdagender dan in Groningen. Dank voor de vriendschap en vele goede herinneringen!

Bram, Frits, René en Joost, ongeveer 35 jaar na onze middelbare school hebben we nog steeds een goede vriendschap en heel veel historie. We gaan nu nog samen naar concerten, doen af en toe een hap en sporten heel soms. Dit alles is bij jou, Joost, niet meer het geval omdat je aan de andere kant van de wereld zit maar na 15 jaar buitenland is nog altijd de band er en voel ik me welkom. Ook toen ik in Nieuw-Zeeland was om een week visiting researcher te zijn bij de University of Waikato, Hamilton. Laten we met z'n allen vooral onze vriendschap behouden.

Al roeien we al 30 jaar niet meer, Berthold, Frank, Ronald en Tom, we komen alsnog jaarlijks bij elkaar om het leven door te nemen. En proberen ook op hoogtepunten in

elkaars leven aanwezig te zijn. Berthold, wij ons jaarlijkse wandelweekend, dat heeft in ieder geval al geleid tot het lopen van het gehele Pieterpad in 15 jaar. En nu tot het lopen van verschillende lange afstandsroutes, altijd met een overnachting en genoeg om voor een jaar bij te praten ☺.

Wiebe en Annelieke, van de 15 jaar dat we hardlopen, hebben jullie acht jaar lang hebben jullie mijn promoveren beslommeringen aangehoord. Dank voor het bieden van een luisterend oor en alle ondersteuning. Het waren soms hoogtepunten maar soms ook klaagzangen Dat stopt (gelukkig?) nu. Wiebe, je hebt ons beiden het boek Nieuwe Wereld van Eckhardt Tolle gegeven. Dit heeft mij geholpen om meer ontspannen met allerlei kwesties om te gaan. Laten we vooral blijven lopen en het dagelijks leven bespreken, de halve marathon van Leek staat alweer in de agenda.

Familie

Lieve nichten, Herma, Janine en Lettie en partners, soms treffen we elkaar wat vaker in het leven, soms wat minder vaak. Maar we houden als kleine familie toch altijd waardevol contact en interesse in elkaar. Dank voor al jullie interesse en ondersteuning. We proberen de rest van de familie eens per jaar bij elkaar te laten komen. Oom Ben & tante Mieke, oom Martin & tante Tini, oom Joop & tante Willy, waarbij helaas oom Anton†, tante Mart† & m'n dierbare neef Martijn† overleden zijn. In het bijzonder wil ik oom Ben & tante Mieke, danken voor jullie steun, interesse en ook praktische ondersteuning als dat weer eens nodig was, en ook oom Martin & tante Tini voor de interesse en ondersteuning.

Annalie en Nico, Annalie, dank voor de prachtige kinderen die we hebben. Volgens de kinderen deed jij de opvoeding en ik de voeding. Ik ben het er niet helemaal mee eens, maar ja… De laatste jaren eten we met de kinderen om de twee weken samen. En soms ook zonder kinderen. We hebben onze relatie weten om te zetten in een waardevolle vriendschap waarbij we er allebei mogen zijn en je ook interesse hebt en ondersteuning hebt geboden bij het promoveren. Nico, jou wil ik ook bedanken voor de belangstelling en het koken op maandagavond ;-).

Mirjam, je hebt me aangespoord dit avontuur van promoveren aan te gaan. Je hebt de laatste fase niet meegemaakt. Maar de start wel mogelijk gemaakt, bedankt hiervoor.

Papa en Mama, zonder jullie was ik er niet geweest en dus ook niet dit proefschrift. Dank voor jullie goede basis, ondersteuning en liefde. Ondanks de ongemakken van een steeds hogere leeftijd is er nog altijd belangstelling voor en ondersteuning van mijn leven en promoveren. Ik hou van jullie.

Lieve Marion, steun en toeverlaat, in het dagelijks leven, mantelzorg en ook in de laatste fase van het promoveren. Je was er in goede tijden en ook als het eens wat minder ging. Dankzij je steun twee jaar geleden heb ik de motivatie weer gevonden om de schouders eronder te zetten. Ik geniet van ons leven samen en jouw kinderen, Jesper en Melle, die mij ruimhartig toelaten in hun leven. En je altijd belangstellende ouders Dick en Gré. Laten we vooral naast gewoon fijn leven, nog veel trips, vakanties en festivals doen. Lief, ik hou van je!

Nanko, lieve zoon, dank voor het op m'n huis passen, als ik weer eens een week aan het schrijven was. Je bent alweer een tijdje uit huis, maar bent betrokken en geïnteresseerd met veel humor waarbij ik ook regelmatig een spiegel voorgeschoteld krijg. Je bent een mooie vent die doorzettingsvermogen laat zien, ik hou van je en onze volgende trip gaat voor de derde keer naar Berlijn.

Janna, lieve dochter, je stapt avontuurlijk en reislustig door het leven. Ook thuis is het altijd levendig als je er bent en geniet ik van je aanwezigheid! Ik hou van je. Laten we nog vele dingen ondernemen. Ik wil afsluiten met een kleine conversatie van twee jaar geleden. Ik: Ik ga werken aan m'n proefschrift op Schiermonnikoog. Jij: Oh, je gaat op vakantie? Ik: Nee, ik ga werken aan m'n proefschrift. Jij: Dat is toch de vakantie voor jou: werken aan je promoveren en daarnaast een beetje relaxen. Dit is tekent voor hoe je hebt aangekeken tegen mijn promoveren. Het promoveren zit er nu op, nu kijken wat de toekomst gaat brengen.

RESEARCH INSTITUTE SHARE

This thesis is published within the **Research Institute SHARE** (Science in Healthy Ageing and HealthcaRE) of the University of Groningen. Further information regarding the institute and its research can be obtained from our internet site: https://umcgresearch.org/w/share.

More recent theses can be found in the list below.

2023

Boersema HJM
The concept of 'Inability to Work Fulltime' in work disability benefit assessment
(Prof S Brouwer, Dr FI Abma, Dr T Hoekstra)

Ots P
The role of individual and contextual factors in paid employment of workers with a chronic disease
(Prof S Brouwer, Dr SKR van Zon)

Kool E
Untangling the elements of midwives' occupational wellbeing: A study among newly qualified and experienced midwives
(Prof ADC Jaarsma, Prof FG Schellevis, Dr EI Feijen-de Jong)

Jansma FFI
Self-management in rehabilitation practice: On the design and implementation of a serious theory-based analogue problem-solving game called 'Think Along?'
(Prof R Sanderman, Dr I Wenzler)

Erpecum CPL van
The role of fast-food outlet exposure in Body Mass Index
(Dr N Smidt, Prof U Bültmann, Dr SKR van Zon)

Kerver N
The effectiveness and cost-effectiveness of upper limb prostheses
(Prof CK van der Sluis, Dr RM Bongers, Dr S van Twillert)

Deviandri R
Management of anterior cruciate ligament injury in lower-middle income countries: Focus on outcomes and health economics in Indonesia
(Dr I van den Akker-Scheek, Prof MJ Postma, Dr HC van der Veen, Dr Andri MT Lubis)

Mangot Sala L
Disruptive Life Events and Health: Longitudinal evidence from a large cohort in the Netherlands
(Prof AC Liefbroer, Dr N Smidt)

Wijk DC
From prosperity to parenthood: How employment, income, and perceived economic uncertainty influence family formation
(Prof AC Liefbroer, Prof HAG de Valk)

Dai Y
Effects of exposure to polycyclic aromatic hydrocarbons and heavy metals on placental trophoblasts and childhood inflammation
(Dr MM Faas, Prof X Xu, Prof X Huo)

Menting SGP
Picking up the pace: The development of pacing behaviour during adolescence
(Dr MT Elferink-Gemser, Prof FJ Hettinga)

Vos M
My name is legion for we are many: Lessons learned from linking and splitting psychiatric Disorder
(Dr CA Hartman, Prof NNJ Rommelse)

Haan-Du J De
Cancer risk, stage, and survivorship among patients with type 2 diabetes
(Prof GH de Bock, Dr GWD Landman, Dr N Kleefstra)

Nieboer P
Teaching and learning in the operating room: Navigating treacherous waters
(Prof SK Bulstra, Prof M Huiskes, Dr M Stevens, Dr F Cnossen)

He Z
Risk factors for elevated blood pressure: focus on perimenopausal women and potential causality
(Prof H Snieder, Dr CHL Thio, Prof QYZ Qingying Zhang)

Peeters CMM
Brace therapy and radiographic imaging in adolescent idiopathic scoliosis; where do we stand?
(Prof PC Jutte, Dr C Faber, Dr FH Wapstra, Dr DHR Kempen)

For earlier theses visit the website: Find Research outputs — the University of Groningen research portal (rug.nl)


Research Institute
SHARE