A Descriptive Statistical Analysis of Overweight and Obesity Using Big Data

Salam Abdulabbas Ganim Ali
Directorate-General for Education
Department of Education REFAEE
The Republic of IRAQ
Ministry of Education
salamalghanim@gmail.com

Hayder Rahm Dakheel AL-Fayyadh
Department of Computer Science
University of Sumer
Thi-Qar, Iraq
haiderrahm@uos.edu.iq

Shaimaa Hadi Mohammed
Department of Computer Science
University of Sumer
Thi-Qar, Iraq
shsumer@yahoo.com

Saadaldeen Rashid Ahmed
Computer Science
Altinbas univerity
University of Tikrit
Karabuk University
Istanbul, Turkey
Saadaljanabi95@gmail.com

2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA) | 978-1-6654-6835-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/HORA55278.2022.9800098

***Abstract*—In this paper, we have obtained the dataset from an open-source repository for obese people by focused on a descriptive statistical analysis of overweight and obesity using big data. We performed the statistical analysis on large scale streaming data for obesity prediction. We have classified the obesity with all categories on the scale of Body Mass Index (BMI) is being calculated i.e., underweight, normal weight, overweight, obese, very obese, and extremely obese using MapReduce technique with the help of Apache Spark and Apache Hadoop engine in pydoop python programming. The MapReduce technique in-volves the updating of cluster centers after arrival of new batch in the stream of data. The streaming of data is produced by the sensors which are classified into six different BMI categories, which are stored and processed through big data tools connected to the statistical analysis system. The Apache spark produces the latency values in accessing the data from dataset. We analyzed any obesity in the people from the normal latency value using the Apache spark and Hadoop which are well known in big data. The methods and techniques by which we can predict obesity efficiently from the large-scale streaming data has been per-formed using python programming. This is applied with the help of Apache Spark and Hadoop. In order to validate the efficiency of MapReduce technique. We have tested it both on single and distributed environment for obesity prediction using the built-in Pydoop package in python.**

Keywords— Big data, Risk Management, process management . Apache, Pydoop.

I. INTRODUCTION

In recent years, obesity has gained recognition as a premier public health issue in the world due to large number of deaths each year. The Center for Disease Control (CDC) reported that more than one-third of adults are obese. We use the cross-sectional obesity prevalence data and account for a source variation for the combined time and age effects, usually known as the cohort effect [1]. The measurement and inclusion of cohort effect in various models have gained much interest recently in academic literature. In this study, we take advantage of the knowledge and academic tools in stochastic modeling to fit the curvilinear relationship of obesity prevalence and age for the cross-sectional observed obesity prevalence. We analyze the obesity trend based on25 years of data for ages 23 to 90 provided by CDC's Behavioral Risk Factor Surveil-lance System (BRFSS) survey and fit proposed models [2]. We assume a polynomial regression and make initial estimates using Ordinary Least Squares (OLS) regression; then we adjust for the cohort effect using an iterative algorithm to fit the quadratic structure of obesity prevalence [3]. Finally, we also make use of time series and fore-casting techniques to forecast obesity prevalence using big data analysis.

Researchers in medical sociology and epidemiology are often interested in big data analysis for the distribution and etiology associated with a large variety of health is-sues. Estimating and predicting obesity prevalence for the world population has been an important concern for researchers working in this area [4]. Significant financial and pension planning consequences could arise due to the changes in mortality rates. This is because mortality and life expectancy affect the financial projections and sustainability of the social security system. Insurance and investment firms use mortality rates to forecast liabilities and amounts required to be booked in periodic financial reporting [5]. Recently, many life and health insurance firms have begun to include obesity as a major risk factor in their ratemaking and life insurance product pricing. It has become very common for the prospective insured to provide measurements of their height and weight during the underwriting phase of the life insurance sales cycle. This information about an individual's obesity is included in the pricing of premiums payable for various life insurance and annuity products.

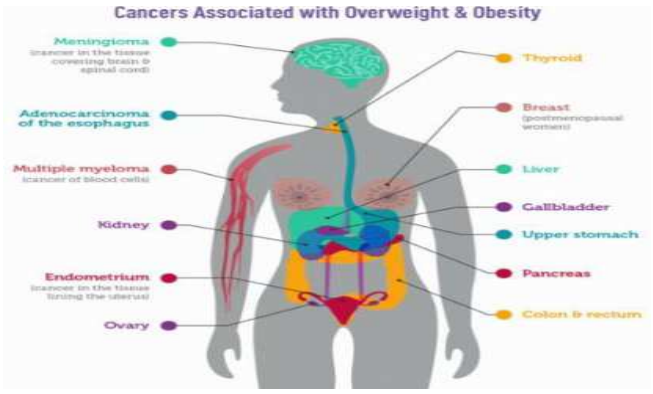

Fig. 1. Sensor operation range and line-of-sight detection [5].

978-1-6654-6835-0/22/$31.00 ©2022 IEEE

In big data analysis, the same individuals are studied over time. However, it is difficult to measure big data not necessarily for the same individuals but for individuals who are the same age. Samples taken each year and grouped by ages and studied over time. Ideally, if the same individuals could be followed over time, the obesity effect would be measured more accurately and inadvertently lead to a measuring of the cohort effect since the same individuals would be followed over time. However, this is usually not the case. Obesity prevalence is estimated through random periodic surveys of samples of individuals from the population who are usually different from one period to another using the big data analysis.

A. Problem Statement

Obesity is a growing concern in both developed and developing countries and it has gained recognition as a premier public health issue in the world due to large number of deaths each year. Findings obesity in obese people, inadequate physical exercise as a potential factor to this, particularly in industrialized countries, and identified seasonal fluctuations in work activity as a potential contributor and declining activity without corresponding increases in obesity of people, the obesity effect would be measured more accurately and inadvertently lead to a measuring of the cohort effect using the Big Data techniques that would be followed over time, we also make use of spark and Hadoop frameworks to evaluate the obesity after the MapReduce has classified the obesity.

B. Research Contribution

In this paper, as spark has become more widely used, performance problems have been exposed in its use in practical applications for prediction of obesity. The dataset for obesity and weight in humans using BMI. For obesity prediction analytics, MapReduce algorithms can be used and feeding it to Apache spark for the evaluation purpose because it has very good pre-defined libraries for the assessment of data on large scale. Visual reports/graphs will be displayed using different statistical analysis of different weights by their BMI index using MapReduce by iteratively running it on different set of data for the prediction of obesity into six different BMI based categories. Big Data methods are seen as a key way for material value creation in various domains such as obesity prediction, but the technology landscape is highly dynamic with varying levels of maturity for obesity. Hence, to understand the technology land-scape better, discuss the models that seem to have significant benefit, and point out the flaws, if any, is one of the major factors for this paper. We also try to understand and identify the current important principles of different paradigms of analytics and different data pipeline methods for obesity prediction using the spark and Hadoop

II. RELATED WORK

Twelve articles focused on factors that contribute to obesity in people. The majority of studies supported physical activity as a factor influencing obesity in obese persons. The level of occupational physical activity (PALs) among obese adults utilizing classic obesity measures remained moderate (1.90) and mean body mass indexes (BMI) remained within normal limits[6]. The introduction of modern agriculture practices and mechanization, however, People with higher BMIs and less physical labor had lower BMIs. Modern technology has changed food patterns as well as physical activity [7]. BMI and non-communicable diseases grew as diets became more fattening and processed than the typical diet of the area

Fig. 2. Different ranges of BMI with the classification of wights [7].

Over the last decade, workplace obesity has become a growing concern. Obesity is on the rise in both industrialized and developing nations, With exceptionally high per-centages in predominantly agrarian countries like Tonga, Samoa, and Kuwait [8]. The present state of information about the prevalence is summarized in this report, Obesity's impact on health, safety, and work performance in obese persons are all aspects to consider. [9]. Obesity prevalence rates and patterns are comparable to obesity rates and patterns in the broader workforce. However, given the recognition of the multi-factorial nature of the etiology of obesity, more study into other occupational-related exposures linked to obesity and obesity-related diseases is required, including job-related stress.

A. Cause and Effects of Obesity

Obese people's health and well-being are critical to a sufficient food supply as well as the economic well-being of communities, nations, and the planet. Despite the fact that obesity rates among obese persons have been rising, Only a small amount of research has been done to determine the consequences of obesity on people's ability to work, People's safety or productivity [10]. Because of the nature of big data labor, it frequently involves physically demanding and time-sensitive tasks, reducing the capacity to schedule, slow down, or postpone work due to physical limitations. More research is needed to fill in the gaps in our understanding of how obesity affects the efficiency and safety of big data work. The impact of obesity could serve as a stimulus for better health habits to help people lose or maintain weight, lead the development of complete worker health programs for the obesity industry, and identify solutions for changing work procedures to improve obesity-friendly safety and efficiency.

Fig. 3. Different causes and effects of obesity [10].

B. *Potential Data Bridges*

Aside from these flaws, the literature also has major flaws that could contribute to bias, such as the following: a) research designs that are poor to moderate; b) samples that aren't representative; c) self-reporting and recollection; and d) misclassification. The majority of the research looked at was cross-sectional, descriptive, or cohort studies involving a variety of populations, limiting the capacity to generalize or compare the results. Obesity's heterogeneity, livestock vs. modern vs. traditional big data approaches, its size creates a variety of demands and risks, limiting the capacity to gen-realize findings. Furthermore, over half of the research were secondary analyses of data from major national or regional surveys that did not have obesity as a key emphasis, Obese people made up a small percentage of the population. This makes it much more difficult to account for or incorporate heterogenic elements that may in-fluence study outcomes into investigations. There were various potential sources of selection bias in the literature that could lead to non-representative samples, according to the review. The AHS database was used in 42% of the Turkey. people-related studies [11], It is only available to pesticide applicators who have been granted a license. Other possible targets of selection bias are however present. Inclusion criteria for research frequently centered on those who were actively employed. As a result, This adds the healthy worker effect, which can have an impact on study outcomes and excludes those who may have stopped being obese owing to obesity-related health consequences. Additionally, Obese people may not be represented in all samples taken from major national surveys when people self-identify as obese. Obesity is not the primary or sole occupation of many fat persons. These fat people work both on and off big data, and their off-big data job may be listed as their principal job. This could lead to their exclusion from the study or their use as a control when they share the same exposure.

Fig. 4. Description of data bridges that need immediate attention in biomedical [11].

III. METHODOLOGY

In this research work, the lack of ability to work, on the other hand, has serious ramifications for obese persons, the fat family, as well as society as a whole. Obese individuals place a low value on their ability to work, linking it to their sense of self, health, well-being, and life quality. In Turkey, the vast majority of individuals (97%) are owned by their families [11]. Two-thirds of the labor force in Turkey is made up of family members and obese people, as a result of the loss of employment ability, family members may be burdened with more tasks. Additionally, Loss of job ability can result in a drop in productivity and money, as well as the loss of personnel. Reduced work ability can have a negative influence on the community and the country owing to decreased job participation, loss of expertise, and revenue loss. People's work abilities must be maintained and promoted in order for the agricultural industry to be sustainable and national and global food security to be ensured. The use of big data as a measure of work competence was substantially validated by hypothesis testing. Given the influence of one's health on one's ability to work, the link with self-reported health status suggests that work ability should be the construct being tested. In addition, the inverse link between measurements of job restrictions and productivity suggests that big data can be used to discriminate amongst workers. The inverse link between measurements of job restrictions and productivity suggests that big data can be used to discriminate amongst workers. Our findings support the use of big data as a tool for detecting and intervening to improve people's job abilities and productivity. However, The homogeneity of the sample and the limited sample size of this study limit the capacity to transfer these findings to other groups for obesity prediction. MapReduce technique is being used for the classification of obesity and the predictions are being fed into the Apache spark and Hadoop for the evaluation. The big data is in large volume stored with velocity, variety and volume however dataset is a collection of relevant sets of rows and columns. There are many types of datasets i.e. machine data, open data, structured, semi-structured, time-stamped and unstructured data.

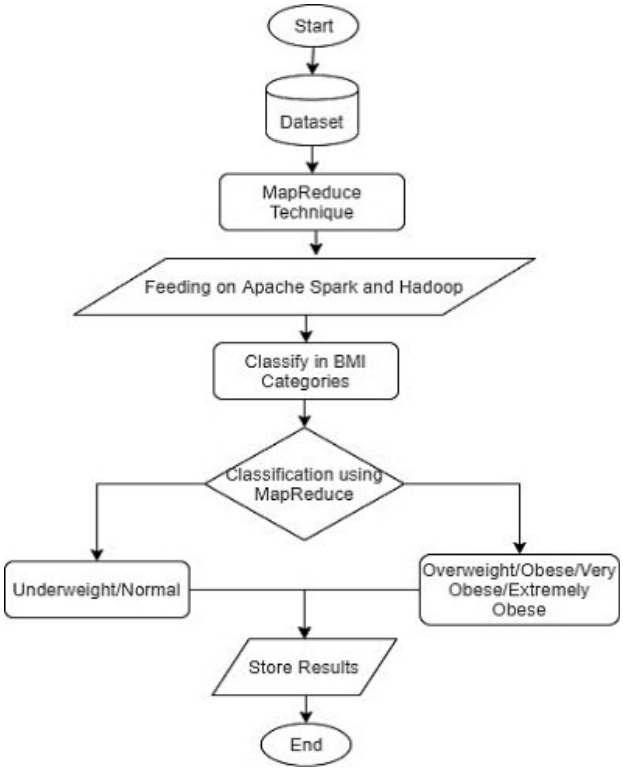

Fig. 5. Flowchart of methodology.

A. *MapReduce in Obesity Prediction*

Prior research has usually revealed characteristics linked to subjective and objective work ability assessments. The factors revealed in this study appear to better reflect the worker's physical and mental resources. As such, approximately half of the reported work ability is explained by these resources, this strengthens the ability to work. The model's comprehension of the worker's resources is the cornerstone of job ability. These finding also support the use of the Big Data in identifying and developing worker resources to meet the demands of work. The most important

libraries that have been used for the statistical analysis of overweight and obesity using big data are Spark Streaming for

cleansing the data, Spark SQL for storing the data, Pydoop for processing the data and GraphX for visualizing the results. All these components can be used on one project. These all components are scheduled, monitored and distributed to spark cluster by spark core. Spark supports its libraries in many languages like Scala, java, python, and R. Spark can be executed over Hadoop clusters, therefore it can also be integrated with other big data tools.

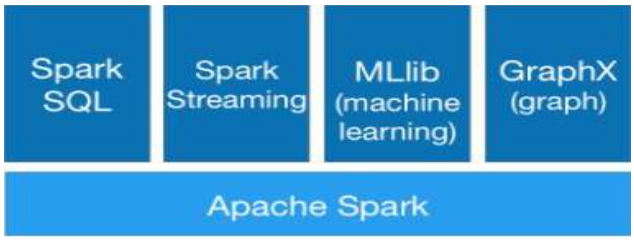

Fig. 6. The secure frameworks of Apache spark being used with distributed system of big data.

B. Experimental Approach

This paper involves both batch and Big-Data analysis using Apache spark and Hadoop. In case of batch processing, we follow below three steps:

- Loading the dataset: The dataset for obesity and weight in humans using BMI. It will be loaded to the spark engine for analysis.
- Processing or analysis: For obesity prediction analytics, MapReduce algorithms can be used and feeding it to Apache spark for the evaluation purpose because it has very good pre-defined libraries for the assessment of data on large scale.
- Visualization/classification: Visual reports/graphs will be displayed using different statistical analysis of different weights by their BMI index using MapReduce.

IV. Apache Spark and Hadoop Engine in Prediction

Apache spark and Hadoop are great big data-based engines for obesity prediction for all of domains and a prediction for the obesity in human beings. Some techniques have already taken place, but the knowledge about spark is like the technology itself in an early phase of adoption for the predication of obesity. In literature, the call for knowledge about Big Data is ever present. Some reports expect a large gap between available data scientists and available places for them in the industry, identifying a deficit in the food in the future. The Hadoop states the obesity age of the technologies that are based on Big Data. More research on the topic is necessary, about both its technical and non-technical aspects. Non-technical papers and reports are mainly aimed at businesses, as they are currently the most interested in Big Data. Research at the dynamics behind data analysis has not been conducted a lot yet.

TABLE I: MEASUREMENT OF DIFFERENT CATEGORIES FOR BODY MASS INDEX IN HUMANS.

CATEGORIES	BODY MASS INDEX
Underweight	10.0 - 17.50
Normal	17.50 – 25.0
Overweight	25.0 - 30.0
Obese	30.0 – 35.0
Obese (Class-1)	35.0 – 40.0
Very Obese (Class-2)	40.0 - 45.0
Extremely Obese (Class-3)	45.0 – 50.0

Given the fact that Big Data comes with high expectations, and the fact that it either is a disruptive technology or an enabler for disruptive technologies of spark and Hadoop, it values for obesity and weight is apparent. It seems the time has come to big data frameworks for the same fate as Blockbuster. Successfully implementing Big Data Analytics is something most obesity reduction will have to do soon. This method indicates inductive research, i.e. there is no hypothesis that is being tested. The result of the research will be a list of factors which obesity reduction will have to consider before starting their Big Data Analytics in obesity and overweight prediction and statistics using BMI.

V. RESULT

This paper provides the Big Data Pre-processor provides the user with a library of operations to perform manipulations and transformations on the selected data. The data is normalized and cleansed automatically using the spark streaming library in python the data gets cleansed and normalized with an additional removing of redundant rows as mentioned in the methodology section. It includes steps such as data-cleaning, normalization, integration, feature sub-set reduction, etc.

Big Data Cleaning: The primary aim of this function is removing inconsistencies, redundant and irrelevant data for obesity. It allows to remove missing values, smoothing out noisy data and correct inconsistent data.

Big Data Integration: It enables integration of data from multiple sources (data files) into a single dataset by joining on attribute values for obesity.

We have analyzed that the user is done with all the steps to infer meaningful results from the big data streams for obesity, the framework provides an option to the user to save all the steps performed. The saved file contains the start-to-end steps per-formed by the user such as which big data stream to select, what operations to perform, which task to perform, which Map-reduce model to select along with model hyper-parameters for obesity prediction. Using this stored file, the user will have the option to perform all the steps with just a single click on the entire dataset, i.e., all the steps will be performed automatically in the background without any user-intervention for obesity prediction.

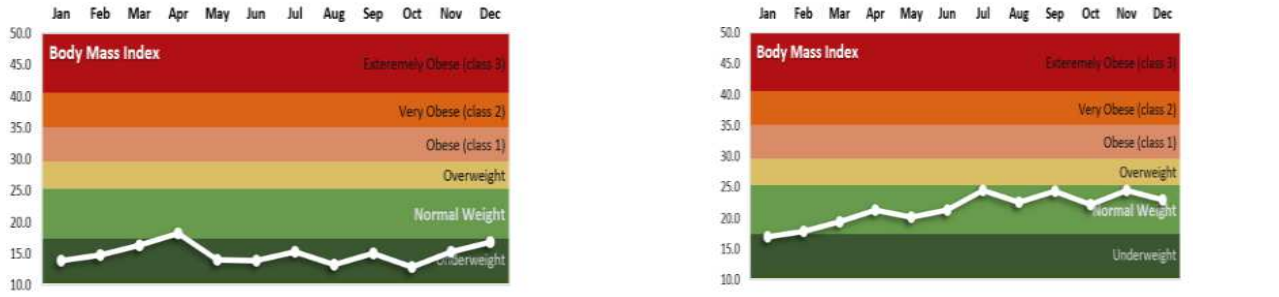

Fig. 7. Line represents the weight tracking for underweight person and normal weight per-son using big data on Apache spark and Hadoop engine.

Fig. 8. Line represents the weight tracking for overweight person and obese person using big data on apache spark and Hadoop engine.

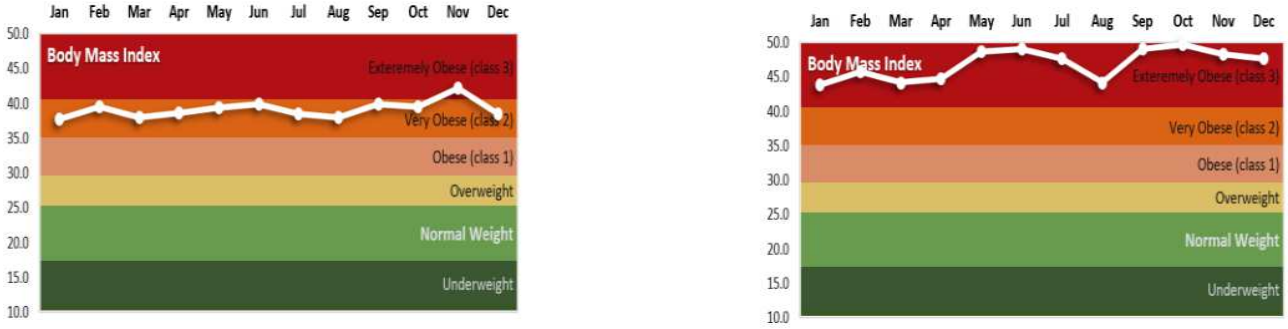

Fig. 9. Line represents the weight tracking for very obese person and extremely obese per-son using big data on apache spark and Hadoop engine.

VI. DISSCUSSION

This research has ramifications for big data analysis, particularly for advanced practice nurses in metropolitan regions, as well as for workplace health coaching and counseling to enhance the health of overweight and obese persons. It is very challenging to make a prediction system for obesity using big data. The paper proposed the system for the Spark and Hadoop-based predications for obesity of different people. This paper indicates the shortcomings of the traditional tools and methods to manage and analyze big data due to its size, heterogeneity, and speed of generation for obesity predication. It also recommends the Hadoop-based map-reduce for minimizing the size of big data. First, findings support the inclusion of the measurement of waist circumference as a standard of care in both primary care and occupational health to ensure that central obesity is addressed as a risk factor for both declining health and declining obesity rate. Second, findings from this study support use of the Big Data for assessment of work ability in the world.

TABLE II. COMPARISON OF DIFFERENT EXISTING SYSTEMS FOR PREDICTING THE BMI BASED OBESITY.

Flink Predictor [12]	Three categories of obesity prediction based on BMI.
K-Means [12]	Five categories of obesity prediction based on BMI.
Multi-MapReduce [13]	Four categories of obesity prediction based on BMI.
Apache Spark and Hadoop for Obesity Prediction	Six categories of obesity prediction based on BMI.

The obesity predictor's design allows it to be used in both research and clinical practice. The world's culture is being reshaped. It is necessary to concentrate on the evaluation, promotion, and enhancement of work ability. Annual evaluations of adult obesity disease by primary care and/or occupational health providers are required to identify factors influencing declining work ability. This move would allow for intervention before obesity disease progresses to the point

of disability, allowing persons with obesity difficulties to exit the system sooner.

VII. Conclusion

Big Data is now a multi-disciplinary field for storing and analyzing enormous amounts of data to generate new data. We used an open-source repository to do statistical analysis on large-scale streaming data for obesity prediction. We have used the MapReduce technique for the classification of obesity. And feed it to the Apache spark and Apache Hadoop frameworks for the assessment. It was very challenging to make a prediction system for obesity using MapReduce. The analysis has been performed utilizing the Apache sparkle and Hadoop frameworks and assessing the prediction into six different BMI based categories i.e., underweight, normal weight, overweight, obese, very obese, and extremely obese. They are basically intended for the MapReduce technique, and appropriate for the continuous spark and Hadoop frameworks. There are some different systems like Apache Mahout, however our MapReduce technique with the help of Apache spark and Hadoop performed re-markable into categorizing the obesity prediction. Moreover, obesity prediction on group information utilizing MapReduce has been introduced firstly in this research.

References

[1] L. Stoner, D. Rowlands, A. Morrison, D. Credeur, M. Hamlin, K. Gaffney, et al., "Efficacy of exercise intervention for weight loss in overweight and obese adolescents: meta-analysis and implications," Sports Medicine, vol. 46, pp. 1737-1751, 2016

[2] T. Brown, T. H. Moore, L. Hooper, Y. Gao, A. Zayegh, S. Ijaz, et al., "Interventions for preventing obesity in children," Cochrane Database of Systematic Reviews, 2019.

[3] R. A. Rigby and D. M. Stasinopoulos, "Smooth centile curves for skew and kurtotic data modelled using the Box–Cox power exponential distribution," Statistics in medicine, vol. 23, pp. 3053-3076, 2004.

[4] D. Conrad, "The Stanford sports to prevent obesity randomized trial (SPORT)," in Sports-Based Health Interventions, ed: Springer, 2016, pp. 261-267.

[5] J. E. Lee, Z. Pope, and Z. Gao, "The role of youth sports in promoting children's physical activity and preventing pediatric obesity: a systematic review," Behavioral Medicine, vol. 44, pp. 62-76, 2018.

[6] S. A. Vella and D. P. Cliff, "Organised sports participation and adiposity among a cohort of adolescents over a two year period," PloS one, vol. 13, p. e0206500, 2018.

[7] N. Chandra, M. Papadakis, and S. Sharma, "Preparticipation screening of young competi-tive athletes for cardiovascular disorders," The Physician and sportsmedicine, vol. 38, pp. 54-63, 2010.

[8] E. Cacciari, S. Milani, A. Balsamo, E. Spada, G. Bona, L. Cavallo, et al., "Italian cross-sectional growth charts for height, weight and BMI (2 to 20 yr)," Journal of endocrinolog-ical investigation, vol. 29, pp. 581-593, 2006.

[9] A. Biffi, P. Delise, P. Zeppilli, F. Giada, A. Pelliccia, M. Penco, et al., "Italian cardiologi-cal guidelines for sports eligibility in athletes with heart disease: Part 1," Journal of Cardi-ovascular Medicine, vol. 14, pp. 477-499, 2013.

[10] T. J. Cole and T. Lobstein, "Extended international (IOTF) body mass index cut - offs for thinness, overweight and obesity," Pediatric obesity, vol. 7, pp. 284-294, 2012.

[11] K. M. Flegal and T. J. Cole, Construction of LMS parameters for the Centers for Disease Control and Prevention 2000 growth charts: US Department of Health and Human Ser-vices, Centers for Disease Control and …, 2013.

[12] G. Vicente-Rodriguez, P. J. Benito, J. A. Casajus, I. Ara, S. Aznar, M. J. Castillo, et al., "Physical activity, exercise and sport practice to fight against youth and childhood obesi-ty," Nutricion hospitalaria, vol. 33, pp. 1-21, 2016.

[13] S. A. Vella, D. P. Cliff, A. D. Okely, M. L. Scully, and B. C. Morley, "Associations be-tween sports participation, adiposity and obesity-related health behaviors in Australian ad-olescents," International Journal of Behavioral Nutrition and Physical Activity, vol. 10, p. 113, 2013.