Kuantumjournals
THE INTERNATIONAL RESEARCH COMMUNITY

Kuantum Journals Of Clinical Studies And Medical Case Reports

Article Information

Received: August 23, 2024
Accepted: September 20, 2024
Published: September 22, 2024

Citation: Diez-Lopez I, Maeso-Mendez S, Sanchez-Merino S. (2024) A new paradigm in the construction of growth charts in pediatrics. Why not use of big data?. Ku J Clin Stud Med Case Rep. 1(2): 019-032.

Research Article

A new paradigm in the construction of growth charts in pediatrics. Why not use of big data?

Ignacio Díez López1*, Sandra Maeso Mendez2, Gaspar Sánchez Merino3

1.Basque Country University UPV-EHU Pediatric Department, Vitoria, BioAraba, Child Growth and Metabolism Group, Spain
2. Child and Adolescent Endocrinology Unit, Department of Pediatrics OSI Araba, Osakidetza Vitória, BioAraba, Child Growth and Metabolism Group, Spain
3. Coordinator of the Innovation Platform, IIS BIOARABA, Vitoria, Spain

***Corresponding Author:** Ignacio Diez Lopez, Basque Country University UPV-EHU Pediatric Department, Vitoria, BioAraba, Child Growth and Metabolism Group, Spain, Email: ignacio.diezlopez@osakidetza.eus

1.Abstract

To date, knowledge of population dynamics and their impact on health has required complex, time-consuming and costly field studies. Big data tools are now a days postulated as a first-rate tool for weighting population changes observed in real time if reliable collection sources and adequate mathematical and computational tools are available for their assessment.

1.1.Main Objective: To carry out a methodological approach to the use of big data applications to develop auxological growth charts in our population with high statistical power. To assess how our population is in terms of auxological variables with respect to the current standards in the country: Orbegozo 2011 and Spanish growth studies 2010.

1.2.Material and Methods: Data collected from episodes of computerised medical records, studying the variables sex, age, weight, height, place of residence (PC, health centre, neighbourhood) of our population between 01/01/2020-31/03/2020 (avoiding pandemic effect). To calculate the curves and percentile tables, we used the Cole-Green LMS algorithm with penalised likelihood, implemented in the RefCurv 0.4.2 software (2020), which can handle large amounts of data. The hyperparameters were selected using the BIC (Bayesian information criterion). To calculate population deviations with respect to the reference population, we have taken as a reference being above 1.5 standard deviations with respect to the mean according to age.

1.3.Results: A total of 66,975 computerised episodes are collected from children under 16 years of age and a total of 1,205,000 variables studied. Although data are available, individuals 16th are excluded due to low N. The graphs of our population with respect to the standards are represented, noting that there are differences with Orbegozo 2011 and Spanish 2010. We present the data and percentages of overweight/obesity by age and sex. There are significant differences of more overweight in the entire sample of men and women in our population than the usual standards.

1.4.Conclusions: Bigdata technology is more powerful than classic population studies and is an innovative tool compared to auxological studies (limited in N) carried out to date. The development of these new strategies in auxology will allow us to know almost in real time the epidemiological situation of the population in different variables, being able to infer health actions in a more effective way.

2.Keywords

Big data; Computerised medical records; Bayesian information criterion; Clinical practice; Medro Period

3.Introduction

The study of human growth is defined as the process by which individuals increase in mass and height as they reach maturity, acquiring the functional characteristics of the adult state [1].

It is therefore considered a relevant indicator of health status in childhood. It is common clinical practice to weigh and measure children throughout the so-called medro period [2].

Growth is not only the expression of a genetic capacity, but is also the phenotypic expression of the state of physical, psychological and social well-being. In short, its health [3].

In order to reflect the situation of a child in relation to individuals of the same age and sex, various growth curves have been developed, both at international level (such as the classic Tanner curves, no longer in use), and others of a multinational nature, such as those produced by the WHO [4-5], which have been postulated as benchmarks of nutritional status and general health; as well as at national level (in our country, the one developed by Carrascosa et al - Spanish studies 2010) and regional level (in the Basque Country, those developed by the Orbegozo Foundation in 1988, 2004 and 2011), as closer representatives of the reality of our environment [6-7]. These studies, if they are of high quality, are longitudinal in nature, so their very nature means that they are long, costly and with a limited number of subjects.

Current electronic medical records include the collection of multiple objective data, clinical and medical-analytical variables and constants as part of routine clinical practice. Among them, aspects of children's somatometry. Different statistical techniques, such as machine learning, would make it possible to exploit these data from a large number of cases (those representing the majority of the population) in a semi-automated way and almost in real time. However, it is important that prior to the analysis of the data, a data cleaning process is carried out. This is due to the fact that the information collected in the medical records is for the purpose of health care, not for research purposes, so there may be variability or errors in the measurement and transcription of the data.

Another important element is the impact that COVID-19 confinement may have had on the physical health of children. Although there are international studies on the subject (Pediatrics 2020, Children 2021) [8-9], there are no studies, at least in our environment and nearby population, on how the pandemic may have influenced the weight/height ratio in children.

Existing studies to date on the somatometric status of the child population, which are considered as a reference for our children, are based on cross-sectional or longitudinal designs with a small sample size [4-7].

This work is seen as an opportunity to use the data already existing in the computerised network to gain first-hand knowledge of the somatometric situation of our environment in minors.

4.Objectives

4.1.Main Objective

To describe the health status of the paediatric population in our setting, Alava, Basque Country, Spain, by extracting variables from the electronic health record and analysing them using a new big data approach.

To check whether these results obtained, current and from a large population (our study), differ significantly (expressed in SDS) for the different variables collected with respect to the normality set by current studies of child auxology (Orbegozo 2011) (old and limited in size) by means of a comparison of paired means.

5.Materials and Methods

5.1.Design

This is a population-based cross-sectional study.

5.2.Study population

All children under 18 years of age under follow-up in the Basque health system, OSAKIDETZA, who present weight and height records in the OSABIDE GLOBAL tool in the area of Alava.

5.3.Inclusion criteria

- Both sexes
- Ages 0-18 years
- To be registered or to present an address of registration (according to data of the GLOBAL of the area of Alava/Araba) belonging to the OSI Araba (which includes all the rural zone of the province except the tributary area of LLODIO) belonging to one of the centres of health of the own OSI, collecting of which one it is in each case.
- To have this data collected in OSABIDE GLOBAL

5.4.Exclusion criteria

- No data recorded in GLOBAL

5.5.Sample size calculation

The study will include all persons between 0 and 18 years of age residing in the historical territory of Araba (except as described). According to data from the Basque Statistics Institute (Eustat), in the year 2021 there were 47,853 persons aged 0-19 in Vitoria (Basque Statistics Institute (Eustat). Population of the Autonomous Community of the Basque Country by territorial area, large age groups, sex and period. Available at: https://www.eustat.eus/bankupx/pxweb/es/DB/-/PX_010154_cepv1_ep06b.px/table/tableViewLayout1/. (Accessed 29/08/2022). As it includes the entire paediatric population of Araba de Vitoria, it is considered that it is not necessary to calculate the sample size. However, it is possible that after the data filtering process there may be a loss in the number of participating subjects, in those cases in which the data necessary to carry out the study have not been recorded or are not well recorded.

In order to eliminate the effect of confinement (COVID-19) experienced by the population, or at least minimise it, the data to be studied would be those collected in the database with two different dates This original article presents the data referring to all the records of the first quarter of the year 2022 existing in the database.

5.6.Variables

- Main variables:
    1. Weight (Kgrs)
    2. Size (cm)
    3. Sex (Male, Female, Binary)
    4. Age (expressed in years and months)
    5. Date of registration

5.7.Data management plan

A data protection impact assessment has been prepared. The data life cycle will involve the IT service of OSI Araba, the principal investigator of the project and the collaborating researchers, including professionals from the Basque Center for Applied Mathematics (BCAM) who are part of the research team (Figure 1). There is a collaboration agreement between BCAM and the Bioaraba Health Research Institute.

The principal investigator requests the extraction of data (date of birth in month and year format, sex, weight, height, date of registration and health centre) from the electronic health record (EHR) from the IT service of OSI Araba.

The BCAM has the data obtained from Osakidetza's electronic medical records for the time necessary to carry out the following actions:

- Data cleaning
- Statistical analysis of the data

Once the research has been completed, the database will be completely destroyed by all persons involved in this study.

Specific security measures were adopted to prevent re-identification and access by unauthorised third parties. The database obtained from the IT Service comes with a patient ID that is neither the CIC, nor the clinical history number, nor any other data that could be used for re-identification of patients. Only the IT Service will have knowledge of this ID.

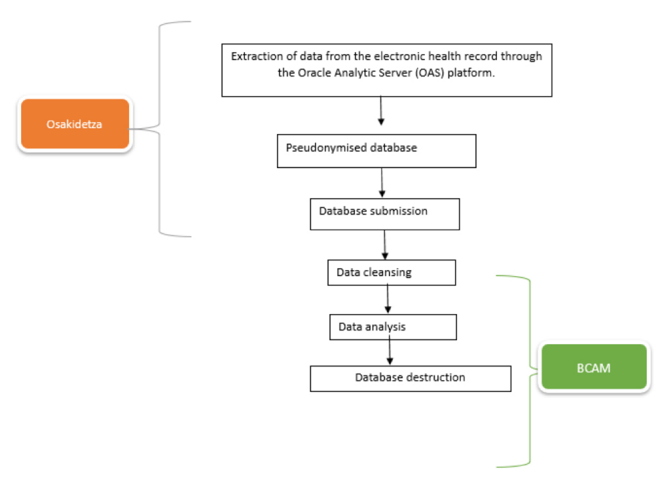

Osakidetza
Extraction of data from the electronic health record through the Oracle Analytic Server (OAS) platform.
Pseudonymised database
Database submission
Data cleansing
Data analysis
Database destruction
BCAM

Figure 1: Data Life Cycle

5.8.Statistical analysis

5.8.1.Hierarchical Dirichlet process mixture model

In the field of machine learning, Dirichlet processes (DPs) [10] are a family of stochastic processes whose realisations (the values they take) are probability distributions. DPs are used in Bayesian inference to describe prior knowledge about the distribution of random variables, i.e., the probability that random variables are distributed according to a particular distribution.

The Dirichlet process is specified by a base distribution and a positive real number, called the concentration parameter. The base distribution is the expected value of the process. A DP draws distributions around the base distribution. Although the base distribution is continuous, the distributions drawn from a DP are discrete. The concentration parameter controls the number of values of the PD: the realisations are discrete distributions with decreasing concentration as the concentration parameter increases.

The Dirichlet process can also be seen as the infinite-dimensional generalisation of the Dirichlet distribution. In the same way as the conjugate Dirichlet distribution for the

categorical distribution, the Dirichlet process is the conjugate for infinite discrete non-parametric distributions. A particularly important application of Dirichlet processes is as a prior probability distribution in infinite mixture models [11]. In this project we will adopt this approach, and the DPs will allow us to build Gaussian mixture models (GM) [11]. These models are known as Dirichlet process Gaussian mixture models (DPGMM).

DPGMMs are especially interesting for modelling populations where the number of clusters is unknown, because they are able to establish the number of components and their parameters (means and covariance matrices) of the Gaussian averaging model automatically. The number of components will be determined by the data and by the value of the concentration parameter. The DPGMM will allow solving two problems simultaneously: performing a probabilistic clustering of the study population and at the same time modelling the underlying distribution of the study population (density estimation) in terms of GMs.

The DPGMM will allow to analyse the study population, classify a new individual into one of the previously identified groups, and make predictions about the variables that characterise it. However, the aim of this study is to analyse different populations that may be segmented according to different criteria, e.g., location or age. For this, we will need to add a higher level of abstraction to the DPs.

A hierarchical Dirichlet process (HDP) is a Bayesian non-parametric approach to data clustering [11]. HDP employs a DP for each of the populations, subject to the constraint that all DPs share a base distribution. This base distribution is drawn from a DP higher in the hierarchy, hence the term hierarchical Dirichlet process. In terms of the DPGMMs, the DPs of each of the populations share the Gaussian components and the top DP in the hierarchy groups the DPGMMs of the populations and in turn establishes the strength of each Gaussian component. In this way, HDPs allow groups to share statistical strength by exchanging clusters between groups, and in turn perform clustering of the different populations. Again, in HDPs the number of GM components, the number of population clusters and the weights of the GM components are set automatically from the data.

5.9.Application to the study of populations

In this project, we will approach the analysis of a set of populations using Hierarchical Dirichlet process Gaussian mixture Gaussian averaging models (HDPGMM) [12]. This will allow us to address the following problems: 1) learn a GM model for the set of populations, 2) establish the different clusters of individuals with a differentiated behaviour within the total population, 3) establish clusters of populations with similar behaviour, 4) determine the strength of the GM components in each population, 5) classify a new individual in one of the GM components and infer the value of one of its variables from the value of the rest, and 6) compare the results obtained at two time instants and analyse the evolution of the populations.

Specifically, by grouping the data according to the different variables, clusterings will be obtained that will inform us about the somatometric similarities and differences of the population according to the date of data collection, age, sex, or health centre [13]. In this way, in addition to being able to draw conclusions about the general population (secondary objectives 1 and 3), it will also be possible to address secondary objectives 2 and 5. The study will also be an opportunity to study and incorporate recent methodological innovations on databases similar to ours [14-16].

Because HDPGMM learning solves the problems of clustering and density estimation, the results can be intuitively visualised using standard visualisation techniques such as linear projections to 2D spaces, e.g., Distributed Stochastic Neighbour Embedding (DSNE) and the Multidimensional Scaling (MDS).

Open source Python libraries are used throughout the process: Pandas for data management and preprocessing, Sklearn for the basic algorithms (DSNE, MDS) and Matplotlib for the visualisation of the results. As for the HDPGMM model, we propose our own open source implementation based on public domain libraries such as https://github.com/blei-lab/online-hdp.

For each variable studied, we proceeded to carry out studies of MEANS and SDS.

These data are also compared with the means and SDS of the studies published to date and reference studies of our population (Orbegozo 2004, 2011 and Spanish 2011).

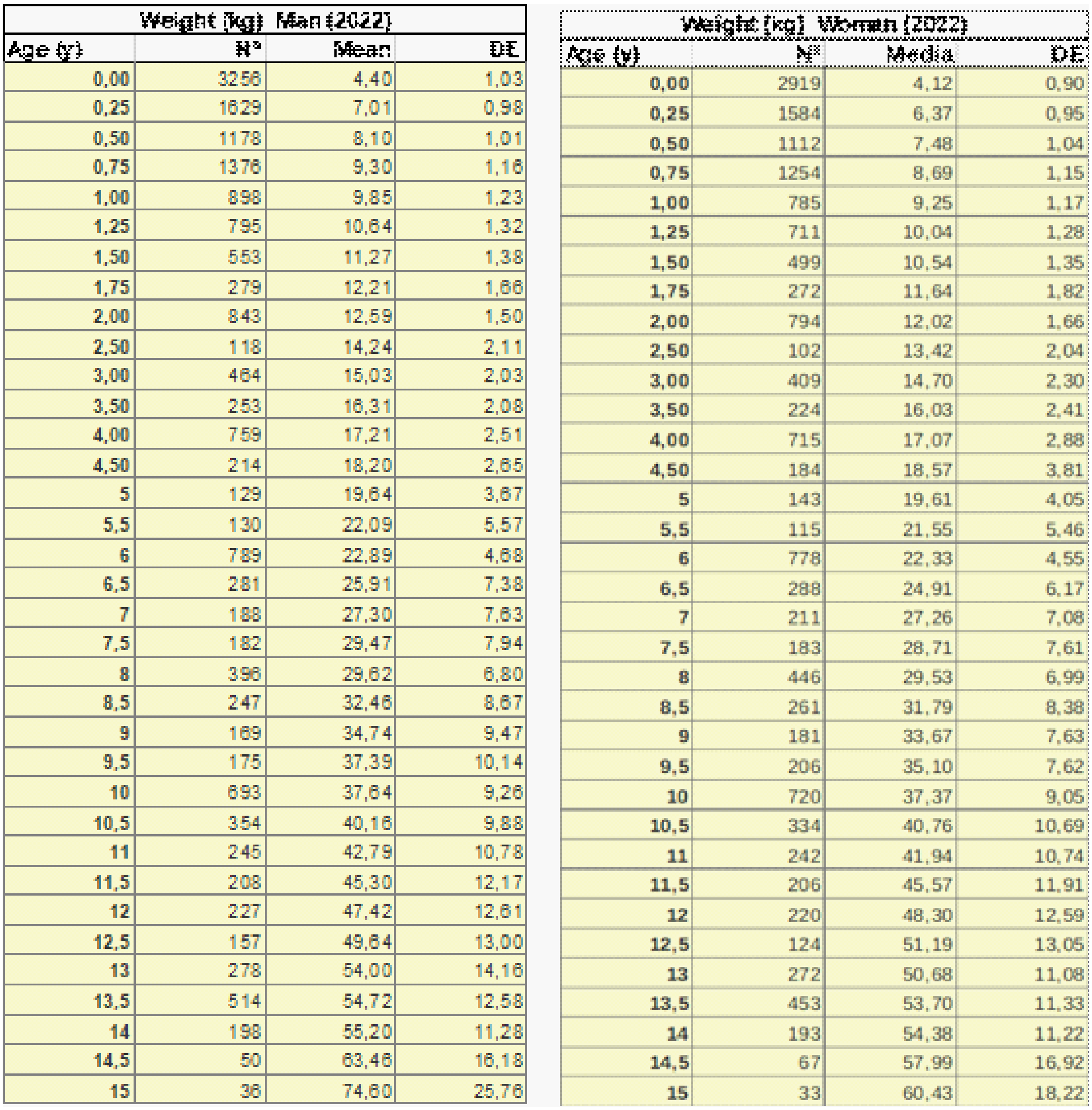
6.Results

Data were collected for a total of 67,270 minors.

The sum of all variables studied (24 per case) given the number of cases is 1,749,020 variables.

Although data are available for the age range 16-18 years, the number of available data being scarcer and with the dispersion of data presented, it was advised by the Collaborative Team of the study, in order to avoid bias, to be eliminated from this presentation.

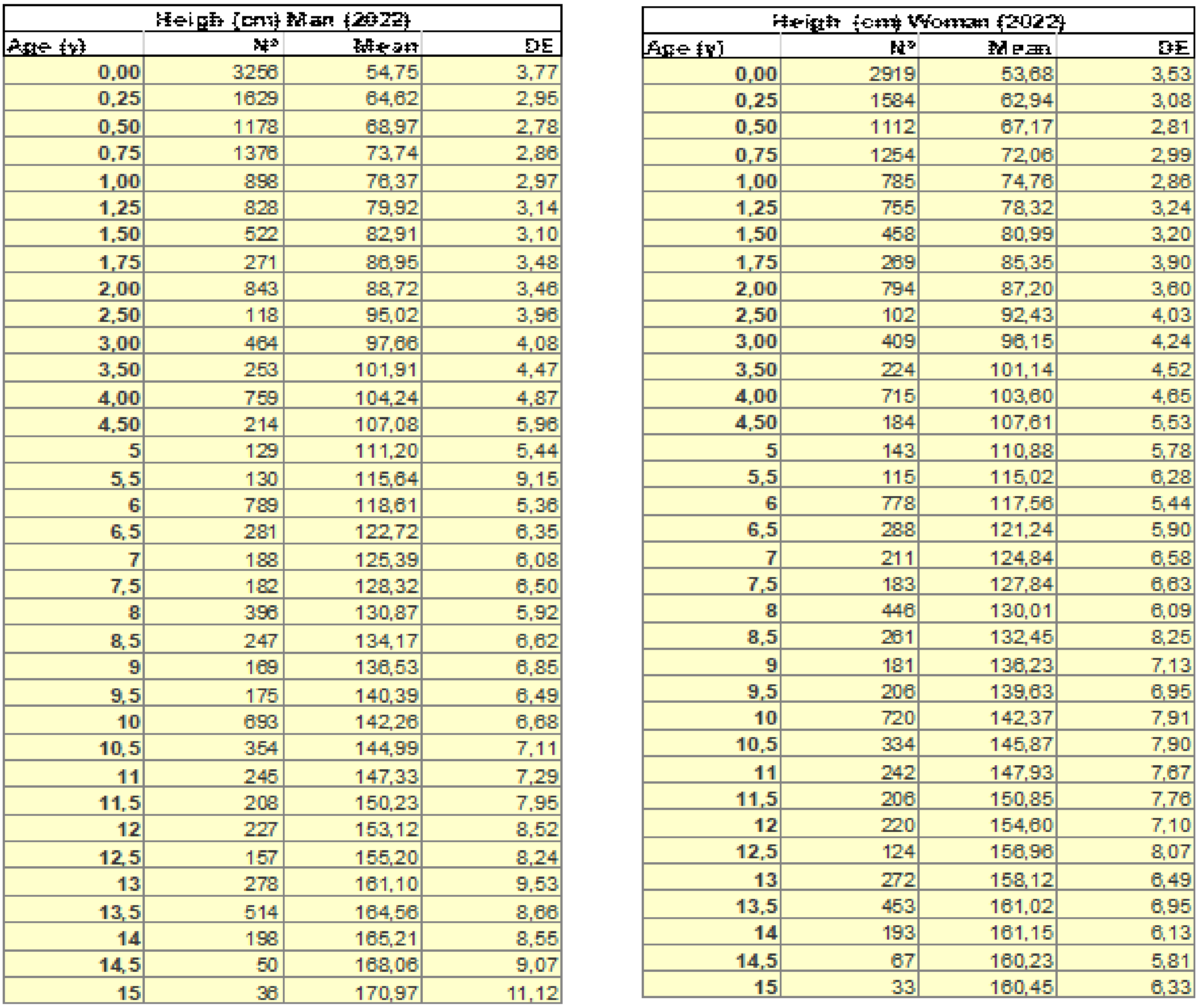
We present the results obtained by sex, age and the variables WEIGHT, HEIGHT and BMI in different tables (Tables 1,2,3).



Weight (kg) Man (2022)			
Age (y)	Nº	Mean	DE
0,00	3256	4,40	1,03
0,25	1629	7,01	0,98
0,50	1178	8,10	1,01
0,75	1376	9,30	1,16
1,00	898	9,85	1,23
1,25	795	10,64	1,32
1,50	553	11,27	1,38
1,75	279	12,21	1,66
2,00	843	12,59	1,50
2,50	118	14,24	2,11
3,00	464	15,03	2,03
3,50	253	16,31	2,08
4,00	759	17,21	2,51
4,50	214	18,20	2,65
5	129	19,64	3,67
5,5	130	22,09	5,57
6	789	22,89	4,68
6,5	281	25,91	7,38
7	188	27,30	7,63
7,5	182	29,47	7,94
8	396	29,62	6,80
8,5	247	32,46	8,67
9	169	34,74	9,47
9,5	175	37,39	10,14
10	693	37,64	9,26
10,5	354	40,16	9,88
11	245	42,79	10,78
11,5	208	45,30	12,17
12	227	47,42	12,61
12,5	157	49,64	13,00
13	278	54,00	14,16
13,5	514	54,72	12,58
14	198	55,20	11,28
14,5	50	63,46	16,18
15	36	74,60	25,76

Weight (kg) Woman (2022)			
Age (y)	Nº	Media	DE
0,00	2919	4,12	0,90
0,25	1584	6,37	0,95
0,50	1112	7,48	1,04
0,75	1254	8,69	1,15
1,00	785	9,25	1,17
1,25	711	10,04	1,28
1,50	499	10,54	1,35
1,75	272	11,64	1,82
2,00	794	12,02	1,66
2,50	102	13,42	2,04
3,00	409	14,70	2,30
3,50	224	16,03	2,41
4,00	715	17,07	2,88
4,50	184	18,57	3,81
5	143	19,61	4,05
5,5	115	21,55	5,46
6	778	22,33	4,55
6,5	288	24,91	6,17
7	211	27,26	7,08
7,5	183	28,71	7,61
8	446	29,53	6,99
8,5	261	31,79	8,38
9	181	33,67	7,63
9,5	206	35,10	7,62
10	720	37,37	9,05
10,5	334	40,76	10,69
11	242	41,94	10,74
11,5	206	45,57	11,91
12	220	48,30	12,59
12,5	124	51,19	13,05
13	272	50,68	11,08
13,5	453	53,70	11,33
14	193	54,38	11,22
14,5	67	57,99	16,92
15	33	60,43	18,22

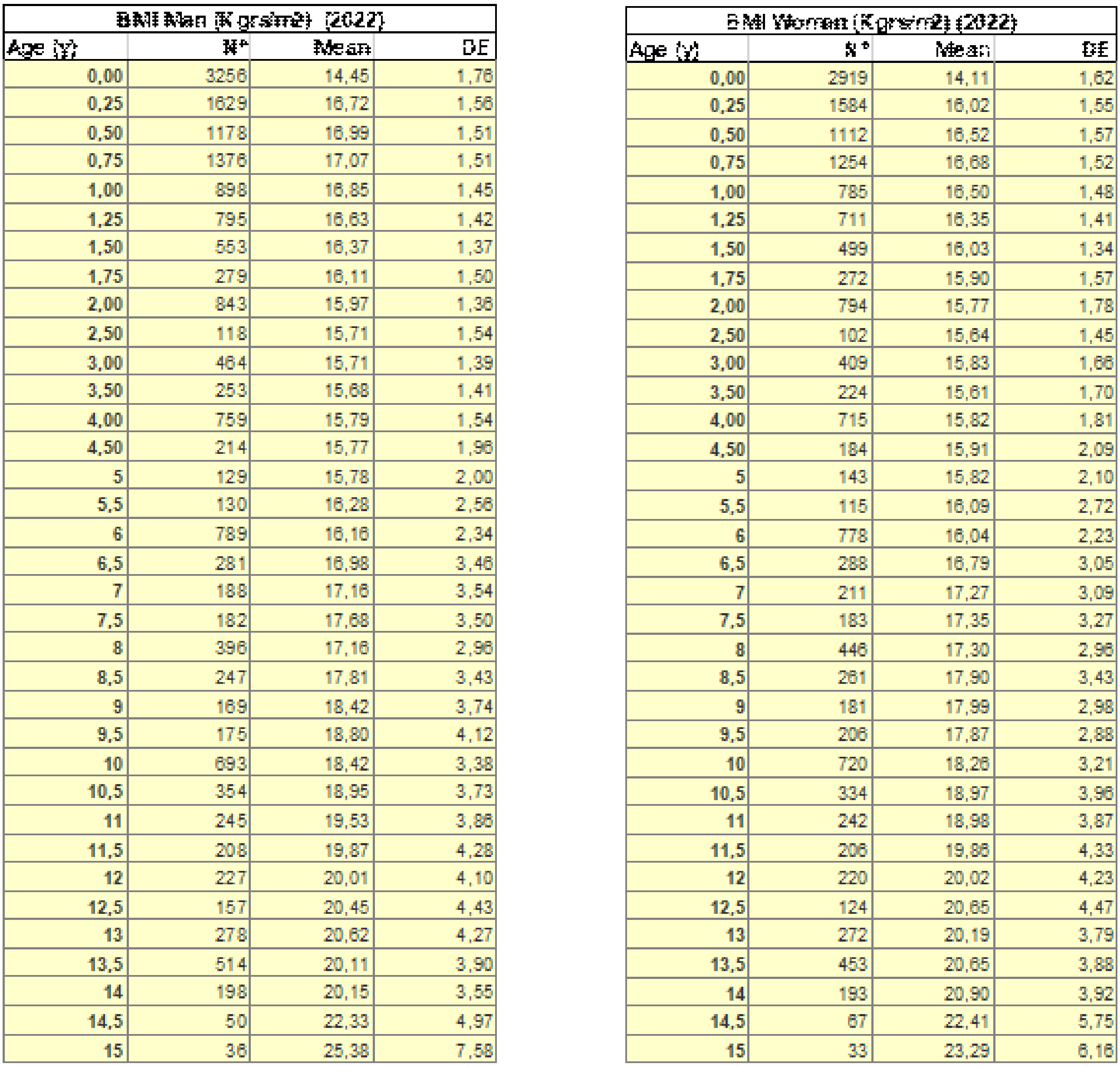
Table 1: Numerical representation of data by age for the variable WEIGHT (Kgrs). Means and SDS



Heigh (cm) Man (2022)			
Age (y)	Nº	Mean	DE
0,00	3256	54,75	3,77
0,25	1629	64,62	2,95
0,50	1178	68,97	2,78
0,75	1376	73,74	2,86
1,00	898	76,37	2,97
1,25	828	79,92	3,14
1,50	522	82,91	3,10
1,75	271	86,95	3,48
2,00	843	88,72	3,46
2,50	118	95,02	3,96
3,00	464	97,66	4,08
3,50	253	101,91	4,47
4,00	759	104,24	4,87
4,50	214	107,08	5,96
5	129	111,20	5,44
5,5	130	115,64	9,15
6	789	118,61	5,36
6,5	281	122,72	6,35
7	188	125,39	6,08
7,5	182	128,32	6,50
8	396	130,87	5,92
8,5	247	134,17	6,62
9	169	136,53	6,85
9,5	175	140,39	6,49
10	693	142,26	6,68
10,5	354	144,99	7,11
11	245	147,33	7,29
11,5	208	150,23	7,95
12	227	153,12	8,52
12,5	157	155,20	8,24
13	278	161,10	9,53
13,5	514	164,56	8,66
14	198	165,21	8,55
14,5	50	168,06	9,07
15	36	170,97	11,12

Heigh (cm) Woman (2022)			
Age (y)	Nº	Mean	DE
0,00	2919	53,68	3,53
0,25	1584	62,94	3,08
0,50	1112	67,17	2,81
0,75	1254	72,06	2,99
1,00	785	74,76	2,86
1,25	755	78,32	3,24
1,50	458	80,99	3,20
1,75	269	85,35	3,90
2,00	794	87,20	3,60
2,50	102	92,43	4,03
3,00	409	96,15	4,24
3,50	224	101,14	4,52
4,00	715	103,60	4,65
4,50	184	107,61	5,53
5	143	110,88	5,78
5,5	115	115,02	6,28
6	778	117,56	5,44
6,5	288	121,24	5,90
7	211	124,84	6,58
7,5	183	127,84	6,63
8	446	130,01	6,09
8,5	261	132,45	8,25
9	181	136,23	7,13
9,5	206	139,63	6,95
10	720	142,37	7,91
10,5	334	145,87	7,90
11	242	147,93	7,67
11,5	206	150,85	7,76
12	220	154,60	7,10
12,5	124	156,96	8,07
13	272	158,12	6,49
13,5	453	161,02	6,95
14	193	161,15	6,13
14,5	67	160,23	5,81
15	33	160,45	6,33

Table 2: Numerical representation of data by age for the variable HEIGHT (cm.). Means and SDS.



BMI Man (Kgrs/m2) (2022)			
Age (y)	N°	Mean	DE
0,00	3256	14,45	1,76
0,25	1629	16,72	1,56
0,50	1178	16,99	1,51
0,75	1376	17,07	1,51
1,00	898	16,85	1,45
1,25	795	16,63	1,42
1,50	553	16,37	1,37
1,75	279	16,11	1,50
2,00	843	15,97	1,36
2,50	118	15,71	1,54
3,00	464	15,71	1,39
3,50	253	15,68	1,41
4,00	759	15,79	1,54
4,50	214	15,77	1,96
5	129	15,78	2,00
5,5	130	16,28	2,56
6	789	16,16	2,34
6,5	281	16,98	3,46
7	188	17,16	3,54
7,5	182	17,68	3,50
8	396	17,16	2,96
8,5	247	17,81	3,43
9	169	18,42	3,74
9,5	175	18,80	4,12
10	693	18,42	3,38
10,5	354	18,95	3,73
11	245	19,53	3,86
11,5	208	19,87	4,28
12	227	20,01	4,10
12,5	157	20,45	4,43
13	278	20,62	4,27
13,5	514	20,11	3,90
14	198	20,15	3,55
14,5	50	22,33	4,97
15	36	25,38	7,58

BMI Women (Kgrs/m2) (2022)			
Age (y)	N°	Mean	DE
0,00	2919	14,11	1,62
0,25	1584	16,02	1,55
0,50	1112	16,52	1,57
0,75	1254	16,68	1,52
1,00	785	16,50	1,48
1,25	711	16,35	1,41
1,50	499	16,03	1,34
1,75	272	15,90	1,57
2,00	794	15,77	1,78
2,50	102	15,64	1,45
3,00	409	15,83	1,66
3,50	224	15,61	1,70
4,00	715	15,82	1,81
4,50	184	15,91	2,09
5	143	15,82	2,10
5,5	115	16,09	2,72
6	778	16,04	2,23
6,5	288	16,79	3,05
7	211	17,27	3,09
7,5	183	17,35	3,27
8	446	17,30	2,96
8,5	261	17,90	3,43
9	181	17,99	2,98
9,5	206	17,87	2,88
10	720	18,26	3,21
10,5	334	18,97	3,96
11	242	18,98	3,87
11,5	206	19,86	4,33
12	220	20,02	4,23
12,5	124	20,65	4,47
13	272	20,19	3,79
13,5	453	20,65	3,88
14	193	20,90	3,92
14,5	67	22,41	5,75
15	33	23,29	6,16

Table 3: Numerical representation of data by age for the variable BODY MASS INDEX (Kgrs/m2). Means and SDS.

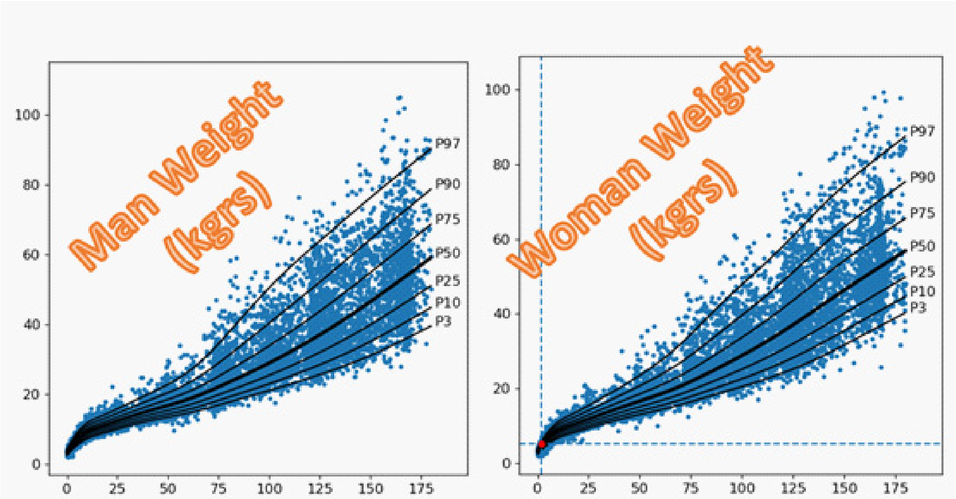


Figure 2: Percentile representation by age of the variable WEIGHT (Kgrs). Abscissae age in months.

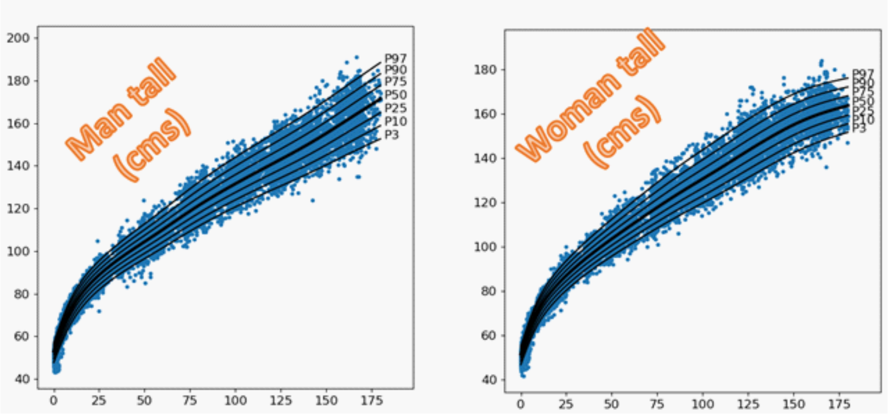


Figure 3: Percentile representation by age of the variable HEIGHT (cm). Abscissae age in months.

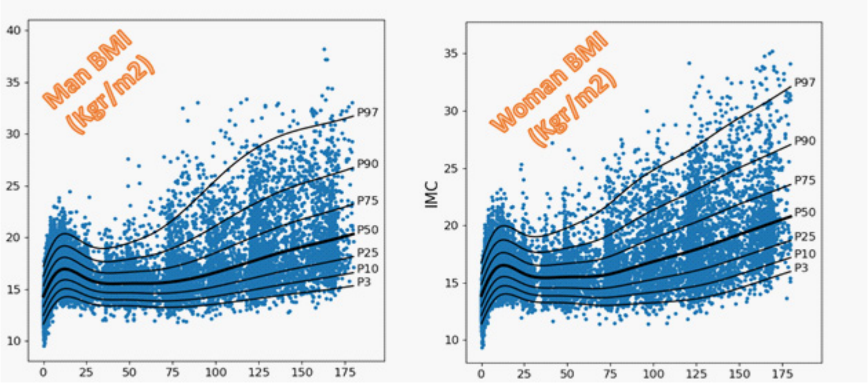


Figure 4: Percentile representation by age of the variable BMI (kg/m2). Abscissae age in months.

Next, we proceed to study using the so-called Hierarchical Dirichlet process Gaussian mixture model or method, applied to our population (big data 2022 study) vs. the most widely used reference graphs to date in the region (Orbegozo 2011) and the most recent and largest study in number of cases (Spanish study 2010) and used in the country (Figures 5,6,7,8,9,10).

Differences are assessed at a significance of p<0.05.

We represent the differences between our study (in black) and the referenced population (in red) for each of the studies and variables by means of the mean +/- 2 SDS.

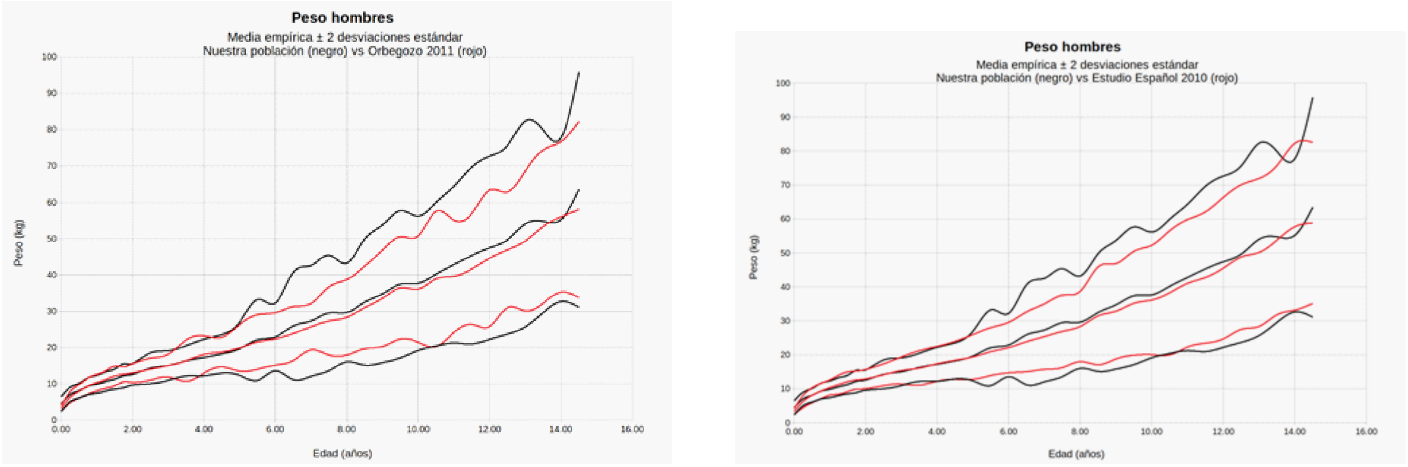


Figure 5: Representation of men AVERAGE +/- 2 SDS by age (years) of the variable WEIGHT (Kgrs). Reference study in red, our population in black.

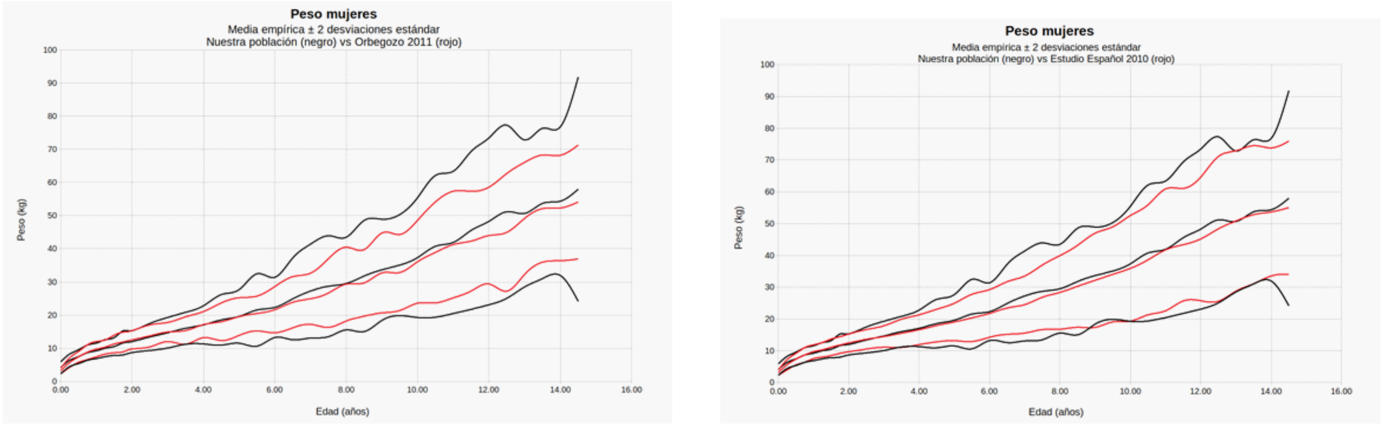


Figure 6: Representation of women AVERAGE +/- 2 SDS by age (years) of the variable WEIGHT (Kgrs). Reference study in red, our population in black.

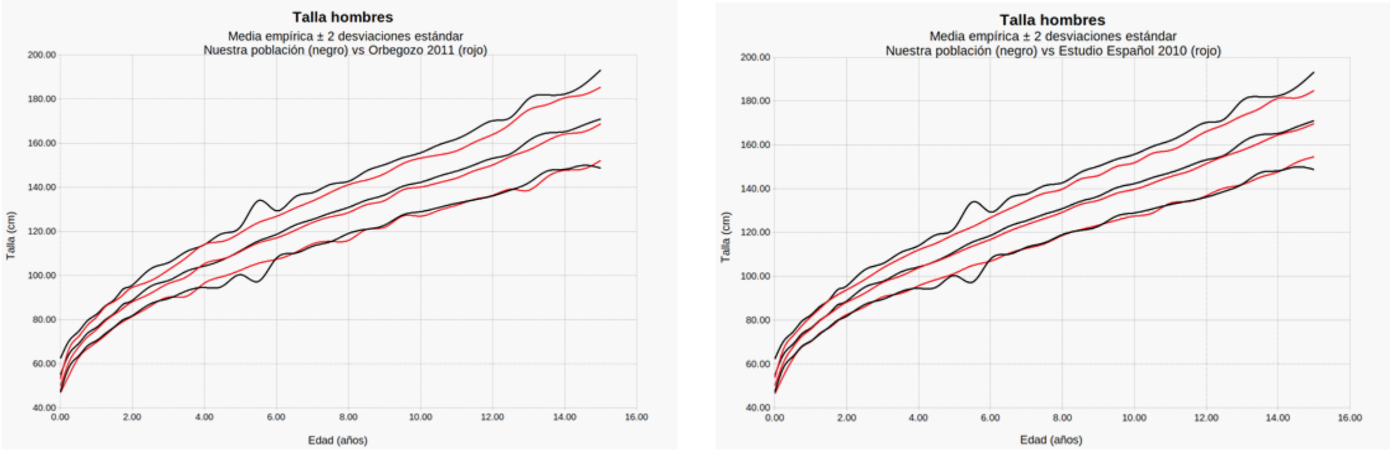


Figure 7: Representation of Men AVERAGE +/- 2 SDS by age (years) of the variable HEIGHT (cm). Reference study in red, our population in black.

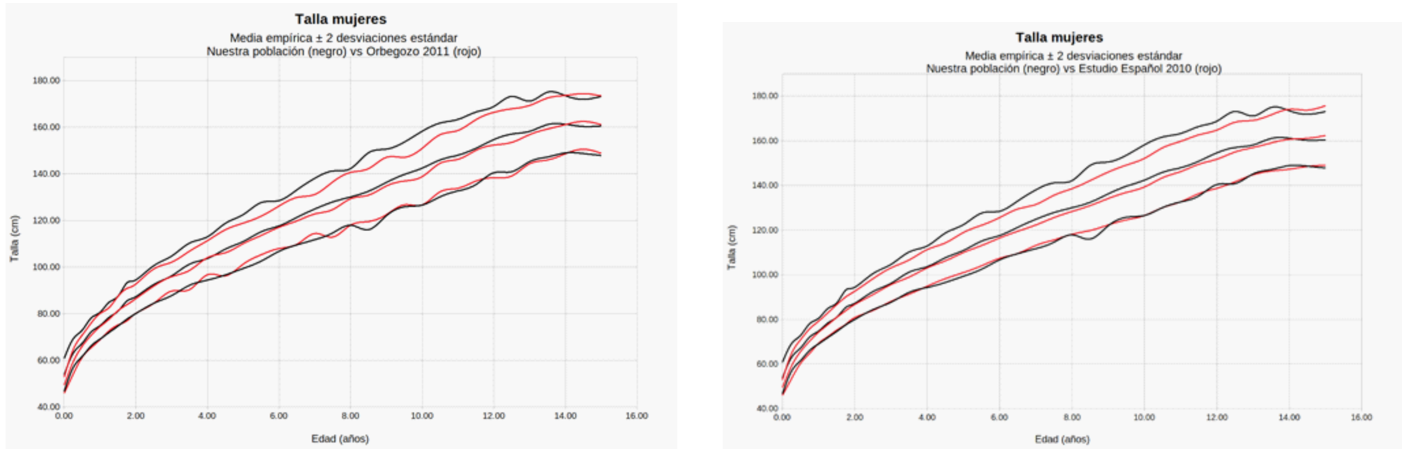


Figure 8: Representation of Women AVERAGE +/- 2 SDS by age (years) of the variable HEIGHT (cms). Reference study in red, our population in black.

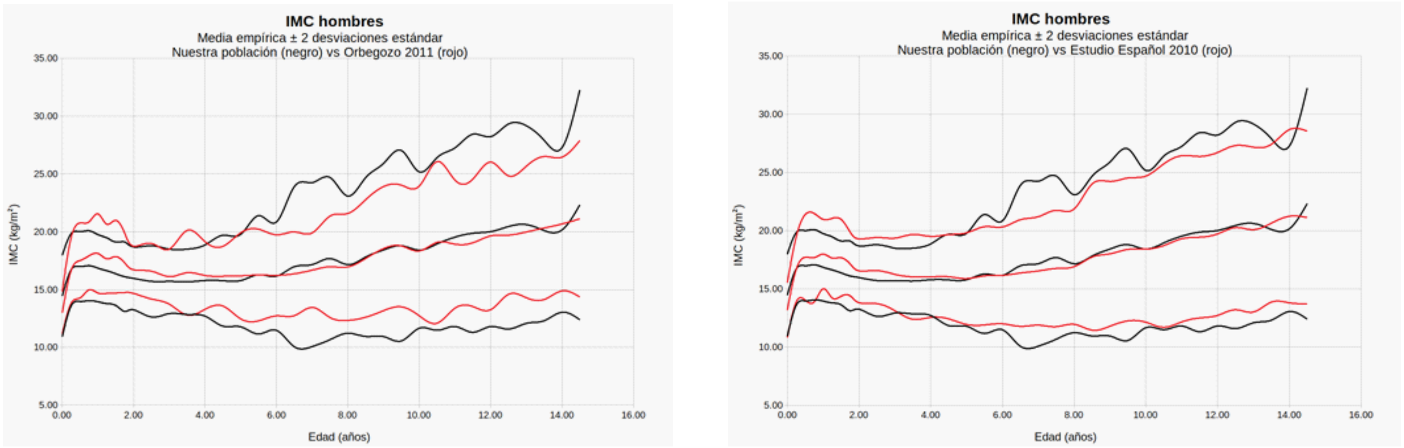


Figure 9: Representation of Men AVERAGE +/- 2 SDS by age (years) of the variable BMI (Kgr/m2). Reference study in red, our population in black.

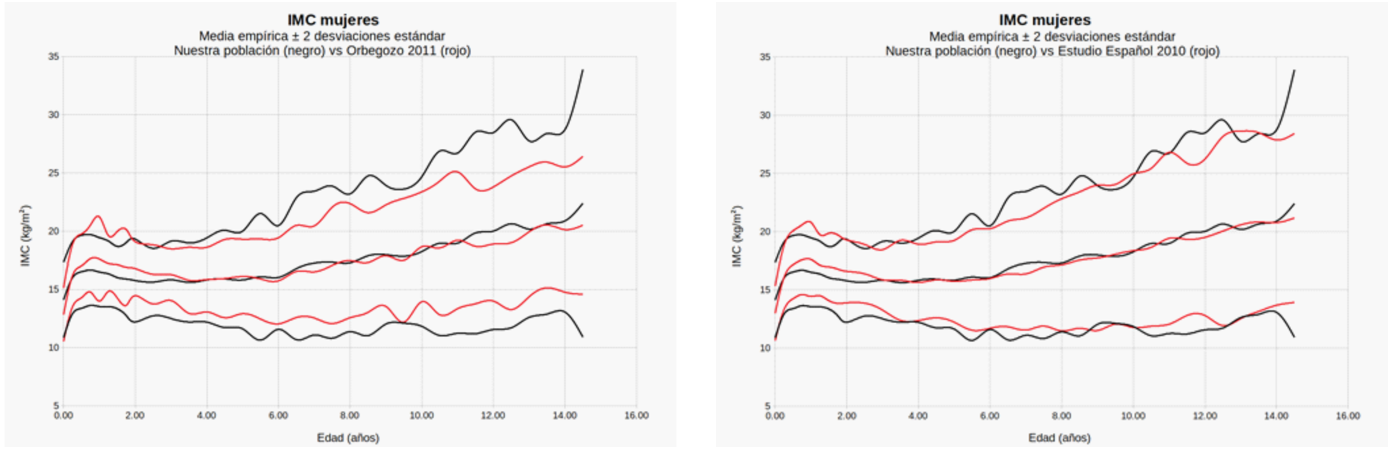


Figure 10: Representation of Women AVERAGE +/- 2 SDS by age (years) of the variable BMI (Kgr/m2). Reference study in red, our population in black.

The curves have not been statistically smoothed to reflect the reality of the observed sample. A significant difference ($p<0.05$) was found in each of the variables for weight and height and for all ages in comparison with our study vs. reference studies. The population studied has more height and weight in general than the reference population using the Hierarchical Dirichlet process Gaussian mixture model (Table 4).

Hierarchical Dirichlet process Gaussian mixture model. Differences found.

Table 4: Numerical representation of data by age of the variable BODY MASS INDEX (Kgrs/m2).

Difference (Vitoria-Estudio Español 2010) HDPGMM	P-value (t-test)
1,28	0,00000
0,03	0,76313
-0,72	0,00000
-0,61	0,00001
-1,14	0,00000
-1,01	0,00000
-1,30	0,00000
-1,04	0,00000
-0,58	0,00000
-0,86	0,00000
-0,53	0,00000
-0,35	0,00235
-0,24	0,00693
-0,27	0,07716
-0,10	0,61262
0,16	0,51493
0,00	0,98708
0,62	0,00807
0,62	0,02816
0,93	0,00117

As an example of the method used, the BMI data are presented here in the form of a table. The differences observed are smaller than for other variables because the differences in weight are compensated for by the greater height of the subjects analysed.

Even so, it is observed that the weight variable is higher than the height variable in our population than the referenced one, which contributes to the fact that the degree of overweight expressed in the form of BMI is also higher.

7.Discussion

The somatometric assessment of a child in relation to individuals of the same age and sex to date has used cross-sectional or semi-longitudinal studies, of a regional, national or international nature, but almost always with a limited number of cases due to the complexity of the technique and the cost of its development [4-5].

However, their importance is vital to have a tool for comparison and assessment of child health and they have been postulated as benchmarks of nutritional status and general health; such as at national level (in our country, the one developed by Carrascosa et al - Spanish studies 2010) and regional level (in the Basque Country, those developed by the Orbegozo Foundation in 1988, 2004 and 2011) [6-7]. These studies, if they are of high quality, are longitudinal in nature, so their very nature means that they are long, costly and with a limited number of subjects.

Current electronic medical records include the collection of multiple data and clinical constants as part of routine clinical practice. Among them, aspects of children's somatometry. Different statistical techniques, such as machine learning, have proven to be effective in other fields [14-16] for the interpretation of a large amount of data generated in real life and to be able to make decisions in this regard.

We postulate with works such as ours the possibility of the real use of this technology to obtain updated and almost real-time growth charts of such a large number of individuals that the power of the studies is very significant.

In this paper we present this possibility as a methodological approach.

Although data are available for the 16-18 age range, the smaller number of cases in relation to the other age groups and the dispersion of these cases means that for the statistical procedure of the study and to avoid bias, they have to be eliminated from this study. This is due to the fact that adolescents go to the doctor less often and therefore the number of records is lower.

In order to assess adolescent populations, we postulate the possibility of conducting ad hoc studies of this population, using databases of educational centres, sports.

The differences with Orbegozo and the Spanish State in the cases of Height, Weight and BMI are statistically significant with respect to our population in the year 2022.

The secular acceleration of weight and height [4-5] is confirmed in our population, since our population is on average 10 years later chronologically. 2010 vs. 2022.

Of all the variables, BMI is the one most affected by this comparative acceleration. This is highlighted in the study and may be due to various causes, such as the childhood obesity pandemic we are experiencing, the effect of confinement/ COVID-19 [8-9] in the year 2022 on child health, changes in diet or even the type of population in the area (immigration rate, socioeconomic level) [2-4].

This BIG DATA survey method is claimed to be a quicker and cheaper way to have up-to-date regional graphs than classical surveys. This should be checked with other studies.

We also add that the problem with the percentile curves is that they are obtained from a normalisation and adjustment method called LMS; Orbegozo and Spanish presents these results in tabular form [6-7]. We have also resorted to empirical graphs that are obtained directly from the means and standard deviations.

The authors encourage work in this direction and this work is the basis for the development of community intervention strategies to be corroborated by our own team.

7.Biases and Limitations of the Study

The main limitation of the study has to do with the fact that the data to be used come from the electronic medical record and therefore have not been generated for research purposes. Therefore, as described in the literature, errors may occur in the measurement and transcription of the data (Heude B et al A big-data approach to producing descriptive anthropometric references: a feasibility and validation study of paediatric growth charts. Lancet Digit Health. 2019 Dec;1(8):e413-e423). To minimise this limitation, the data extracted from the electronic medical records will be cleaned before proceeding with the statistical analysis of the data.

8.Project Impact

The expected impact of the project results, in terms of the ability to modify healthcare processes to improve the health and quality of life of patients, is of great importance.

It is estimated that the current cost of carrying out an updated, regional, longitudinal growth study, with the consequent limit of cases (<1,000) is more than 8-10 years per project and with an economic cost of more than 60,000 euros in this period, taking into account published studies (Orbegozo) in their methodology.

This project has developed in a much shorter time and at a lower cost.

Moreover, the data obtained are not limited to a limited (though supposedly representative) population but are quasi-real by encompassing most of the data available on the area's computer servers.

The nature of this study allows it to be repeated periodically, detecting areas for improvement in different subpopulations.

On the other hand, by having variables associated with other types of medro, such as health centres, it will be possible to detect areas of social and health risk, where other types of studies or intervention measures can be implemented.

9.Ethical Aspects

The study has been carried out in accordance with the principles established in the Declaration of Helsinki (1964), latest version Fortaleza, Brazil 2013, in the Council of Europe Convention on Human Rights and Biomedicine (1997), and in the regulations on biomedical research and personal data protection. Law 14/2007 on Biomedical Research.

Study approved by the CEIC on 24/03/2023 with CODE EXPTE 2022-058.

10.Economic Report

The study will be carried out without funding. The tasks described in the project are carried out by the principal investigator and his collaborators.

11.Acknowledgements

This original study has been support thanks to the work of the Collaborative Group from Basque Centre of Applied Mathematics (BCAM). Bilbao, Bizkaia Basque Country, Spain.

- Jose A. Lozano Basque Center for Applied Mathematics BCAM
- Ioar Casado Tellechea Basque Center for Applied Mathematics BCAM
- Aritz Pérez Postdoctor Fellow BCAM - Basque Center for Applied Mathematics

12.References

1. Zamlout A, Alwannous K, Kahila A, et al. (2022) Syrian national growth references for children and adolescents aged 2-20 years. BMC Pediatr. 22(1): 282.
2. Tarupi W, Lepage Y, Felix ML, et al. (2020) Growth references for weight, height, and body mass index for Ecuadorian children and adolescents aged 5-19 years. Arch Argent Pediatr. 118(2): 117-124.
3. Heude B, Scherdel P, Werner A, et al. (2019) A big-data approach to producing descriptive anthropometric references: a feasibility and validation study of paediatric growth charts. Lancet Digit Health. 1(8): e413-e423.
4. WHO Multicentre Growth Reference Study Group. (2006) WHO Child Growth Standards based on length/height, weight and age. Acta Paediatr Suppl. 450: 76-85.
5. de Onis M, Onyango AW, Borghi E, Siyam A, Nishida C, Siekmann. (2007) Development of a WHO growth reference for school-aged children and adolescents. Bull World Health Organ. 85(9): 660-667.
6. González ES, Lezcano AC, Garcia JMF, Longás AF, de Lara DL, López-Siguero JP. (2011) Spanish growth studies: current situation, usefulness and recommendations for use. An Pediatr (Barc). 74(3): 193.e1-16.
7. Lezcano AC, García JMF, Ramos CF, et al. (2008) Spanish cross-sectional growth study 2008. Part II: values of height, weight and body mass index from birth to adult height. An Pediatr (Barc). 68(6): 552-569.
8. Ashikkali L, Carroll W, Johnson C. (2020) The indirect impact of COVID-19 on child health Paediatrics and Child Health. Paediatr Child Health (Oxford). 30(12): 430-437.
9. Stavridou A, Kapsali E, Panagouli E, et al. (2021) Obesity in Children and Adolescents during COVID-19 Pandemic. Children (Basel). 8(2): 135.
10. Ferguson TS. (1973) A Bayesian analysis of some nonparametric problems. Ann statist. 1(2): 209-230.
11. Rasmussen CE. (1999) The infinite Gaussian mixture model. Advances in neural information processing systems. 12: 554-560.
12. Teh YW, Jordan MI. (2010) Hierarchical Bayesian nonparametric models with applications. Bayesian nonparametrics. 1: 158-207.
13. Van der Maaten L, Hinton, G. (2008) Visualizing data using t-SNE. J Machine Learning Res. 9(11): 2579-2605.

14. Kruskal JB. (1964) Non metric multidimensional scaling: a numerical method. Psychometrika. 29(2): 115-129.

15. Gilholm P, Mengersen K, Thompson H. (2020) Identifying latent subgroups of children with developmental delay using Bayesian sequential updating and Dirichlet process mixture modelling. PloS one. 15(6): e0233542.

16. Diana A, Matechou E, Griffin J, Johnston A. (2020) A hierarchical dependent Dirichlet process prior for modelling bird migration patterns in the UK. Ann Appl Statisti. 14(1): 473-493.