**Criteria and procedure for data cleaning in the NCD-RisC database**

This document is the protocol for cleaning multiple variables in the NCD-RisC database, which has the following steps:

1. Applying univariate cleaning criteria
2. Applying multivariate cleaning criteria
3. Multivariate outlier detection

**Applying univariate cleaning criteria**

Data outside the following ranges are removed to minimise the possibility of entry error.

*Anthropometry (age-specific)*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Height (cm) | Weight (kg) | BMI (kg/m2) | Waist (cm) | Hip (cm) | Waist-to-hip ratio | Waist-to-height ratio |
| 5-9 years | 60-180 | 5-90 | 6-40 | 20-150 | 30-180 | 0.4-1.8 | 0.2-1.5 |
| 10-14 years | 80-200 | 8-150 | 8-60 | 20-200 | 30-200 | 0.4-1.8 | 0.2-1.5 |
| 15+ years | 100-250 | 12-300 | 10-80 | 30-300 | 40-300 | 0.4-2.0 | 0.2-2.0 |

BMI: body-mass index

*Blood pressure*

|  |  |  |  |
| --- | --- | --- | --- |
|  | SBP  (mmHg) | DBP  (mmHg) | Pulse (beats/s) |
| All ages | 70-270 | 30-150 | 30-220 |

SBP: systolic blood pressure; DBP: diastolic blood pressure

*Lipids*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| (mmol/L) | TC | LDL | HDL | Triglycerides |
| All ages | 1.75-20 | 0.5-10 | 0.4-5 | 0.2-20 |

TC: total cholesterol; LDL: low-density lipoprotein cholesterol; HDL: high-density lipoprotein cholesterol

*Glucose*

|  |  |  |  |
| --- | --- | --- | --- |
|  | FPG (mmol/L) | OGTT (mmol/L) | HbA1c (%) |
| All ages | 2-30 | 2-30 | 3-18 |

FPG: fasting plasma glucose; OGTT: 2hour oral glucose tolerance test

*Kidney function related variables*

|  |  |  |
| --- | --- | --- |
|  | Serum creatinine  (µmol/L) | eGFR  (mL/min/1.73m2) |
| All ages | 8-1800 | 5-250 |

eGFR: estimated glomerular filtration rate

**Applying multivariate cleaning criteria**

The following constraints are applied after applying univariate cleaning criteria listed above:

* SBP > DBP

If multiple measurements of blood pressure are taken, the average SBP and DBP are calculated by discarding the first measurement and averaging the remainders. The constraint is applied to the average SBP and DBP. The averaging should be done after excluding measurements outside of the ranges in the univariate cleaning criteria noted in the first section.

* TC > LDL
* TC > HDL
* TC – (LDL + HDL) ≥ margin of error

The “margin of error” is determined by using the Cholesterol Reference Method Laboratory Network permitted measurement error limits for TC (8.9%), HDL (13%) and LDL (12%) as follows:

* Calculate errors in worst case scenario, i.e. TC is underestimated and HDL/LDL overestimated, each by the largest error permitted;
* For simplification the limit of errors by quartiles of TC using US NHANES data:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| TC range (mmol/L) | <4.040 | 4.040 to <4.714 | 4.714 to <5.465 | ≥5.465 |
| Margin of error (mmol/L) | -0.7125 | -0.8821 | -0.9700 | -1.7520 |

**Multivariate outlier detection**

After applying both univariate and multivariate cleaning criteria, multivariate outlier detection is carried out on all pairs of variables within the same risk factor family (see lists in the first section). In each pair, a method based on Mahalanobis distance1 is used to calculate the distance of the data points to the centre of all available data, and those that lie outside a certain threshold will be removed (both variables) . Although called “outlier detection”, the threshold is chosen to catch only extreme values that are implausible. The points detected and removed should be examined in scatter plots.

*Data normalization*

Mahalanobis outlier detection method assumes the data to be Normally distributed. We suggest comparing the distribution of each variable to its log-transformed form, and pick the one closer to Normal distribution. For reference, based on the data in NCD-RisC database, all variables are log-transformed except height and DBP.

In case neither the variable nor its log transformation is close to be Normally distributed, alternative non-parametric methods should be used. Such methods are being tested as of March 2022 and will be added to this document once finalised.

*Pairs of variables considered*

All pairs of variables of interest from the same risk factor family are examined (see lists in the first section of this document). Exceptions are noted below.

* Some variables, e.g., FPG, OGTT, LDL and non-HDL cholesterol, are very skewed even after log-transformation, therefore we do not recommend cleaning any pair that includes any of them using the method based on Mahalanobis distance. Alternative outlier detection methods are currently being explored for NCD-RisC work.
* For anthropometrics variables, the procedure should be applied separately for children (5-9 years), adolescents (10-14 years), and adults (15+ years).
* eGFR is calculated from serum creatinine using a regression-based model (eg. CKD-EPI) hence multivariate cleaning is not needed.

*Cleaning process*

For all pairs that multivariate cleaning are applied to, the cleaning process is to:

* Compute the Mahalanobis of distance of all the data points after normalisation
* Exclude points with distance larger than 68.5 (quantile of chi2 distribution equivalent to 8SD for normal distribution)

Code and an example of the multivariate outlier detection procedure are available at

<https://github.com/NCD-RisC/Data-cleaning-protocol>

*References:*

1 Filzmoser, P. *A multivariate outlier detection method*. (Citeseer, 2004).