mathematics

MDPI

Article

Statistical Considerations for Analyzing Data Derived from Long Longitudinal Cohort Studies

Rocío Fernández-Iglesias [1,2,3,*], Pablo Martínez-Camblor [4,5], Adonina Tardón [1,2,3] and Ana Fernández-Somoano [1,2,3]

[1] Spanish Consortium for Research on Epidemiology and Public Health (CIBERESP), Monforte de Lemos Avenue 3-5, 28029 Madrid, Spain
[2] University Institute of Oncology of the Principality of Asturias (IUOPA)—Department of Medicine, University of Oviedo, Julian Clavería Street s/n, 33006 Oviedo, Asturias, Spain
[3] Instituto de Investigación Sanitaria del Principado de Asturias (ISPA), Roma Avenue s/n, 33001 Oviedo, Asturias, Spain
[4] Biomedical Data Science Department, Geisel School of Medicine at Dartmouth, Lebanon, NH 03756, USA
[5] Faculty of Health Sciences, Universidad Autonoma de Chile, Providencia 7500912, Chile
* Correspondence: rocio.fdez.iglesias@gmail.com

Abstract: Modern science is frequently based on the exploitation of large volumes of information storage in datasets and involving complex computational architectures. The statistical analyses of these datasets have to cope with specific challenges and frequently involve making informed but arbitrary decisions. Epidemiological papers have to be concise and focused on the underlying clinical or epidemiological results, not reporting the details behind relevant methodological decisions. In this work, we used an analysis of the cardiovascular-related measures tracked in 4–8-year-old children, using data from the INMA-Asturias cohort for illustrating how the decision-making process was performed and its potential impact on the obtained results. We focused on two particular aspects of the problem: how to deal with missing data and which regression model to use to evaluate tracking when there are no defined thresholds to categorize variables into risk groups. As a spoiler, we analyzed the impact on our results of using multiple imputation and the advantage of using quantile regression models in this context.

Keywords: missing data; quantile regression; tracking; cohort studies; children's health; cardiovascular risk

MSC: 62P10; 92B15; 92D30

Citation: Fernández-Iglesias , R.; Martínez-Camblor, P.; Tardón, A.; Fernández-Somoano, A. Statistical Considerations for Analyzing Data Derived from Long Longitudinal Cohort Studies. *Mathematics* **2023**, *11*, 4070. https://doi.org/10.3390/math11194070

Academic Editors: Miguel Ángel Montero-Alonso and Juan De Dios Luna del Castillo

Received: 7 August 2023
Revised: 18 September 2023
Accepted: 21 September 2023
Published: 25 September 2023

1. Introduction

Modern science is frequently based on the exploitation of large volumes of information stored in datasets and involving complex computational architectures [1]. Sometimes, these datasets compromise a huge number of participants. That is the case for those studies based on large registries, which frequently include hundreds of thousands or even millions of patients [2]. In this situation, despite some aspects of the statistical analyses becoming unuseful (i.e., p values), the main challenge is the computational capacity for handling the number of subjects. The so-called "omic sciences", including genomics, transcriptomics, proteomics, and metabolomics, among other technologies, represent a clear example of research requiring a high computational capacity but usually involving few subjects. In these studies, the researchers collected a number of variables and had to cope with several specific methodological challenges. Among those, we have examples such as preserving the security of the data, the difficulty of cleaning and checking their consistency, or the presence of missing values. The loss of subjects between follow-ups in the case of longitudinal studies, the data harmonization when the information comes from different records or systems, and apparently trivial aspects such as sometimes being able

to know the information contained in each variable are just a few examples of the issues that researchers have to deal with. Schmitt et al. [3] presented an interesting document in which the authors described the cohort, quality assurance procedures, and results of the Successful Aging after Elective Surgery (SAGES) study, highlighting the relevance of the processes related to data collection for having a successful project.

We consider here studies in which a relevant amount of information is systematically collected with the goal of studying the evolution of the enrolled participants in an undefined future, trying to delineate the associations between exposure to potential risk factors and posterior health status. Cohort designs, such as the landmark Framingham study [4], which was originally aimed to identify the determinants of cardiovascular disease (CVD) and whose collected data have been used with different goals; the European Prospective Intake and Cancer (EPIC) study [5], designed to investigate the relationships between nutritional, lifestyle, and environmental factors and the incidence of different types of cancer and other chronic diseases; or the Environmental Influences On Child Health Outcomes (ECHO) program [6], a network of pediatric cohorts that aims to understand the effects of a broad range of early environmental influences on child health and development, are just few examples. Particularly, there are a number of them that enrolled pregnant women and had active follow-ups with themselves and their children to determine whether pre-, peri-, or post-natal exposures may influence childhood or even adulthood health outcomes. Examples of these so-called birth cohorts are the Infancia y Medio Ambiente (INMA) (Environment and Childhood) project [7], a network concerned with the relation of environmental exposures with growth, health, and development from early fetal life until puberty, or the New Hampshire Birth Cohort (NHBC) study [8], which investigated the effect of several factors such as environment contaminants on the health outcomes of pregnant women and their children.

Usually, related subprojects involve part of the subjects and a limited number of variables. They suffer from the same problems. The use of multivariate statistical techniques implies that even if a subject is only missing one of the required variables, then it should be completely excluded from the analysis. Additionally, in longitudinal studies in which large numbers of variables are collected at different follow-ups, subjects having missing information at one follow-up can differ from those having missing information in another. This can result in a drastic reduction in the available sample size and, perhaps worse, the potential introduction of systematic biases. Aside from that, the study of risk factors in health populations, and particularly in children, copes with unclear or controversial threshold definitions. As a result, children thresholds are chosen as a specific percentile of the variable of interest [9], usually assuming that it is normally distributed with parameters estimated in healthy children; that is, there is not enough knowledge about the targets and clinically meaningful thresholds.

In this work, we aim to provide some statistical insight for longitudinal cohort studies involving controversial threshold definitions. Despite some of the considered techniques being new, we put the focus on their utilization in this particular setting. Dealing with missing data or selecting the adequate regression methodology implies making a number of decisions which could impact the final conclusions. Published documents are overwhelmingly focused on describing the obtained results and, in general, do not present in detail each decision made. Here, we pay more attention to those methodological details, analyzing the impact of the made decisions on the final results and discussing their suitability in relation to the possible alternatives.

2. Materials and Methods

2.1. The INMA-Asturias Cohort

In 2004, the INMA-Asturias cohort [10] was established as a prospective, population-based cohort study. As part of the INMA project [7], its aim is to examine the potential impact of environmental exposures on maternal and child health outcomes, with special emphasis on exposure to environmental pollutants and genetic and nutritional factors.

The cohort is located in a 483 km^2 area in northern Spain, with San Agustín University Hospital (Avilés, Asturias) serving as the reference hospital. The economy of this region historically relied on industries characterized by important environmental pollution. Originally, the area included a population of 165,201 inhabitants (reduced to 144,875 in 2021), and the reference hospital is a public health center with 436 beds, providing primary care as well as central, medical, and surgical services to this population.

From May 2004 to June 2007, pregnant women attending their first prenatal visits at the obstetrics service of San Agustín University Hospital or the Las Vegas health center (Corvera, Avilés) were consecutively selected if they met the following criteria: mother's age ≥ 16 years, singleton pregnancy, scheduled delivery at San Agustín University Hospital, no assisted conception, and no communication handicap. Extensive data were collected by trained staff through questionnaires, medical records, biological and environmental samples, and anthropometric measures. Follow-up visits took place at the first and third trimesters of pregnancy, at birth, and when the children's ages were 18 months and 4, 8, and 12 years.

The availability of blood samples enabled the measurement of markers of adult cardiovascular risk factors, including serum lipid, glucose, insulin, blood pressure, and anthropometric measures. These markers have expanded the scope of research beyond the initial objectives, allowing study of the tracking of cardiovascular-related measures. Here, we use the work by Fernández-Iglesias et al. [11] to illustrate the motivation behind specific methodological decisions and their potential impact on the results obtained.

2.2. Tracking of Cardiovascular-Related Measures

In epidemiology, predictability or maintenance of the range of a biological variable (or specifically of risk factors for chronic diseases) within a specific population is referred to as *tracking*. Particularly in children, early studies of growth established that some measures are relatively stable over time periods [12]. This phenomenon has interested both biologists and statisticians since the early 1980s, although there is no widely accepted definition of the term. Attempts to put the underlying concept into practice have resulted in the two main conceptions shown in Box 1.

Box 1. Tracking definitions.

- The ability to predict subsequent observations $(t+1)$ from earlier observations $(1, \ldots, t)$ [13]. If, in a cohort of n children, we measured their heights $y_{i,t}$, with $1 \leq i \leq n$ and $1 \leq t \leq k$, then *tracking* is the ability to predict $y_{i,t+1}$ from $y_{i,1}, \cdots, y_{i,t}$.
- The maintenance of a relative position within a distribution of values in the observed population through time [14,15]. Therefore, in the children's height example, the question is whether children at higher percentiles at time t will also be at higher percentiles at time $t+1$.

Here, we focus on this second conception in an attempt to explore the relationship between longitudinal measurements.

Considering that atherosclerosis is a progressive accumulation process that can begin in childhood and youth [16,17], in Fernández-Iglesias et al. [11], we studied the tracking between 4 and 8 years of the following cardiovascular-related variables that reflect well-established CVD risk factors in adulthood: waist-to-height ratio (WC/Height ratio) for central obesity, mean arterial pressure (MAP) for hypertension, triglycerides (TG), high-density lipoprotein cholesterol (HDL-c), and the atherogenic coefficient (AC) for dyslipidemia, and the homeostatic model assessment of insulin resistance (HOMA-IR) for insulin resistance.

Operationally, tracking is challenging [18], particularly when examining risk factors. The most commonly used statistical techniques in the literature include logistic regression, correlation coefficients, or linear regression models. Logistic regression models require the use of thresholds to categorize risk factors that are inherently continuous, typically

using specific quantiles. This is an extremely common approach in epidemiology research, but it has major limitations. It may lead to a loss of statistical power, to less precise estimates, or to difficulty in comparing results between studies when the thresholds are sample-dependent [19–21]. Choosing them arbitrarily can be a pitfall, especially when studying adult risk factors in generally healthy children. In such cases, it is advisable to use a methodology that allows for the use of continuous measures. However, commonly used continuous approaches, such as correlation coefficients or linear regression models [22–24], also have important limitations. These methods primarily concentrate on assessing the impact within the central part of the variable's distribution. However, in the context of variables denoting risk factors, a shift in the variable's mean often does not imply a meaningful clinical or health-related impact. Instead, it is the consequences observed at the extreme part of the distribution that hold a relevant significance. Consequently, the insights yielded by these techniques may not contribute substantial valuable knowledge. To overcome this challenge, in Section 2.4, we propose the use of quantile regression models to overcome two challenges: (1) to analyze the tracking of risk factors while avoiding the use of thresholds and (2) to maintain the focus on the extreme parts of the distribution.

2.3. Missing Data: Multivariate Imputation

Missing data is a recurrent problem in statistics which is especially impactful on longitudinal studies. Little and Rubin [25] proposed a missing data classification based on the underlying loss mechanism (Box 2).

Box 2. Types of missing data according to missingness mechanisms.

Let $\{\boldsymbol{X}, Y\}$ be a k-dimensional random matrix. For the sake of simplicity, we will assume univariate missing data; that is, Y is the only variable containing missing values. Let R be the response indicator vector; that is, $R = 1$ if Y is observed, and we have $R = 0$ otherwise. Then, the following apply:

- **The missing completely at random (MCAR) model satisfies**

$$\mathcal{P}\{R|(Y, \boldsymbol{X})\} = \mathcal{P}\{R\},$$

That is, the probability of being missing does not depend either on Y or $\boldsymbol{X}$. This means that there are no systematic differences between the missing and observed values. For example, serum lipid measurements may be missing because some samples have been lost in transit to the laboratory.

- **The missing at random (MAR) model satisfies**

$$\mathcal{P}\{R|(Y, \boldsymbol{X})\} = \mathcal{P}\{R|\boldsymbol{X}\},$$

That is, the probability of being missing depends on the observed data. For example, serum lipid measures may be more likely to be missing in young people, as they tend to be less concerned and do not attend visits for blood collection.

- **The missing not at random (MNAR) model satisfies**

$$\mathcal{P}\{R|(Y, \boldsymbol{X})\} = \mathcal{P}\{R|Y\},$$

That is, the probability of being missing depends on the missing values themselves or on unobserved information. For example, in a study to assess the effect of a hypertensive treatment, hypertensive subjects may present greater collaboration that results in a lower number of missingness.

The statistical analysis approach depends on each of these situations. Under the MCAR model, the observed data can be considered a random sample from the original target sample. In such cases, a complete-case analysis does not introduce bias in the estimated parameters but implies a sample size reduction with the associated loss of power. When missing data are not MCAR, as observed, the data do not represent the full population, and the complete-case approach may provide biased results. Multiple imputation (MI) methods

can produce unbiased estimations and preserve the original sample size under the MAR situation [26]. However, under the MNAR model, as long as the missingness depends on unobserved information, MI could fail [27]. Strategies to handle the MNAR model include collecting more information about the causes for the missingness or performing sensitivity analyses to evaluate the results under various scenarios [26].

The MI method, proposed in Rubin [28], does not focus on imputing the "closest" possible values to the actual missing values but rather making valid and efficient inferences about the parameters of interest. The key concept of MI is to use the distribution of the observed data to estimate a set of plausible values for the missing ones. Random components are incorporated into these estimated values to reflect their uncertainty. Multiple datasets are created and then analyzed individually. Finally, the individual estimations are combined to obtain the overall estimates, their standard errors, and adequate confidence intervals.

MI procedures consider the MAR model and the relationship

$$Y = g(\boldsymbol{X}) + \epsilon, \tag{1}$$

where $g(\cdot)$ and ϵ are the link function and random white noise, respectively. Box 3 summarizes the MI algorithm.

Box 3. Steps of the MI method.

Let $\{\boldsymbol{X}_n, Y_n\}$ be a random sample drawn from $\{\boldsymbol{X}, Y\}$, and let β be the target parameter. We assume that the values $y_{i_1}, \cdots, y_{i_m}$ $(1 \leq i \leq n,\ \ m < n)$ are missing.

- **Step 1**. From the non-missing values, we compute the function $\hat{g}(\cdot)$ which estimates $g(\cdot)$ (Equation (1)). For each missing value, y_{i_j} $(1 \leq i \leq n,\ \ 1 \leq j \leq m)$ generates a pseudo-value $\hat{y}_{i_j} = \hat{g}(\boldsymbol{X}_{i_j,n}) + \epsilon_{i_j}$, where ϵ_{i_j} is randomly generated. With this dataset, we estimate the target parameter $\hat{\beta}$ and its variance, $\hat{V}^2$.
- **Step 2**. We repeat Step 1 B times (where B is a large enough number) and obtain a vector of estimations $\{\hat{\beta}_1, \cdots, \hat{\beta}_B\}$ and another with their respective variabilities $\{\hat{V}_1^2, \cdots, \hat{V}_B^2\}$. Notice that in each repetition, the error (ϵ) is randomly generated. Therefore, each repetition provides a different dataset.
- **Step 3**. We use Rubin's rules to combine the vectors obtained in Step 2 into a single estimation with its variability. This estimation reflects both the uncertainty due to the sample variation and the uncertainty due to the missing data. The m $\hat{\beta}^k$ estimates and $\hat{SE}^k$ standard errors are combined using Rubin's rules to produce an overall estimate and standard error that reflect both the uncertainty due to the sample variation and the uncertainty due to the missing data.

Different algorithms have been proposed for estimating Equation (1) [29]. For instance, if we consider the linear model

$$Y = \boldsymbol{\beta} \cdot \boldsymbol{X} + \epsilon, \tag{2}$$

then we have the imputation process

$$\hat{y}_{i_j} = \hat{\boldsymbol{\beta}} \cdot X_{i_j} + \epsilon_{i_j}, \quad (1 \leq i \leq n,\ \ 1 \leq j \leq m)$$

where ϵ_{i_j} is randomly generated.

In many MI algorithms, a Bayesian perspective is often adopted, treating the parameters associated with the link function $g(\cdot)$ as random variables rather than fixed constants. This approach introduces uncertainty about missing values not only by incorporating random noise through the error term ϵ_{i_j}, as noted in Step 1 of Box 3, but also by introducing uncertainty into the link function parameters, whose state of knowledge is represented through a posterior distribution [26]. For instance, if we consider the same linear model (Equation (2)), then we have the imputation process

$$\hat{y}_{i_j} = \hat{\boldsymbol{\beta}} \cdot X_{i_j} + \epsilon_{i_j}, \quad (1 \leq i \leq n,\ \ 1 \leq j \leq m)$$

where ϵ_{i_j} is randomly generated and $\hat{\boldsymbol{\beta}}$ is sampled from its posterior distribution based on the available data.

Rubin's rules [28] combine the results of the B analysis performed to obtain

$$\bar{\hat{\beta}} = \frac{1}{B}\sum_{k=1}^{B}\hat{\beta}_k.$$

The variance of $\bar{\hat{\beta}}$, denoted as V_T^2, is calculated by

$$V_T^2 = V_W^2 + V_B^2 \cdot \left(1 + \frac{1}{B}\right), \tag{3}$$

where V_W^2 is the within-imputation variance and represents the sample variation and V_B^2 is the between-imputations variance and represents the extra variance due to the uncertainty around the imputed data; that is, we have

$$V_W^2 = \frac{1}{B}\sum_{k=1}^{B}\hat{V}_k^2, \quad \text{and}$$
$$V_B^2 = \frac{1}{B-1}\sum_{k=1}^{B}(\hat{\beta}_k - \bar{\hat{\beta}})^2$$

Inflating the between-imputation variance in Equation (3) by the factor $1/B$ reflects the extra variability as a consequence of imputing the missing data using a finite number of imputations instead of an infinite number. For constructing $100 \times (1-\alpha)\%$ confidence intervals, we assume $\bar{\hat{\beta}}$ is normally distributed and use the general formula

$$\bar{\hat{\beta}} \pm z_{\alpha/2}\sqrt{V_T^2},$$

where $z_{\alpha/2}$ is the critical value of the standard normal distribution.

Different indexes have been proposed for measuring the severity of the missing data problem. We consider here the so-called fraction of missing information (FMI), which estimates the proportion of the total variance due to the imputations and is defined by

$$\text{FMI} = \frac{V_B^2(1+1/B)}{V_T^2}.$$

The FMI ranges between 0 and 1. It is equal to zero only if the missing data do not add extra variation to the sample variance, an exceptional situation which implies perfect imputation models. And it is equal to one when the whole variation is caused by the missing data. In practice, this is equally unlikely since it means that there is no variation in the observed information [30]. The higher the value of this indicator, the greater the influence of the imputation model on the final results. Another index is the relative efficiency (RE), which represents the relative efficiency of using B rather than an infinite number of imputations:

$$\text{RE} = \frac{1}{1+\text{FMI}/B}. \tag{4}$$

It ranges from 0.5 to 1, where the higher the value, the less efficiency would be gained by increasing the number of imputed datasets.

The INMA-Asturias Cohort Example

In our study, we had a total of 416 children, but just 154 (37.02%) had all the required information. The missing percentage for cardiovascular-related variables oscillated between 6.97% and 44.47%. In the models, measures at age 4 play the role of the independent variable, and the same measures at age 8 play the role of the dependent variable. We excluded children who lacked data at the 4 and 8 year time points simultaneously. The final

considered sample (307 children) showed missing percentages which ranged from 2.3% to 25.1% (see Table S1).

The first decision related to missing data is the plausibility of the MAR assumption. We can reject data to be MCAR using the Little [31] test or by exploring if there are variables associated with missingness. But there is no way to distinguish whether the data are MAR or MNAR without additional information. In general, the MAR assumption will be more reasonable the more variables are included in the imputation model that are related to the missingness of the data on the variable of interest or to the variable of interest itself. In our case example, assuming the data were MAR, we had extra related auxiliary variables that could be incorporated into the imputation model, suggesting that the imputation methods could perform considerably well.

The second point is to specify the imputation model. To avoid unnecessary complexity, we have represented in this section the multiple imputation theory for univariate missing data. In our case, missing data occurred in more than one variable, and thus we applied a multiple imputation strategy for imputing multivariate missing data. In particular, we applied the *multivariate imputation by chained equations (MICE)* method [32] using the predictive mean matching algorithm. A detailed description and definition of this algorithm, which is based on Bayesian imputations in the MICE package, can be found in the work of Stef van Buuren [33]. Initially, we specified the imputation model with five (=B) imputed datasets [32,34–36]. Regarding to imputation model diagnosis, we assessed the maintenance of the observed relationship between the dependent and independent variables in the complete datasets. Figure 1, for example, shows that the distributions of observed and imputed data for TG were quite similar, as expected under the MAR approach.

The next decision was to determine the final B value. As the rate of missing information was below 0.5, we applied the criteria suggested by White et al. [36], Graham et al. [37], and Bodner [38]. We started with B equal to the maximum percentage of missing data observed ($B = 26$). Then, we applied the corresponding analysis to each generated dataset and combined the results. The FMI was calculated and verified whether $100 \cdot \text{FMI} \leq B$. B should be adjusted to the minimum number that satisfies this criterion otherwise. Of the 81 quantile regression models performed, the FMI median was 0.25 (interquartile range (IR): 0.12; 0.29), but the maximum was 0.46. As the computation time and storage capacity were not a concern, we finally selected $B = 50$.

After that, MI was repeated with the new number of imputations ($B = 50$), the models were estimated for each of the 50 datasets created, and the overall estimates and variances were calculated using Rubin's rules. The influence that the imputation had on these estimates was checked. Table 1 summarizes the corresponding indicators for each of the cardiovascular-related measures. The proportion of the total variance due to the imputation procedure was around 28% in the models involving measures with higher percentages of missing data and around 10% in those measures with low percentages. Note that by using a number of imputations B satisfying $100 \cdot \text{FMI} \leq B$ and taking into account Equation (4), it is expected to obtain REs higher than 99%, as we observed in Table 1. Therefore, minimal variation would occur just by increasing the number of imputations.

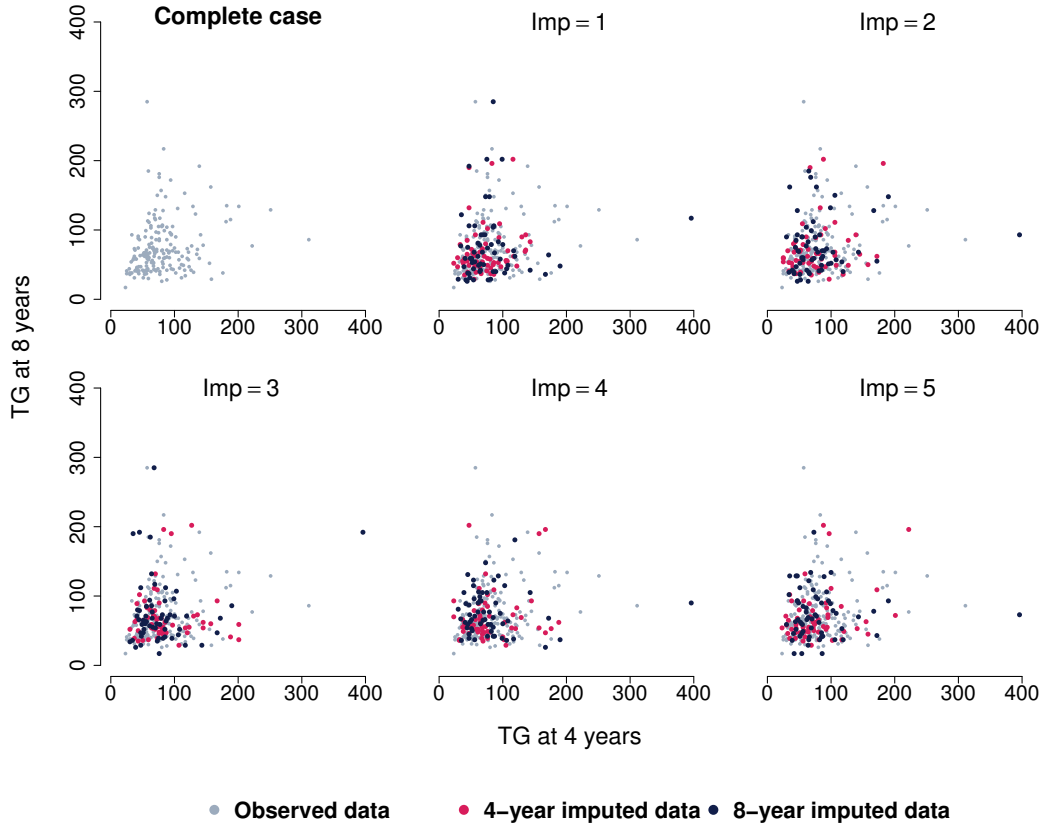

Figure 1. Scatter plots for each of the initial five imputed datasets of TG measure at 4 vs. TG measure at 8 years. TG = triglycerides; Imp = imputation.

Table 1. Median, first, and third quartiles for the indicators of the impact of the missing data, expressed as percentages.

Measure	FMI	RE
TG	30.2 (28.7; 32.6)	99.4 (99.4; 99.4)
HDL-c	25.9 (23.1; 28.9)	99.5 (99.4; 99.5)
AC	28.6 (24.3; 29.7)	99.4 (99.4; 99.5)
WC/Height ratio	11.6 (9.6; 14.6)	99.8 (99.7; 99.8)
MAP	8.9 (7.4; 10.9)	99.8 (99.8; 99.9)
HOMA-IR	29.2 (27.6; 32.5)	99.4 (99.4; 99.5)

FMI = fraction of missing information; RE = relative efficiency; TG = triglycerides; HDL-c = high-density lipoprotein cholesterol; AC = atherogenic coefficient; WC/Height ratio = waist-to-height ratio; MAP = mean arterial pressure; HOMA-IR = homeostatic model assesment of insulin resistance.

2.4. Quantile Regression

Quantile regression models (QRMs) were introduced in 1978 by Koenker and Bassett [39]. They offer a natural extension of the classical linear regression models in which, instead of specifying the change in the conditional mean of the dependent variable's distribution associated with a change in the independent variables, the change in any conditional quantile of the distribution is specified. In longitudinal studies, QRMs have been applied in a wide variety of problems. For instance, Lipsitz et al. [40] used this technique for analyzing the changes in the distribution of CD4 cell counts in patients with human immunodeficiency virus. They are also commonly used for identifying risk factors in particular populations. Fenske et al. [41] applied a QRM for detecting obesity risk factors in childhood.

Mathematically, given the dependent variable Y, the k-dimensional independent variable X, and the τth quantile with $\tau \in (0,1)$, the QRM can be specified as follows:

$$Y = \beta_\tau \cdot X + \epsilon_\tau,$$

where the residuals verify that $\mathcal{P}(\epsilon_\tau \leq 0|X) = \tau$; that is, its conditional τth quantile, $q_\tau(\cdot|\cdot)$ is zero. Therefore, we have

$$q_\tau(Y|X) = \beta_\tau \cdot X + q_\tau(\epsilon_\tau|X) = \beta_\tau \cdot X.$$

Let $\{X_n, Y_n\}$ be a random sample from $\{X, Y\}$ (sample size n). The estimator $\hat{\beta}_\tau$ is obtained by minimizing a sum of weighted absolute residuals that gives asymmetric penalties depending on whether the values of the dependent variable are being overestimated or underestimated:

$$\tau \cdot \sum_{\epsilon_{\tau_i} \geq 0} |\epsilon_{\tau_i}| + (1-\tau) \cdot \sum_{\epsilon_{\tau_i} < 0} |\epsilon_{\tau_i}| \quad (1 \leq i \leq n). \tag{5}$$

This means that the proportion of data points below the τth estimating regression line $\hat{y}_i = \hat{\beta}_\tau \cdot X_i$ $(1 \leq i \leq n)$ is τ and the proportion lying above it is $1 - \tau$. Equation (5) can be minimized using different algorithms based on linear programming [42].

The interpretation of the coefficient estimates is analogous to those in classical linear regression, except that instead of referring to the effect on the conditional mean of the dependent variable, we refer to the conditional quantile. Each $\hat{\beta}_\tau$ can be interpreted as the increment of the τth quantile of the dependent variable per unit of change in the value of the corresponding independent variable, while the rest of the independent variables are fixed.

There are several procedures for computing both the standard errors and confidence intervals for the quantile regression coefficients. Under certain conditions, the usual coefficient estimators are asymptotically normally distributed [42]. However, asymptotic standard errors are complex, and resampling approaches are frequently employed [43].

QRMs overcome some limitations of classical linear regression tools, even if the researcher is only interested in a central position and its behavior. Box 4 provides some guidance on the situations for which a QRM may be appropriate. The last two points are the keys to its usefulness in evaluating tracking. But it is worth noting that the last point also makes these models highly suitable for assessing whether the effects of an exposure are the same in all quantiles. And the third point also solves the incredibly common cases where exposures follow skewed distributions.

Box 4. Situations in which quantile regression is useful.

1. **In the presence of outliers**. It is able to cope better with outliers, since it is based on the estimation of a position measure such as the quantile. Outliers only have an influence on the estimation of the quantile close to them.
2. **In case of heteroscedasticity**. If the variance depends on the independent variables, quantile regression can capture this effect.
3. **When distributional assumptions are not satisfied**. QRMs do not make assumptions about the distribution of errors, and thus they can be used when the conditions for applying other regression models are not satisfied.
4. **When the interest is at the extremes of the distribution**. Sometimes the real interest of the research question lies in what happens in the tails of the distribution. The QRM allows one to answer this question by estimating the extreme quantiles.
5. **When there is no known threshold defining the at-risk population**. As the model can be estimated for any quantile, it becomes possible to evaluate the impact of the independent variables on a specific section of the distribution without having to select a particular point.

The INMA-Asturias Cohort Example

Taking the TG measure as an example, we estimated the following QRMs for τ values ranging from 0.1 to 0.9 in 0.05 intervals:

$$q_\tau(TG_8|rank(TG_4)) = \hat{\beta}_{\tau_0} + \hat{\beta}_{\tau_1} \cdot rank(TG_4),$$

where TG_8 represents the TG measure at 8 years and $rank(TG_4)$ is the rank transformation of the TG measure at 4 years. As previously mentioned, the tracking conception is based on the relative positions of subjects within the distribution of the variable of interest. In order to incorporate this relative position within the independent variable, a rank transformation was applied. Here, we use the crude analysis as an example for simplicity, but as in any regression model, adjustment variables can be included.

Our aim is studying the impact on the upper tail of the TG at the 8 year distribution, (i.e., to estimate $\hat{\beta}_{\tau_1}$ for high τ values). However, estimating the effect for quantiles across the whole distribution and plotting $\hat{\beta}_{\tau_1}$ estimates against τ serves as a useful exploratory tool to assess whether the size and nature of the effect remains constant. Figure 2 shows that the association differed for high-risk subjects (those at the highest quantiles of TG at the 8 year distribution) compared with average subjects (those around the 0.5 quantile), reflecting an increasing trend in the the association's effect. This observation would not have been possible using classical linear regression models.

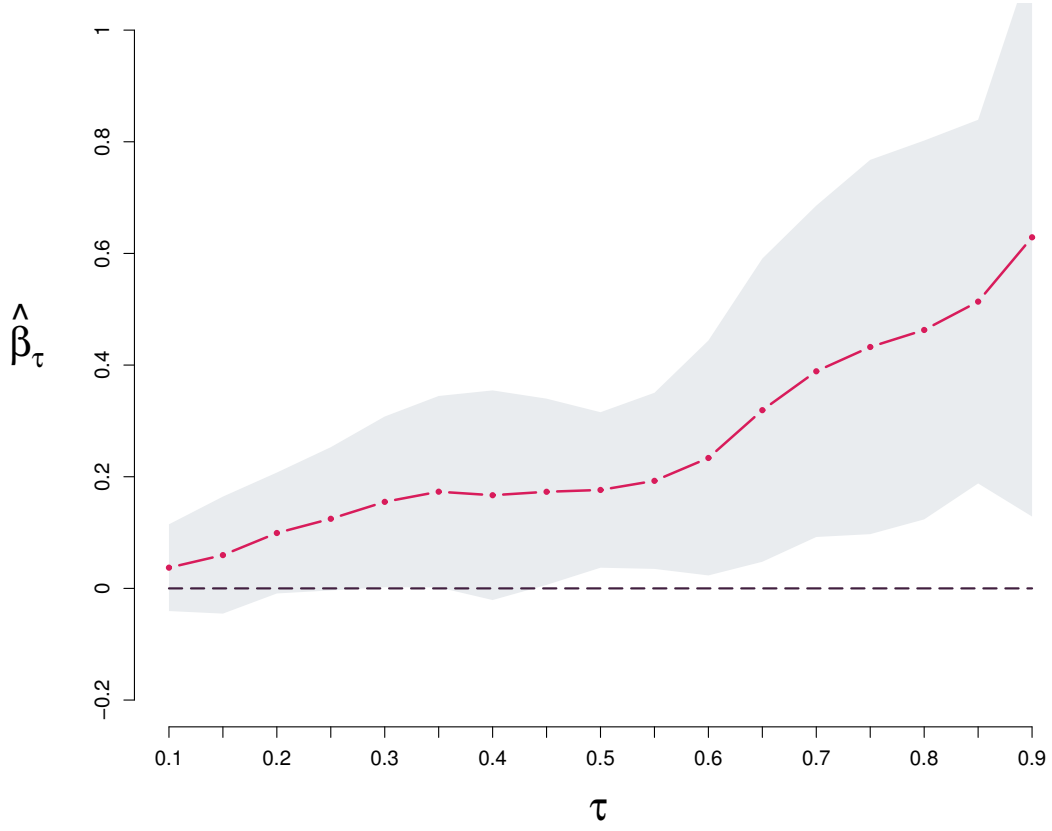

Figure 2. Quantile regression parameters ($\hat{\beta}_\tau$) per quantile (τ) for the effect of the rank transformation of TG at 4 on TG at 8. The red dots and lines represent point estimate of the parameters, while the grey bounds represent the confidence interval estimate. TG = triglycerides.

3. Results

The final results of the analysis may depend on the methodological decisions made. Regarding the missing data, the possible alternative here would be to conduct a complete case analysis. We compared the results between these two approaches and did not observe any systematic differences. However, contrary to what might be expected, not all

confidence intervals were narrower when applying MI in contrast to the complete case analysis. Figure 3 shows an example of the variances in the parameter estimates for several TG at 8 years quantiles by analysis type, and we can observe that the estimates were not more accurate for all parameters when using MI ($\tau = 0.3$ and $\tau = 0.5$). We observed this phenomenon in all the models involving measures with a high percentage of missing data (TG, HDL-c, AC, and HOMA-IR) but not in models involving the WC/Height ratio and MAP, which had less than 10% missing data.

Regarding the statistical model, a binary response model such as logistic regression could have been considered as an alternative to quantile regression. With this approach, we would still focus on the upper tail of the distribution and explore the probability of being in a high-risk TG category at 8 years, depending on the TG values at 4 years. For that purpose, we considered the 0.9 quantile, which is both age- and sex-specific, to calculate the binary variable that divided TG into the normal category (TG $<$ 0.9 quantile) and the risk category (TG $\geq$ 0.9 quantile). In our sample, without imputation, 72.3% of the 4-year-old children had normal TG levels, 8.2% had risk levels, and 19.5% had missing data. At 8 years, 71.6% of the children had normal levels, 8.5% had risk levels, and 19.9% had missing data.

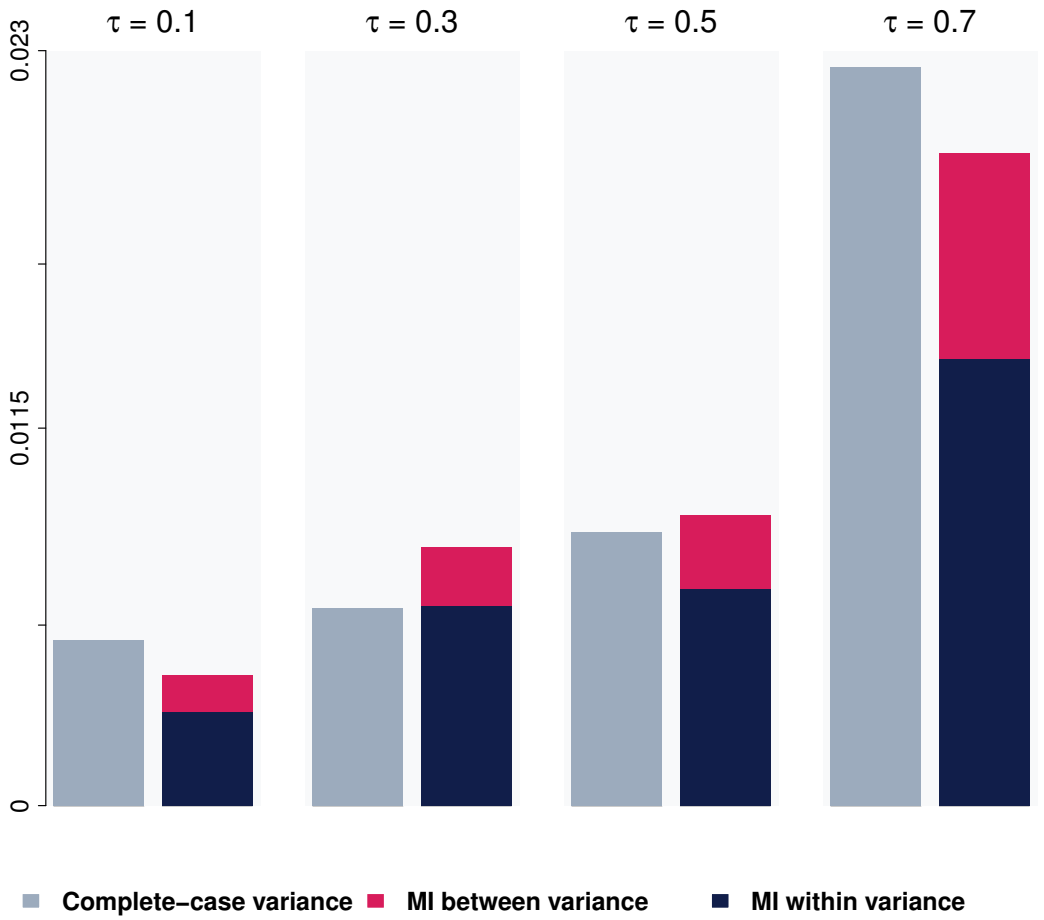

Figure 3. Variance of parameter estimates in quantile regression models for TG at 8 years and TG at 4 years by type of analysis: complete case or MI analysis. TG = triglycerides; MI = multiple imputation; τ = quantile.

Using a QRM, we observed a positive association between the rank of TG at 4 years and the 0.9 quantile of TG at 8 years ($\hat{\beta}_{0.9}$: 0.629, 95%CI: 0.129–1.129). The logistic regression model showed a positive association between the rank of TG at 4 years and the odds of being in the risk category of TG at 8 years (odds ratio (OR): 1.009, 95%CI: 0.995–1.023). While the observed association and overall conclusion were the same, the estimated parameters were not directly comparable. In the QRM, $\hat{\beta}_{0.9}$ represents an additive effect on the dependent

variable, whereas the OR in the logistic regression model represents a multiplicative effect. More specifically, for the same one-unit increase in the rank of TG at 4 years, in the first case, we estimated a 0.629 mg/dL increase in the 0.9 TG quantile at 8 years, while in the second case, we estimated there to be 1.009 times the risk of being in the TG risk category at 8 years. Moreover, the outcome did not represent the same construct. In the QRM, the outcome was a specific point of the TG distribution at 8 years, while in logistic regression, it was a section of the distribution, assuming no variation in the effect within that section. Another option would be to use the binary TG variable for both 4 and 8 years in the logistic regression. In this case, we found that children who were in the risk category at 4 years were 3.287 times (95%CI: 1.173–9.212) more likely to be in the risk category at age 8 than those who were not. Again, the evidence on the nature of the association between variables was the same, as high TG values at 4 years were positively associated with high TG values at 8 years, but the estimated effects were not comparable.

4. Discussion

The epidemiology literature has plenty of statistical analysis. Despite these usually being briefly explained, it is never clear what the impact on the observed results would be if a different decision was made. Lack of space is a common problem in specialized journals, and deep explanations are relegated to Supplementary Materials or directly omitted. Here, we explored the impact of the decisions taken, particularly with regard to missing data and the selection of the most appropriate statistical model for the study of variables involving controversial thresholds.

In recent years, MI has become a quite popular method for dealing with missing data. As we saw in Section 2.3, the most appropriate approach depends on the data missingness mechanism and on the amount of missing data or the role played by the involved variables (dependent or independent variables, adjustment variables, etc.) [44]. Several authors recommend the use of MI procedures regardless of the mechanism of missingness [45]. They argue that, under the MCAR condition, it is preferred against a complete case analysis because it results in more power. Under the MAR condition, it is preferred because, aside from more power, it will give unbiased results, whereas complete case analysis may not. And under the MNAR model, some authors suggest that it will provide less biased results than complete case analysis [46]. However, the decision is not always straightforward, and using MI only to maximize the sample size is a kind of artificial approach, which may not always be successful when it is correctly performed. Here, we presented an example where utilizing MI resulted in a higher sum of within-imputation variance and between-imputation variance and, consequently, total variance for certain quantiles compared with the variance obtained through complete case analysis. This may occur in cases where the proportion of imputed data is large and there are no variables closely related to the missing data or to the variables containing the missing data themselves. The MI model would reflect the high uncertainty around the missing data, and the target parameter estimation would be highly dependent on the generated datasets. This adds extra noise and increases uncertainty when combining the results, and it potentially leads to higher between-imputation variance values that offset the gain in the within-imputation variance resulting from the increase in sample size. Another example in which MI might yield to less precise confidence intervals, despite an increased sample size, is when there is a large proportion of missing data in the explanatory variables, and these are highly correlated with the response variable. In this case, MI can affect the precision of the estimates.

We considered tracking analysis of the cardiovascular-related measures—particularly TG—in healthy children as an example of analysis that requires avoiding the use of arbitrary thresholds while focusing on the extreme parts of the outcome distribution. Correlation coefficients and linear regression models are frequently used to explore tracking while preserving the continuous nature of the variables. These methods would focus on estimating the effect of TG at 4 years on the average TG at 8 years. Nevertheless, this does not provide us with any information on the magnitude or direction of the association

in the upper part of the TG distribution. In contrast, quantile regression addresses this constraint and allows us to assess the impact within the region of interest without relying on quantile-based categorization. We compared our approach with two variations of the classical logistic regression analysis using thresholds. The overall finding was the same: There was a positive association between the high TG values at 4 and 8 years of age. However, quantile regression provided much richer information. If there were a clear threshold enabling the categorization of a cardiovascular-related variable into normal and risk values, then logistic regression would allow us to estimate the effect on the probability of being in the risk category at 8 years associated with an increase in TG values at 4 years, thus providing an estimation of the tracking of the variable between these ages. However, in the specific case of cardiovascular-related variables in children, where consensus on the threshold values is lacking, we are truly estimating the effect on the probability of being in a category that holds no clinical significance. And we are also assuming homogeneity of risk within categories. In other words, the risk is the same for all individuals within the normal category and the same for all individuals within the risk category. On the contrary, quantile regression allows us to estimate this effect across all quantiles, thus covering the entire part of the distribution that may imply potential risk. In our example, using quantile regression, we were able to observe that the effect was not constant across all quantiles of the distribution at age 8. Instead, it increased as the quantile increased. Using logistic regression, we would not be able to see this behavior.

This suggests that the magnitude of tracking increases the more extreme the values are, providing relevant insights. While there is no established risk threshold for TG in pediatric ages, our findings indicate that increasing TG levels at 4 years may lead not only to a higher average at 8 years but also to a longer upper tail of the TG distribution at 8 years. Although it is not the purpose of this article, it should be mentioned that this could imply difficulty in normalizing TG values in the future for those children who present extreme values at 4 years of age and a progressive increase in TG values at 8 years of age. These results have potential implications for children's health, as the consequences associated with such changes in TG levels are not yet known. These findings can also inform the identification of cardiovascular-related measures that should be considered as targets for screening and monitoring in clinical practice, as well as in the development of public health guidelines and recommendations for children [11].

Quantile regression has gained widespread popularity in social science, economics, environmental modeling, public health research [47–50], and in recent years, in the field of environmental pollutant exposure [51–55]. In longitudinal data analysis, which suffers from a high level of complexity due to the intercorrelation among repeatedly measured observations, QRMs have also gained increasing popularity. Most longitudinal modeling methods primarily focus on mean regression, concentrating solely on the average effects of covariates and the mean trajectory of longitudinal outcomes. Consequently, similar to independent data, quantile regression has also been extended and applied to longitudinal data. Quantile regression for longitudinal data possesses the capacity, at both the population and individual levels, to identify heterogeneous covariate effects, elucidate variations in longitudinal changes across different quantiles of the outcome, and offer more robust estimates when heavy-tailed distributions and outliers are present [56]. Despite this, its application in longitudinal cohort studies for tracking purposes has been limited [57]. This work serves as an example of its potential for investigating risk variables without known thresholds or when research interests lie in non-central areas of the distribution, as occurs in tracking studies or also when evaluating the possible effects of exposures. Even in other cases, it can complement traditional analysis methods by estimating a family of conditional quantile functions, providing a more nuanced understanding of variable effects.

5. Conclusions

Details are important in statistical analysis, as they can impact the final results. In our data, the findings seemed to be robust with respect the to main decisions taken but led

to differences in terms of accuracy and richness of information obtained. Here we point out that although multiple imputation methods are generally useful for mitigating biases in estimates, they may not necessarily improve the precision of standard error estimates. Moreover, we illustrate that quantile regression can be a powerful tool in addressing challenges associated with controversial threshold definitions and tracking analyses in cohort studies, providing valuable additional information. Given the strengths of these models, they should be considered in analyses of continuous outcomes, at least as a first step for making future modeling decisions. Finally, it is always unclear what impact different decisions would have on the obtained results, and there are always numerous alternatives to choose from. Therefore, it is essential to describe and report precisely how the analysis was conducted, including its limitations and strengths, even if it has to be included in Supplementary Materials.

6. Computational Considerations

Nowadays, there are many resources that allow a wide range of statistical analyses to be performed, including those that may require a high computational capacity, such as the ones presented here. In this work, we used R statistical software (version 4.2.1; R Foundation for Statistical Computing, Vienna, Austria, www.r-project.org). In particular, we used the package *MICE* [32] developed by van Buuren and Groothuis-Outshoorn, which includes several different imputation model options to perform multivariate imputation with chained equations. The package *quantreg* [58] was used for QRM estimation and inference, which provides several alternative methods to estimate model parameters and to compute standard errors.

Supplementary Materials: The following supporting information can be downloaded at www.mdpi.com/article/10.3390/math11194070/s1. Table S1: Number of participants with missing data for each variable, expressed in absolute and relative frequencies, for the final sample composed by 307 children.

Author Contributions: Conceptualization, R.F.-I., A.F.-S. and P.M.-C.; data curation, R.F.-I.; formal analysis, R.F.-I.; funding acquisition, A.T.; investigation, R.F.-I., A.F.-S. and P.M.-C.; methodology, R.F.-I., A.F.-S. and P.M.-C.; supervision, A.F.-S., P.M.-C. and A.T.; visualization, R.F.-I. and P.M.-C.; writing—original draft, R.F.-I. and P.M.-C.; writing—review and editing, A.F.-S. and P.M.-C. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by grants from CIBERESP (PhD employment contract and fellowship), ISCIII: PI04/2018, PI09/02311, PI13/02429, PI18/00909 co-funded by FEDER, "A way to make Europe"/"Investing in your future", Fundación Cajastur, and Universidad de Oviedo.

Institutional Review Board Statement: The study was conducted while conforming to the principles of the Declaration of Helsinki, and its protocol was approved by the Asturias Regional Ethics Committee.

Informed Consent Statement: Written informed consent was obtained from every participanting woman and, in such cases, her partner.

Data Availability Statement: The data and computing code are available for replication from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SAGES	Successful Aging after Elective Surgery
CVD	Cardiovascular disease
EPIC	European Prospective Intake and Cancer
ECHO	Environmental Influences On Child Health Outcomes
INMA	INfancia y Medio Ambiente (Environment and Childhood)

NHBC	New Hampshire Birth Cohort
WC/Height ratio	Waist-to-height ratio
MAP	Mean arterial pressure
TG	Triglycerides
HDL-c	High-density lipoprotein cholesterol
AC	Atherogenic coefficient
HOMA-IR	Homeostatic model assessment of insulin resistance
MCAR	Missing completely at random
MAR	Missing at random
MNAR	Missing not at random
MI	Multiple imputation
FMI	Fraction of missing information
RE	Relative efficiency
MICE	Multivariate imputation by chained equations
QRMs	Quantile regression models

References

1. Andreu-Perez, J.; Poon, C.C.Y.; Merrifield, R.D.; Wong, S.T.C.; Yang, G.Z. Big Data for Health. *IEEE J. Biomed. Health Inform.* **2015**, *19*, 1193–1208. [CrossRef] [PubMed]
2. Rodriguez-Martinez, A.; Zhou, B.; Sophiea, M.K.; Bentham, J.; Paciorek, C.J.; Iurilli, M.L.; Carrillo-Larco, R.M.; Bennett, J.E.; Di Cesare, M.; Taddei, C.; et al. Height and body-mass index trajectories of school-aged children and adolescents from 1985 to 2019 in 200 countries and territories: A pooled analysis of 2181 population-based studies with 65 million participants. *Lancet* **2020**, *396*, 1511–1524. [CrossRef] [PubMed]
3. Schmitt, E.; Saczynski, J.; Kosar, C.; Jones, R.; Alsop, D.; Fong, T.; Metzger, E.; Cooper, Z.; Marcantonio, E.R.; Travison, T.; et al. The successful aging after elective surgery study: Cohort description and data quality procedures. *J. Am. Geriatr. Soc.* **2015**, *63*, 2463–2471. [CrossRef] [PubMed]
4. Tsao, C.W.; Vasan, R.S. Cohort Profile: The Framingham Heart Study (FHS): Overview of milestones in cardiovascular epidemiology. *Int. J. Epidemiol.* **2015**, *44*, 1800–1813. [CrossRef] [PubMed]
5. Riboli, E.; Kaaks, R. The EPIC Project: Rationale and study design. European Prospective Investigation into Cancer and Nutrition. *Int. J. Epidemiol.* **1997**, *26*, 6–14. [CrossRef]
6. Blaisdell, C.J.; Park, C.; Hanspal, M.; Roary, M.; Arteaga, S.S.; Laessig, S.; Luetkemeier, E.; Gillman, M.W. The NIH ECHO Program: Investigating how early environmental influences affect child health. *Pediatr. Res.* **2022**, *92*, 1215–1216. [CrossRef]
7. Guxens, M.; Ballester, F.; Espada, M.; Fernández, M.F.; Grimalt, J.O.; Ibarluzea, J.; Olea, N.; Rebagliato, M.; Tardon, A.; Torrent, M.; et al. Cohort profile: The INMA–INfancia y Medio Ambiente–(Environment and Childhood) Project. *Int. J. Epidemiol.* **2012**, *41*, 930–940. [CrossRef]
8. Gilbert-Diamond, D.; Cottingham, K.L.; Gruber, J.F.; Punshon, T.; Sayarath, V.; Gandolfi, A.J.; Baker, E.R.; Jackson, B.P.; Folt, C.L.; Kargas, M.R.; et al. Rice consumption contributes to arsenic exposure in US women. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 20656–20660. [CrossRef]
9. Lurbe, E.; Agabiti-Rosei, E.; Cruickshank, J.K.; Dominiczak, A.; Erdine, S.; Hirth, A.; Invitti, C.; Litwin, M.; Mancia, G.; Pall, D.; et al. 2016 European Society of Hypertension guidelines for the management of high blood pressure in children and adolescents. *J. Hypertens.* **2016**, *34*, 1887–1920. [CrossRef]
10. Fernández-Somoano, A.; Estarlich, M.; Ballester, F.; Fernández-Patier, R.; Aguirre-Alfaro, A.; Herce-Garraleta, M.D.; Tardón, A. Outdoor NO2 and benzene exposure in the INMA (Environment and Childhood) Asturias cohort (Spain). *Atmos. Environ.* **2011**, *45*, 5240–5246. [CrossRef]
11. Fernández-Iglesias, R.; Martinez-Camblor, P.; Fernández-Somoano, A.; Rodríguez-Dehli, C.; Venta-Obaya, R.; Karagas, M.R.; Tardón, A.; Riaño-Galán, I. Tracking between cardiovascular-related measures at 4 and 8 years of age in the INMA-Asturias cohort. *Eur. J. Pediatr.* 2023, *online ahead of print*. [CrossRef] [PubMed]
12. Binkin, N.J.; Yip, R.; Fleshood, L.; Trowbridge, F.L. Birth weight and childhood growth. *Pediatrics* **1988**, *82*, 828–834. [CrossRef]
13. Rosner, B.; Hennekens, C.H.; Kass, E.H.; Miall, W.E. Age-specific correlation analysis of longitudinal blood pressure data. *Am. J. Epidemiol.* **1977**, *106*, 306–313. [CrossRef] [PubMed]
14. Berenson, G.S.; Foster, T.A.; Frank, G.C.; Frerichs, R.R.; Srinivasan, S.R.; Voors, A.W.; Webber, L.S. Cardiovascular disease risk factor variables at the preschool age. The Bogalusa heart study. *Circulation* **1978**, *57*, 603–612. . [CrossRef] [PubMed]
15. Clarke, W.R.; Schrott, H.G.; Leaverton, P.E.; Connor, W.E.; Lauer, R.M. Tracking of blood lipids and blood pressures in school age children: The Muscatine study. *Circulation* **1978**, *58*, 626–634. [CrossRef] [PubMed]
16. Milei, J.; Ottaviani, G.; Lavezzi, A.M.; Grana, D.R.; Stella, I.; Matturri, L. Perinatal and infant early atherosclerotic coronary lesions. *Can. J. Cardiol.* **2008**, *24*, 137–141. [CrossRef]
17. Mcgill, H.C.; Mcmahan, C.A.; Herderick, E.E.; Malcom, G.T.; Tracy, R.E.; Strong, J.P. Origin of atherosclerosis in childhood and adolescence. *Am. J. Clin. Nutr.* **2000**, *72*, 1307S–1315S. [CrossRef]

18. Wang, Y.; Wang, X. How do statistical properties influence findings of tracking (maintenance) in epidemiologic studies? An example of research in tracking of obesity. *Eur. J. Epidemiol.* **2003**, *18*, 1037–1045. [CrossRef]
19. Ragland, D.R. Dichotomizing continuous outcome variables: Dependence of the magnitude of association and statistical power on the cutpoint. *Epidemiology* **1992**, *3*, 434–440. [CrossRef]
20. Altman, D.G.; Royston, P. The cost of dichotomising continuous variables. *BMJ* **2006**, *332*, 1080. [CrossRef]
21. Bennette, C.; Vickers, A. Against quantiles: Categorization of continuous variables in epidemiologic research, and its discontents. *BMC Med. Res. Methodol.* **2012**, *12*, 21. [CrossRef] [PubMed]
22. Sarganas, G.; Schaffrath Rosario, A.; Niessner, C.; Woll, A.; Neuhauser, H.K. Tracking of Blood Pressure in Children and Adolescents in Germany in the Context of Risk Factors for Hypertension. *Int. J. Hypertens.* **2018**, *2018*, 8429891. [CrossRef] [PubMed]
23. Joshi, S.M.; Katre, P.A.; Kumaran, K.; Joglekar, C.; Osmond, C.; Bhat, D.S.; Lubree, H.; Pandit, A.; Yajnik, C.S.; Fall, C.H. Tracking of cardiovascular risk factors from childhood to young adulthood—The Pune Children's Study. *Int. J. Cardiol.* **2014**, *175*, 176–178. [CrossRef] [PubMed]
24. De Wilde, J.A.; Middelkoop, B.J.; Verkerk, P.H. Tracking of thinness and overweight in children of Dutch, Turkish, Moroccan and South Asian descent from 3 through 15 years of age: A historical cohort study. *Int. J. Obes.* **2018**, *42*, 1230–1238. [CrossRef]
25. Little, R.J.A.; Rubin, D.B. *Statistical Analysis with Missing Data*, 3rd ed.; Wiley: Hoboken, NJ, USA, 2019.
26. Schafer, J.L.; Graham, J.W. Missing data: Our view of the state of the art. *Psychol. Methods* **2002**, *7*, 147–177. [CrossRef]
27. Kristman, V.L.; Manno, M.; Côté, P. Methods to account for attrition in longitudinal data: Do they work? A simulation study. *Eur. J. Epidemiol.* **2005**, *20*, 657–662. [CrossRef]
28. Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys*; Wiley: New York, NY, USA, 1987.
29. Laqueur, H.S.; Shev, A.B.; Kagawa, R.M.C. SuperMICE: An Ensemble Machine Learning Approach to Multiple Imputation by Chained Equations. *Am. J. Epidemiol.* **2021**, *191*, 516–525. [CrossRef]
30. Van Buuren, S. *Flexible Imputation of Missing Data*, 2nd ed.; Chapman & Hall: London, UK, 2018.
31. Little, R.J. A test of missing completely at random for multivariate data with missing values. *J. Am. Stat. Assoc.* **1988**, *83*, 1198–1202. [CrossRef]
32. Van Buuren, S.; Groothuis-Oudshoorn, K. MICE: Multivariate imputation by chained equations. *J. Stat. Softw.* **2011**, *45*, 1–67. [CrossRef]
33. Van Buuren, S.. *Flexible Imputation of Missing Data*; Chapman & Hall/CRC Interdisciplinary Statistics: London, UK, 2012; p. 342.
34. Austin, P.C.; White, I.R.; Lee, D.S.; van Buuren, S. Missing Data in Clinical Research: A Tutorial on Multiple Imputation. *Can. J. Cardiol.* **2021**, *37*, 1322–1331. [CrossRef]
35. Lee, K.J.; Roberts, G.; Doyle, L.W.; Anderson, P.J.; Carlin, J.B. Multiple imputation for missing data in a longitudinal cohort study: A tutorial based on a detailed case study involving imputation of missing outcome data. *Int. J. Soc. Res. Methodol.* **2016**, *19*, 575–591. [CrossRef]
36. White, I.R.; Royston, P.; Wood, A.M. Multiple imputation using chained equations: Issues and guidance for practice. *Stat. Med.* **2011**, *30*, 377–399. [CrossRef] [PubMed]
37. Graham, J.W.; Olchowski, A.E.; Gilreath, T.D. How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prev. Sci.* **2007**, *8*, 206–213. [CrossRef]
38. Bodner, T.E. What improves with increased missing data imputations? *Struct. Equ. Model. A Multidiscip. J.* **2008**, *15*, 651–675. [CrossRef]
39. Koenker, R.; Bassett, G. Regression Quantiles. *Econometrica* **1978**, *46*, 33–50. [CrossRef]
40. Lipsitz, S.R.; Fitzmaurice, G.M.; Molenberghs, G.; Zhao, L.P. Quantile Regression Methods For Longitudinal Data with Drop-Outs: Application to CD4 Cell Counts of Patients Infected with the Human Immunodeficiency Virus. *J. R. Stat. Soc. Ser. C* **1997**, *46*, 463–476. [CrossRef]
41. Fenske, N.; Fahrmeir, L.; Rzehak, P.; Höhle, M. *Detection of Risk Factors for Obesity in Early Childhood with Quantile Regression Methods for Longitudinal Data*; Technical Report; University of Munich: Munich, Germany, 2008. [CrossRef]
42. Koenker, R. *Quantile Regression*; Cambridge University: Cambridge, UK, 2005.
43. Hao, L.; Naiman, D.Q. *Quantile Regression*; Sage Publications: Thousand Oaks, CA, USA, 2007.
44. Enders, C.K. *Applied Missing Data Analysis*; Guilford Press: New York, NY, USA, 2010.
45. Van Ginkel, J.R.; Linting, M.; Rippe, R.C.; van der Voort, A. Rebutting Existing Misconceptions About Multiple Imputation as a Method for Handling Missing Data. *J. Personal. Assess.* **2020**, *102*, 297–308. [CrossRef]
46. Schafer, J.L. *Analysis of Incomplete Multivariate Data*, 1st ed.; Chapman & Hall: London, UK, 1997.
47. Yu, K.; Lu, Z.; Stander, J. Quantile regression: Applications and current research areas. *J. R. Stat. Soc.* **2003**, *52*, 331–350. [CrossRef]
48. Staffa, S.J.; Kohane, D.S.; Zurakowski, D. Quantile Regression and Its Applications: A Primer for Anesthesiologists. *Anesth. Analg.* **2019**, *128*, 820–830. [CrossRef]
49. Oconnor, C. Robust estimates of vulnerability to poverty using quantile models. *Econ. Model.* **2023**, *123*, 106274. [CrossRef]
50. Amjad, M.; Akbar, M. The Association between Fruit and Vegetable Intake and Socioeconomic Factors in the Households of Pakistan Using Quantile Regression Model. *Soc. Work Public Health* **2023**, *38*, 248–258. [CrossRef] [PubMed]
51. Wei, Y.; Kehm, R.D.; Goldberg, M.; Terry, M.B. Applications for Quantile Regression in Epidemiology. *Curr. Epidemiol. Rep.* **2019**, *6*, 191–199. [CrossRef]

52. Peralta, A.A.; Schwartz, J.; Gold, D.R.; Vonk, J.M.; Vermeulen, R.; Gehring, U. Quantile regression to examine the association of air pollution with subclinical atherosclerosis in an adolescent population. *Environ. Int.* **2022**, *164*, 107285. [CrossRef] [PubMed]
53. Strickland, M.; Lin, Y.; Darrow, L.; Warren, J.; Mulholland, J.; Chang, H. Associations Between Ambient Air Pollutant Concentrations and Birth Weight: A Quantile Regression Analysis. *Epidemiology* **2019**, *30*, 624–632. . [CrossRef]
54. Cowell, W.; Jacobson, M.H.; Long, S.E.; Wang, Y.; Khan, L.G.; Ghassabian, A.; Naidu, M.; Torshizi, G.D.; Afanasyeva, Y.; Liu, M.; et al. Maternal urinary bisphenols and phthalates in relation to estimated fetal weight across mid to late pregnancy. *Environ. Int.* **2023**, *174*, 107922. [CrossRef]
55. Kapwata, T.; Wright, C.Y.; Reddy, T.; Street, R.; Kunene, Z.; Mathee, A. Environmental Science and Pollution Research Relations between personal exposure to elevated concentrations of arsenic in water and soil and blood arsenic levels amongst people living in rural areas in Limpopo, South Africa. *Environ. Sci. Pollut. Res.* **2023**, *30* , 65204–65216. [CrossRef]
56. Huang, Q.; Zhang, H.; Chen, J.; He, M. Quantile Regression Models and Their Applications: A Review. *J. Biom. Biostat.* **2017**, *8*, 354. [CrossRef]
57. Norris, T.; Bann, D.; Hardy, R.; Johnson, W. Socioeconomic inequalities in childhood-to-adulthood BMI tracking in three British birth cohorts. *Int. J. Obes.* **2020**, *44*, 388–398. [CrossRef]
58. Koenker, R. *Quantreg: Quantile Regression*, R Package Version 5.94; R Foundation for Statistical Computing: Vienna, Austria, 2017. Available online: https://CRAN.R-project.org/package=quantreg (accessed on 25 August 2023).