European Journal of Clinical Nutrition www.nature.com/ejcn

ARTICLE

Detecting potential outliers in longitudinal data with time-dependent covariates

Lazarus K. Mramba[1]✉, Xiang Liu[1], Kristian F. Lynch[1], Jimin Yang[1], Carin Andrén Aronsson[2,3], Sandra Hummel[4], Jill M. Norris[5], Suvi M. Virtanen[6,7,8,9], Leena Hakola[8,9], Ulla M. Uusitalo[1] and Jeffrey P. Krischer[1]

BACKGROUND: Outliers can influence regression model parameters and change the direction of the estimated effect, over-estimating or under-estimating the strength of the association between a response variable and an exposure of interest. Identifying visit-level outliers from longitudinal data with continuous time-dependent covariates is important when the distribution of such variable is highly skewed.
OBJECTIVES: The primary objective was to identify potential outliers at follow-up visits using interquartile range (IQR) statistic and assess their influence on estimated Cox regression parameters.
METHODS: Study was motivated by a large TEDDY dietary longitudinal and time-to-event data with a continuous time-varying vitamin B_{12} intake as the exposure of interest and development of Islet Autoimmunity (IA) as the response variable. An IQR algorithm was applied to the TEDDY dataset to detect potential outliers at each visit. To assess the impact of detected outliers, data were analyzed using the extended time-dependent Cox model with robust sandwich estimator. Partial residual diagnostic plots were examined for highly influential outliers.
RESULTS: Extreme vitamin B_{12} observations that were cases of IA had a stronger influence on the Cox regression model than non-cases. Identified outliers changed the direction of hazard ratios, standard errors, or the strength of association with the risk of developing IA.
CONCLUSION: At the exploratory data analysis stage, the IQR algorithm can be used as a data quality control tool to identify potential outliers at the visit level, which can be further investigated.

European Journal of Clinical Nutrition; https://doi.org/10.1038/s41430-023-01393-6

INTRODUCTION

In any statistical analysis, data are examined for unusual observations [1]. Outliers can be defined as observations that are unusually larger or smaller as compared to other data points of the same variable [2]. Several studies have suggested ways of defining outliers: they are observations that deviate extensively from the overall pattern or expectation of the other data points (see [3, 4]), they are observations with large residual values [5] or they are data values falling outside of an expected range [6].

The cause of the extreme values may be the result of uncorrected data entry, system errors, self-reporting bias or they can be true observations that are due to rare events. Regardless, these outliers may have a large influence on the regression model parameters that can change the direction of the effect, mask the effect, underestimate the effect, or overestimate the effect [7, 8].

Detection of outliers from longitudinal and time to event studies with continuous time-varying covariates provides a challenge in most applied research areas [9]. Typically, the process of data cleaning and management takes a considerable amount of time due to the complexity of large datasets. Yet, conducting exploratory data analysis to detect extreme values at an earlier stage of the statistical analysis is necessary to avoid misleading conclusions that are based on only a few data points. There are several common approaches to detecting outliers including the use of knowledge and experience of investigators, using other published cross-sectional references or eyeballing using graphical outputs [10] to come up with single upper and/or lower values-cut-offs across the dataset. Some studies [11] have used conditional growth percentiles to identify outliers in growth trajectory data and defined outliers to be observations 4 standard deviations (σ) away from the expected (conditional) value. Parameters, σ and mean (μ) are affected by extreme observations thus may not be suitable measures of spread and location, respectively, where the data are skewed [12]. In addition, conditional growth method cannot be applied to a subject's first measurement (visit). Robust regression methods have also been

[1]Health Informatics Institute, Morsani College of Medicine, University of South Florida, Tampa, FL, USA. [2]Department of Clinical Sciences, Lund University, Malmö, Sweden. [3]Department of Pediatrics, Skåne University Hospital, Malmö, Sweden. [4]Institute of Diabetes Research, Helmholtz Zentrum and Forschergruppe Diabetes, Klinikum rechts der Isar, Technische Universität and Forschergruppe Diabetes e.V, Munich, Germany. [5]Department of Epidemiology, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [6]Finnish Institute for Health and Welfare, Health and Well-Being Promotion Unit, Helsinki, Finland. [7]Center for Child Health Research, University of Tampere and Tampere University Hospital, Tampere, Finland. [8]Faculty of Social Sciences, Unit of Health Sciences, Tampere University, Tampere, Finland. [9]Tampere University Hospital, Wellbeing Services County of Pirkanmaa, Tampere, Finland. ✉email: Lazarus.Mramba@epi.usf.edu

Received: 14 August 2023 Revised: 6 December 2023 Accepted: 12 December 2023
Published online: 03 January 2024

used to detect outliers on a longitudinal electronic health data records [13]. Other methods include the use of Jackknife residuals [14] where they defined an outlier with a cut-off of ±5 in their longitudinal childhood growth data, and studentized residuals [15] with a cut-off of >2σ to define outliers, whereas [16] used a cut-off of ±6 for a longitudinal study on obesity prevalence and weight change in children and adolescents together with cross-sectional *z*-scores thresholds as defined by Centers for Disease Control.

Non-parametric statistics such as interquartile range (IQR) and median absolute deviation (MAD) are simple and alternative useful methods that can be applied to detect potential outliers during the exploratory data analysis stage. One study [17] calculated weight for length percentiles and then used the IQR method to detect and remove outliers in a longitudinal childhood genome study. The numerical IQR method, followed by graphical assessment using box plots, was described as an approach to successfully identify potential outliers in an educational achievement study to improve student learning [18]. MAD method has been used to detect and remove outliers by several studies (see [12, 19, 20]. Both IQR and MAD statistics are robust measures of dispersion that are more resilient to outliers than standard deviation. The IQR method measures the central half of the data for any shape of the distribution and is a great alternative measure of dispersion that does not require symmetry and has no distribution assumptions since it uses percentiles making it more robust to the presence of outliers.

The primary aim of this study was to describe how to identify potential outliers for longitudinal data with skewed distributions at follow-up visits using a non-parametric interquartile range statistic method. This method may be a specific tool/procedure for the statistical analyst with no prior additional knowledge or policy at hand for dealing with extreme values.

MATERIALS AND METHODS

Robustness of IQR and MAD statistics

The robustness of IQR and MAD has been shown by [21] that if T_n is a statistic on an ordered sample of size n, then, T_n has breakdown value b, $0 \le b \le 1$, if for every $\epsilon>0$, $lim_{\{X(\{(1-b)n\})\to\infty\}} T_n<\infty$ and $lim_{\{X(\{(1-(b+\epsilon))n\})\to\infty\}} T_n = \infty$. The sample median ($m$) remains unchanged in the presence of extreme low or high values. If less than 50% of the sample $\to \infty$, then m and MAD will remain the same. If more than 50% of the sample $\to \infty$, then $m \to \infty$ and so does MAD. The median in that case will be located within the outliers and thus MAD has a breakdown value of 50%. Similarly, IQR has a breakdown value of 25%, breaking down when Q_1 is located within the outliers [22].

Interquartile range algorithm

For each continuous time-dependent variable of interest, the following defines the IQR algorithm and can be applied overall or stratified by important subject level factors such as country or longitudinal variables such as age or follow-up visit:

(i) Calculate IQR $= Q_3 - Q_1$.
(ii) Define lower and upper limits of outliers as $[Q_1 - k\times \text{IQR}, Q_3 + k\times \text{IQR}]$ or $[Q_1 - h, Q_3 + h]$ where $h = k\times$ IQR and k>0 is a scale factor.
(iii) Flag observations outside the limits as potential outliers.

IQR lower-limits below zero can be set to zero, for instance, in food record data where intake cannot be negative.

Motivating TEDDY dietary data

This study was motivated by a large longitudinal and time-to-event dataset with time-varying covariates. Dietary intake data was obtained from The Environmental Determinants of Diabetes in the Young (TEDDY), which is an observational longitudinal study that investigates factors associated with Diabetes (T1D) in children [23]. Of the 8676 children that were enrolled in the TEDDY study, 120 were HLA ineligible, 22 had no food records, and 33 had no record on Islet Autoimmunity and were dropped, leaving a total of 8501 subjects with 152,426 records followed up from birth up until censoring at 10 years of age for this study. Diet was assessed by 24-h dietary recall at the age of 3 months, by 3-day food record at the age of 6, 9, and 12 months and every 6 months thereafter until the subject developed islet autoimmunity (IA) or was censored at the end of the study period. For the purpose of illustrating the IQR method, focus was on exposure to daily intake of vitamin B_{12} (μg/day), including intake from foods and dietary supplements [24] and the risk of developing IA.

Statistical analysis

The basic Cox proportional hazards model [25] (see also [26, 27]) assumes that exposure covariates are fixed. This model can be extended to introduce variables that vary continuously with time in the form of $z_i(t) = z_i g(t)$ and expressed as

$$h(t) = h_o(t)\exp\{\beta_1 x_1 + \cdots + \beta_k x_k + g(t)(\gamma_1 z_1 + \cdots + \gamma_m z_m)\}$$
$$= h_o(t)\exp\left(\sum_{i=1}^{k}\beta_i x_i + \sum_{i=1}^{m}\gamma_i z_i(t)\right)$$

where $\boldsymbol{Z} = \{z_1, \ldots, z_m\}$ are time-varying covariates and γ_i are regression coefficient for a covariate $g(t)z_i$, which is a function of time. With the data arranged using the counting process, the extended time-dependent cox model associating the risk of islet autoimmunity and exposure variables is expressed as

$$h(t|\boldsymbol{X},\boldsymbol{Z}) = h_o(t)\exp(\beta_1 \text{vitaminB}_{12}(t) + \beta_2 \text{sex} + \beta_3 \text{fdr} + \beta_4 \text{hla} + \beta_5 \text{country})$$

where vitamin B_{12} (μg/day) is the continuous time-varying exposure of interest, adjusted for the fixed effects covariates: sex, HLA DR3/4, FDR, and country. Standard errors (SE) were calculated using robust (empirical) variance sandwich estimator to account for correlations within subjects [28, 29]. Vitamin B_{12} (μg/day) was energy adjusted by country and visit using Willett's residual method [30].

Detecting highly influential observations using residual diagnostics

Residual diagnostics plots were obtained after fitting Cox regression models on the full TEDDY dietary data to further ascertain the influence of highly influential observations on the estimated parameters. Partial DFBETA measure of influence for vitamin B_{12} was plotted against analysis time and partial efficient score residuals were also plotted against analysis time to identify observations with disproportionate influence. Partial residuals were calculated for each observation within the subject. These are the additive contributions to a subject's overall residual (see [26, 27] for details). The partial DFBETA value estimates the change in the regressor's coefficient due to deletion of that individual record.

Although the primary aim of this study is to describe methods to identify outliers, we provide some guidelines on how to handle potential outliers. The steps below were followed in this study:

(i) The full dataset was analyzed with the presence of these outliers for sensitivity analysis.
(ii) The variable was log-transformed to base 2 to avoid dropping observations.
(iii) Removed outliers based on IQR scale factors $k = 3, 5, 7, 10$.

Detected outliers for $k = 3$ were inversely weighted so that they can lie within the lower and upper bounds using the following procedure: For each x_i value that is outside the limits, compute weight $w_i = \max(|x_i\text{-upper limit}|, |x_i\text{-lower limit}|)$. Calculate a pseudo value $x_j = \frac{x_i}{w_i} + v$, where v>0 is any constant (such as lower limit, Q3, etc.) such that x_i outlier is replaced with x_j that is bounded within the limits. An alternative is to replace x_i with the group-specific quartile value such as the median or a combination of quartiles $(Q1 + Q2 + Q3)$. Statistical analysis was conducted using SAS® software version 9.4 [31], R Core Team (2023) version 4.3.1 [32] and Stata statistical software (release 18) [33].

RESULTS

During the first 10 years of follow-up, 778 out of 8501 children developed persistent confirmed IA. The median age (IQR) at risk was 36.3 (18.1, 72.5) months. Of the 778 children with IA, 590 had

Table 1. Demographics of study participants and the risk of getting Islet Autoimmunity where subjects are followed up for the first 10 years of life.

Variable	Total subjects followed from enrollment at age 3 months ($N = 8501$)	Developed IA by age 10 years ($N = 590$)
Incidence rate (per 1000 person/ years); 95% CI		0.042 (0.039, 0.045)
IA age at risk month; median (Q1, Q3)		36.3 (18.1, 72.5)
Country		
USA	3616 (42.5%)	207 (35.1%)
Finland	1803 (21.2%)	139 (23.5%)
Germany	580 (6.8%)	30 (5.1%)
Sweden	2502 (29.4%)	214 (36.3%)
Family history with T1D		
No	7540 (88.7%)	477 (80.9%)
Yes	961 (11.3%)	113 (19.1%)
Human Leukocyte Antigen (HLA) genotype DR3/4		
No	5178 (60.9%)	299 (50.7%)
Yes	3323 (39.1%)	291 (49.3%)
Sex		
Male	4305 (50.6%)	336 (56.9%)
Female	4196 (49.4%)	254 (43.1%)

complete records. The incidence rate of Islet Autoimmunity was estimated to be 0.042 with 95% CI: (0.039, 0.045).

Descriptive statistics of the demographics are displayed by country, FDR status, HLA DR3/4 status and sex as shown in Table 1 with the number of subjects at enrollment (age 3 months), number of person-years follow-up, number of children developing IA with incidence, that were analyzed in the Cox model. Supplementary Table S1 displays additional descriptive statistics of the distribution of vitamin B_{12} intake including interquartile range (IQR), median, standard deviations and standard errors by country for several datasets used.

The trend of vitamin B_{12} levels (ug) by country and visit from 8501 children in the TEDDY cohort followed regularly till 10 years of age while at risk of developing Islet Autoimmunity is shown in Fig. 1, which shows that the intake/day of vitamin B_{12} varied on average between countries and by visit with the intake being consistently higher in the USA and lowest in Germany. A box plot of the vitamin B_{12} intakes (Supplementary Fig. S1) shows the presence of potential outliers at some of the follow-up visits.

Table 2 provides Cox regression estimates both in the log scale (log-hazard ratios) and in the exponential form (HR) together with the sandwich robust SE and the 95% CI. The HR represents the risk associated with a unit standard deviation (SD) change in take. The two outstanding extreme observations (vitamin B_{12} of 670.81 and 1666.83 µg) were scrutinized further by dropping each one of them in turn and re-analyzing the data, then later, dropping both values and repeated the analysis, and then replaced them with their medians calculated by country and visit and re-analyzed the data. Results indicated that dropping an observation from the full data with vitamin B_{12} intake of 670.81 µg/day at the 6th year visit (visit 72 months) from the USA had the largest impact on both the standard errors and the direction of the effect. It was noted that this observation was a confirmed Islet Autoimmunity positive case. The other outlier with vitamin B_{12} intake value of 1666.83 µg/day was not an Islet Autoimmunity positive case and had minimal

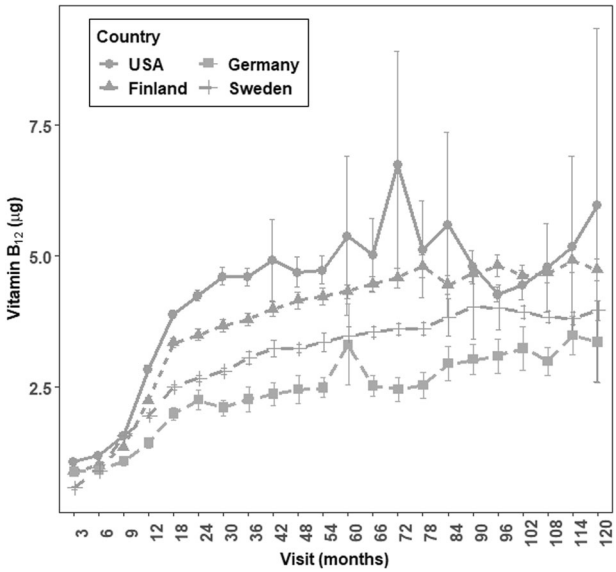

Fig. 1 Mean intake by country and visit. Line graph of average vitamin B_{12} (µg/day) intake by country in first 10 years of life.

influence on the Cox regression model as shown in Table 2 under the sensitivity analysis sub-title.

After fitting the extended Cox regression models, residual diagnostics plots were examined for highly influential observations on the model parameters and are given in Figs. 2–4. Figure 2 displays eight residual diagnostics plots from panels (a)–(h) to find out any extreme observations that are influential. Figure 2a, b shows standardized DFBETA residuals against vitamin B_{12} intake for full data and IQR-5 data. Extreme observations are noticeable from the full data. Similarly, Fig. 2c, d displays martingale residuals with potential outliers seen on the plot of full data. Deviance residual plots are provided in Fig. 2e, f while score residuals are shown in Fig. 2g, h for models fitted using the full data and IQR-5 datasets. All plots from the full data indicated the presence of at least two observations that have larger residuals than expected.

Figures 3 and 4 display partial DFBETA residuals by country for data analyzed based on the full data and IQR-5 data, respectively. We compared the magnitudes of the largest DFBETA values to the Cox regression coefficients and labeled the points by their vitamin B_{12} intake values. Results from Fig. 3 showed that the USA's vitamin B_{12} intake/day values of 670.81 µg at visit 72 months and 1666.83 µg at visit 120 months were disproportionately influential for the data from the full data but not from the IQR-5 dataset (Fig. 4).

Similar patterns were observed when looking at the partial deviance residuals (Supplementary Figs. S2 and S3), and partial score residuals plots (Supplementary Figs. S5 and S6) by country. The goodness of fit graphs based on Cox-Snell residuals plotted against cumulative hazard for modeling the full data, IQR-5 data and $\log_2$-transformed data are provided in the Supplementary files (Supplementary Figs. S4a–c, respectively). There are light-heavy tails, but the models fitted the data reasonably well.

DISCUSSION

The study has provided practical illustrations where IQR method can be used to identify potential outliers in longitudinal datasets by follow-up visits.

Analysis of TEDDY dietary data following the IQR method showed the impact of extreme observations on the model (Table 2). Results from the full data analysis indicated a significant increase in risk of developing IA with higher intakes of vitamin B_{12}. Analysis from the log-transformed and IQR-k-reduced datasets showed that an increase in intake of vitamin B_{12} reduces the risk of IA although the association

Table 2. Association of vitamin B_{12} (μg/day) intake/day with risk of Islet Autoimmunity (IA) from TEDDY dietary data.

Dataset (obs.)	Vitamin B_{12} obs. dropped, *n* (%)	*β* (SE)	HR (95% CI)	*P* value
Full ($n = 152,426$)	0 (0%)	0.024 (0.009)	1.025 (1.006, 1.044)	0.010
IQR-k = 10 ($n = 152,426$)	77 (0.05%)	−0.030 (0.044)	0.971 (0.891, 1.057)	0.496
IQR-k = 7 ($n = 152,426$)	134 (0.09%)	−0.023 (0.044)	0.977 (0.897, 1.065)	0.596
IQR-k = 5 ($n = 152,400$)	243 (0.16%)	−0.013 (0.044)	0.987 (0.905, 1.077)	0.773
IQR-k = 3 ($n = 152,217$)	811 (0.53%)	−0.006 (0.045)	0.994 (0.910, 1.086)	0.897
IQR-k = 3 wtd ($n = 152,217$)	0 (0%)	−0.004 (0.045)	0.996 (0.912, 1.088)	0.930
Log2 transformed ($n = 152,426$)	0 (0%)	−0.019 (0.054)	0.982 (0.884, 1.09)	0.728
Sensitivity analysis after residual diagnostics				
Full data-case extreme obs. ($n = 152,425$)[a]	1 (0 %)	−0.240 (0.211)	0.787 (0.52, 1.189)	0.255
Full-non-case with extreme obs. ($n = 152,425$)[b]	1 (0 %)	0.035 (0.012)	1.035 (1.011, 1.06)	0.004
Full less two extremes obs. ($n = 152,424$)[c]	2 (0 %)	−0.215 (0.189)	0.807 (0.557, 1.168)	0.255
Full data extremes replaced with medians ($n = 152,426$)[d]	0 (0 %)	−0.252 (0.215)	0.777 (0.509, 1.185)	0.242

Hazard ratios (HR) and 95% confidence intervals (CI) represent the risk associated with a unit standard deviation (SD) change in intake. The estimate (*β*) is the log-hazard ratio shown with robust standard errors (SE). Several datasets were used under different IQR-K conditions.
For sensitivity analysis: [a]Vitamin $B_{12} = 670.81$ μg from visit 72 months in the USA was dropped from the full data. [b]Vitamin $B_{12} = 1666.83$ μg value from visit 120 months from the USA was dropped from the full data. [c]Both values of vitamin B_{12} were dropped (670.81 and 1666.83 μg). [d]The two extreme values were replaced with their country and visit specific median values.

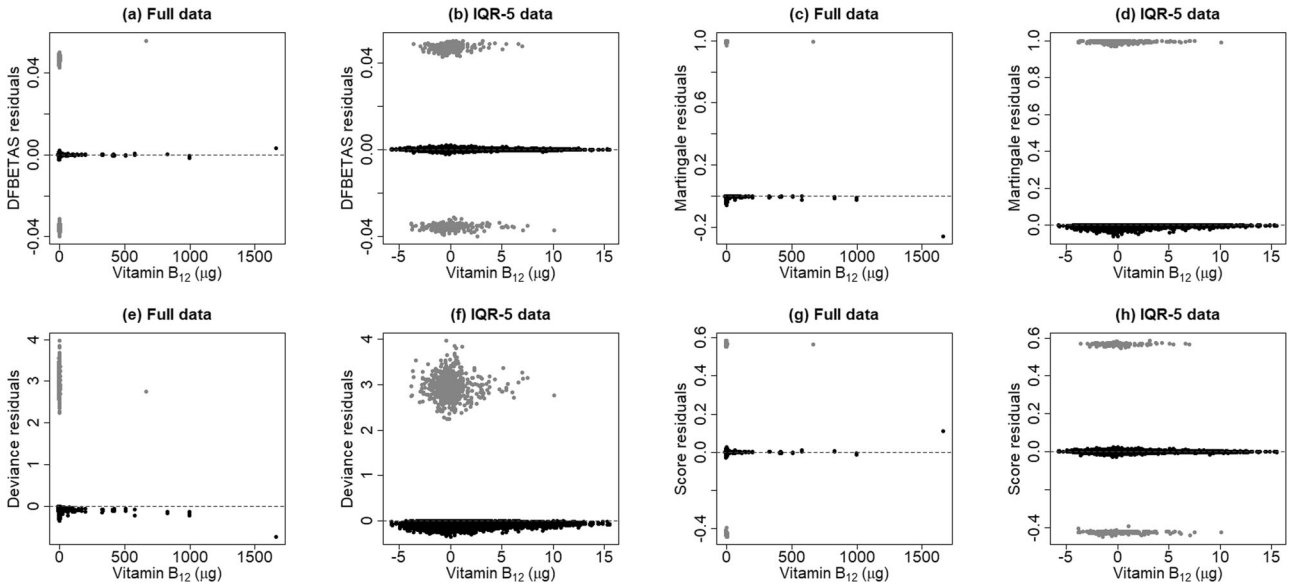

Fig. 2 Residuals diagnostics plots from fitted models for full and IQR-5 TEDDY data where residuals are plotted against average vitamin B_{12} (μg/day) intake/day. a, **b** The standardized DFBETA residuals, **c**, **d** the martingales residuals, **e**, **f** deviance residuals and **g**, **h** score residuals against the vitamin B_{12} intake. Extreme observations that appear to be further away from other points are clearly seen on the full data plots.

between exposure and outcome in the Cox model was not statistically significant. Some of the detected outliers by the IQR method were also found to be highly influential observations by the residual diagnostic plots. The most influential extreme observation was found to be a case where the unusual large value of vitamin B_{12} of 670.81 μg was on the 6th year visit in the USA when the subject was diagnosed with Islet Autoimmunity. Results in Table 2 show the impact of this outlier on the model. When this unusual observation is removed, the time-to-event analysis handles this subject as a non-case up to the last visit that appears in the dataset when data is arranged in the counting process format. As a result, the direction of the HR, the SE and the significance of the association changed drastically.

In many studies, researchers would not want to drop observations. We have illustrated the procedure of not having to drop outliers using these two extreme observations (vitamin B_{12} of 670 and 1666.83 μg) where we replaced them with their group-specific medians, and in using the log transformation and inverse weighting methods (section "Detecting highly influential observations using residual diagnostics"). This made the 1 outlier out of the 590 cases remain as a case in the survival model with an intake value that was within the group range. Results for this analysis were seen to be in the same direction as for the IQR-k methods and $\log_2$-transformed method, showing a reduced risk of IA with intake of vitamin B_{12} and larger standard errors (0.216) than when the outlier are not handled (0.009).

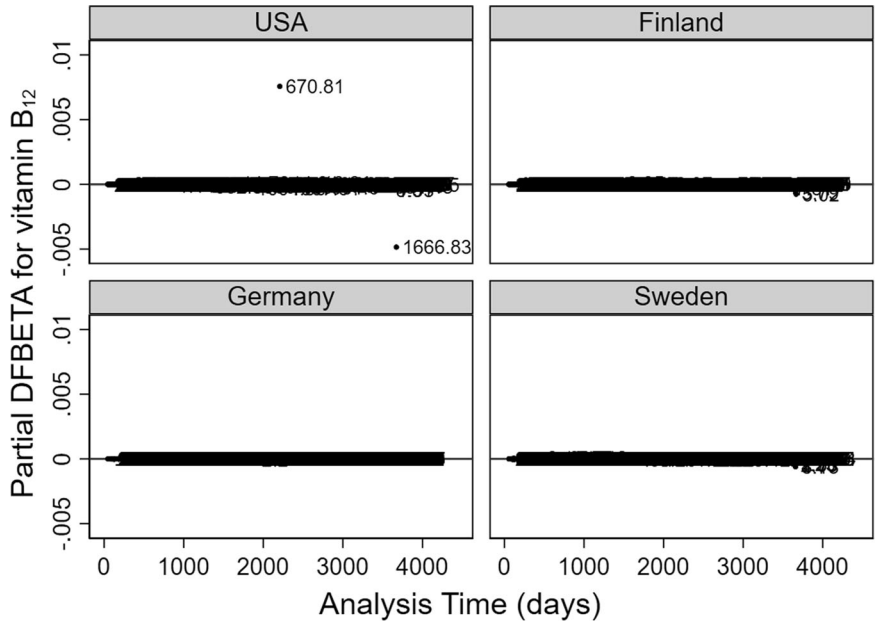

Fig. 3 Partial DFBETA residuals from full TEDDY dietary data showing vitamin B_{12} (μg/day) values that have larger or unusual partial DFBETA residuals compared to other data points. The extreme observations from the full data come from the USA in visits 72 months for the 670.81 μg/day and visit 120 months for the 1666.83 μg/day intakes indicate to have a disproportionate influence on the fitted model.

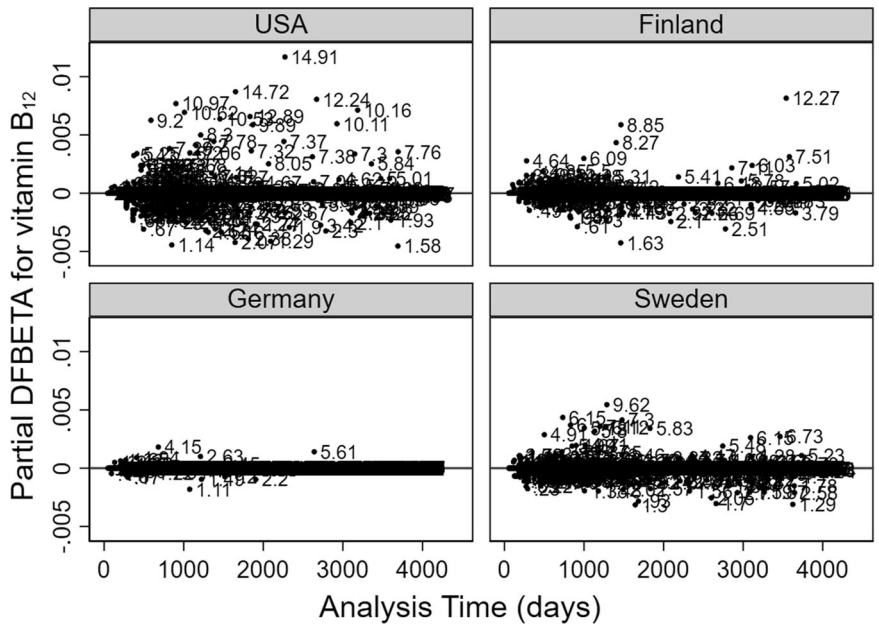

Fig. 4 Partial DFBETA residuals from IQR-5 TEDDY data with labeled vitamin B_{12} (μg/day) values. There are no obvious extreme values shown on the residual plot.

The extreme value of vitamin $B_{12} = 1666.83$ μg was found to be on the 10th year visit in the USA and although it is such an extremely large value compared to the specific country and visit values, this observation had mild impact on the survival model since it was not a case. When only this outlier is removed, in the presence of the other case-outlier, the HR still indicated an increased risk of IA with intake of vitamin B_{12}.

Our study has shown that 1 outlier out of 590 cases of IA (Table 1) has a different impact on the model compared to 1 outlier out of 7911 (8501–590) non-cases. We could have had, say, 40 outliers out of 590 cases compared to 40 outliers out of 7911 non-cases. IQR-k algorithm can identify these potential outliers that could represent a genuine subgroup related to disease/case status. In survival regression models, cases with extreme values can be highly influential compared to non-cases.

Choices for the IQR scale factor *k* may depend on the distribution of the data and how far away from the median the researchers would like to keep the data points. Smaller *k* values are more stringent and make the distribution of the variable to be more precise than bigger *k* values which give room for larger variances. If no prior information on the distribution of the variable of interest is available, the analyst can examine several *k* values to identify potential outliers to flag them off then conduct sensitivity analyses to see if the results from the model parameters (estimate, standard errors, strength of association) change with or without the detected outliers.

Examining residual diagnostic plots for the full models compared to those from the IQR-5 models revealed similar patterns of the presence of potential outliers that could be highly influential in the full models but not in the IQR-5 models.

We have illustrated the use of IQR method to detect potential outliers of time-varying continuous variables in longitudinal datasets at follow-up visits. It can be used as a data quality control procedure to identify unusual observations. Once outliers are identified, they can be flagged off and investigated further, including conducting sensitivity analysis to ascertain their influence in the regression model.

DATA AVAILABILITY

Data from The Environmental Determinants of Diabetes in the Young (https://doi.org/10.58020/y3jk-x087) reported here will be made available for request at the NIDDK Central Repository (NIDDK-CR) website, Resources for Research (R4R), https://repository.niddk.nih.gov/.

CODE AVAILABILITY

Statistical analysis code can be provided by the corresponding author upon a reasonable request.

REFERENCES

1. Agresti A, Franklin CA, Klingenberg B. Statistics: the art and science of learning from data. 5th ed. Pearson; Essex, England; 2021.
2. McClave JT, Sincich TT. Statistics. 13th ed. Pearson Higher Ed; New Jersey, USA; 2017.
3. Aguinis H, Gottfredson RK, Joo H. Best-practice recommendations for defining, identifying, and handling outliers. Organ Res Methods. 2013;16:270–301.
4. Jones PR. A note on detecting statistical outliers in psychophysical data. Attention, perception, and psychophysics. Vol. 81. Springer New York LLC; New York, USA, 2019. p. 1189–96.
5. Leys C, Delacre M, Mora YL, Lakens D, Ley C. How to classify, detect, and manage univariate and multivariate outliers, with emphasis on pre-registration. Int Rev Soc Psychol. 2019;32:5.
6. Van den Broeck J, Cunningham SA, Eeckels R, Herbst K. Data cleaning: detecting, diagnosing, and editing data abnormalities. PLoS Med. 2005;2:966–70.
7. Stasinopoulos MD, Rigby RA, Heller GZ, Voudouris V, Bastiani F De. Flexible regression and smoothing using GAMLSS in R. CRC Press; Boca Raton, FL, USA. 2017.
8. Rigby RA, Stasinopoulos MD, Heller GZ, Bastiani F De. Distributions for modeling location, scale, and shape: using GAMLSS in R. CRC Press; Boca Raton, FL, USA. 2020.
9. Yang J, Rahardja S, Fränti P. Outlier detection: how to threshold outlier scores? In: ACM International Conference Proceeding Series. Association for Computing Machinery; New York, USA, 2019.
10. Van der Meer T, Te Grotenhuis M, Pelzer B. Influential cases in multilevel modeling: a methodological comment. Am Socio Rev. 2010;75:173–8.
11. Yang S, Hutcheon JA. Identifying outliers and implausible values in growth trajectory data. Ann Epidemiol. 2016;26:77–80.e2.
12. Leys C, Ley C, Klein O, Bernard P, Licata L. Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. J Exp Soc Psychol. 2013;49:764–6.
13. Phan HTT, Borca F, Cable D, Batchelor J, Davies JH, Ennis S. Automated data cleaning of paediatric anthropometric data from longitudinal electronic health records: protocol and application to a large patient cohort. Sci Rep. 2020;10:10164.
14. Shi J, Korsiak J, Roth DE. New approach for the identification of implausible values and outliers in longitudinal childhood anthropometric data. Ann Epidemiol. 2018;28:204–11.e3.
15. Dugravot A, Sabia S, Shipley MJ, Welch C, Kivimaki M, Singh-Manoux A. Detection of outliers due to participants' non-adherence to protocol in a longitudinal study of cognitive decline. PLoS One. 2015;10:e0132110.
16. Boone-Heinonen J, Tillotson CJ, O'Malley JP, Marino M, Andrea SB, Brickman A, et al. Not so implausible: impact of longitudinal assessment of implausible anthropometric measures on obesity prevalence and weight change in children and adolescents. Ann Epidemiol. 2019;31:69–74.e5.
17. Hazrati S, Hourigan SK, Waller A, Yui Y, Gilchrist N, Huddleston K, et al. Investigating the accuracy of parentally reported weights and lengths at 12 months of age as compared to measured weights and lengths in a longitudinal childhood genome study. BMJ Open. 2016;6:11653. https://doi.org/10.1136/bmjopen-2016-011653.
18. Farooqui T, Mustafa I, Christie T. Outliers in educational achievement data: their potential for the improvement of performance. Pak J Stat. 2014;30:71–82.
19. Voloh B, Watson MR, König S, Womelsdorf T. MAD saccade: statistically robust saccade threshold estimation via the median absolute deviation. J Eye Mov Res. 2019;12:1–12.
20. Chen Z, Song S, Wei Z, Fang J, Long J. Approximating median absolute deviation with bounded error. Proc VLDB Endow. 2021;14:2114–26. https://doi.org/10.14778/3476249.3476266.
21. Casella G, Berger RL. Statistical inference. 2nd ed. Duxbury; USA. 2002.
22. Rousseeuw PJ, Croux C. Explicit scale estimators with high breakdown point. In: Dodge Y, editor. L_1-Statistical analysis and related methods. Y. Dodge, Amsterdam; North-Holland; 1992. p. 77–92.
23. TEDDY Study Group. The Environmental Determinants of Diabetes in the Young (TEDDY) Study. Ann N Y Acad Sci. 2008;1150:1–13. https://doi.org/10.1196/annals.1447.062.
24. Uusitalo U, Kronberg-Kippila C, Aronsson CA, Schakel S, Schoen S, Mattisson I, et al. Food composition database harmonization for between-country comparisons of nutrient data in the TEDDY Study. J Food Compos Anal. 2011;24:494–505.
25. Cox DR. Regression models and life tables (with discussion). J R Stat Soc B 1972;74:187–220.
26. Klein JP, Moeschberger ML. Survival analysis: techniques for censored and truncated data. 2nd ed. Springer; New York, USA. 2003.
27. Hosmer DW, Lemeshow S, May S. Applied survival analysis: regression modeling of time-to-event data. 2nd ed. John Wiley & Sons, Inc.; New Jersey, USA; 2008.
28. Lin DY, Wei LJ. The robust inference for the cox proportional hazards model. J Am Stat Assoc. 1989;84:1074–8.
29. Zeger SL, Liang K-Y. Longitudinal data analysis for discrete and continuous outcomes. Biometrics. 1986;42:121–30.
30. Willett WC, Howe GR, Kushi LH. Adjustment for total energy intake in epidemiologic studies. Am J Clin Nutr. 1997;65:1220S–1228S. discussion 1229S–1231S.
31. SAS Institute Inc. SAS Software 9.4 (SAS/STAT 15.2). Cary, NC, USA; 2016. http://www.sas.com/.
32. R Core Team. R: A language and environment for statistical computing. Vienna, Austria; 2023. https://www.r-project.org/.
33. StataCorp LLC Stata Statistical Software. College Station, TX: StataCorp LLC; 2023.

ACKNOWLEDGEMENTS

The authors would like to thank Sarah Austin-Gonzalez of the University of South Florida (USF)-Health Informatics Institute for editing and providing study information and support. We would also like to thank the reviewers for helping us to improve on the manuscript.

AUTHOR CONTRIBUTIONS

Conceptualization: LKM and JY. Methodology: LKM, KFL, and XL. Software: LKM. Formal analysis: LKM. Resources: JPK. Data curation: LKM, JY, and UMU. Writing—original draft preparation: LKM. Writing—review and editing: LKM, XL, KFL, JY, CAA, SH, JMN, SMV, LH, UMU, and JPK. Supervision: KFL, XL, and JPK. Project administration: UMU and JMN. Funding acquisition: JMN, UMU, and JPK. All authors have read and agreed to the published version of the manuscript.

FUNDING

The TEDDY Study is funded by U01 DK63829, U01 DK63861, U01 DK63821, U01 DK63865, U01 DK63863, U01 DK63836, U01 DK63790, UC4 DK63829, UC4 DK63861, UC4 DK63821, UC4 DK63865, UC4 DK63863, UC4 DK63836, UC4 DK95300, UC4 DK100238, UC4 DK106955, UC4 DK112243, UC4 DK117483, U01 DK124166, U01 DK128847, and Contract No. HHSN267200700014C from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of Allergy and Infectious Diseases (NIAID), Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), National Institute of Environmental Health Sciences (NIEHS), Centers for Disease Control and Prevention (CDC), and JDRF. This work is supported in part by the NIH/NCATS Clinical and Translational Science Awards to the University of Florida (UL1 TR000064) and the University of Colorado (UL1 TR002535). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

COMPETING INTERESTS

The authors declare no competing interests.

ETHICAL APPROVAL

The TEDDY study was conducted in accordance with the Declaration of Helsinki, and approved by local US Institutional Review Boards and European Ethics Committee Boards, including the Colorado Multiple Institutional Review Board, Medical College of Georgia Human Assurance Committee (2004–2010), Georgia Health Sciences University Human Assurance Committee (2011–2012), Georgia Regents University Institutional Review Board (2013–2015), Augusta University Institutional Review Board (2015–present), University of Florida Health Center Institutional Review Board, Washington State Institutional Review Board (2004–2012), Western Institutional Review Board (2013–present), Ethics Committee of the Hospital District of Southwest Finland, Bayerischen Landesärztekammer (Bavarian Medical Association) Ethics Committee, Regional Ethics Board in Lund, Section 2 (2004–2012), and Lund University Committee for Continuing Ethical Review (2013–present).

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41430-023-01393-6.

Correspondence and requests for materials should be addressed to Lazarus K. Mramba.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.