* Complete status of treatment (binary) (%) [Missing] 9 ( 0.0)
* Educational Attainment (%)[Missing] 384 ( 0.5)
* Primary Substance (admission to treatment) (%) Other 1585 ( 1.9)
* Primary Substance (admission to treatment) (%) [Missing] 1 ( 0.0)
* Frequency of Substance Use (Primary Substance) (%) 1 day a week or more 5610 ( 6.6)
* Frequency of Substance Use (Primary Substance) (%) Less than 1 day a week 4178 ( 4.9)
* Frequency of Substance Use (Primary Substance) (%) [Missing] 421 ( 0.5)
* ~~Corrected Occupational Status (f) (%) Looking for a job for the first time 193 ( 0.2)~~
* ~~Corrected Occupational Status (f) (%) Not seeking for work 832 ( 1.0)~~
* ~~Corrected Occupational Status (f) (%) [Missing] 1 ( 0.0)~~
* Number of Children (dichotomized) (%) [Missing] 745 ( 0.9)
* Housing Situation (Tenure Status) (%) Illegal Settlement 906 ( 1.1)
* Housing Situation (Tenure Status) (%) Others 2354 ( 2.8)
* Housing Situation (Tenure Status) (%) [Missing] 4679 ( 5.5)
* SUD Severity (Dependence status) (%) [Missing] 1 ( 0.0)
* Urbanicity (%) [Missing] 2 ( 0.0)
* Primary Substance (initial diagnosis) (%) Other 1987 ( 2.3)
* Primary Substance (initial diagnosis) (%) [Missing] 6421 ( 7.5)
* Cohabitation status (Recoded) (f) (%) [Missing] 1 ( 0.0)
* Biopsychosocial compromise (%) [Missing] 1557 ( 1.8)
* Treatment Admission Motive (%) Other 4514 ( 5.3)

For instance – under occupational status, could “inactive”, “not active” and “unemployed” be collapsed into one category? If not – see the comment under methods above – what these categories mean should be explained

**Employment Status\_corr24** n(%)=”Looking for a job for the first time" were changed for "Unemployed", "No activity" for "Inactive", and "Not seeking for work", for "Inactive".

**Ordered frequency of consumption**: 1. Less than 1 day a week 2. 1 day a week or more 3. 2 to 3 days a week 4. 4 to 6 days a week 5. Daily

**También saqué otras\_sus1\_mod, otras\_sus2\_mod, ya que eran las basales**

The imputation was based on chained random forest with predictive mean matching across all individuals with data at that age  
- Khera, R., Kondamudi, N., Liu, M., Ayers, C., Spatz, E. S., Rao, S., Essien, U. R., Powell-Wiley, T. M., Nasir, K., Das, S. R., Capers, Q., & Pandey, A. (2023). Lifetime healthcare expenses across demographic and cardiovascular risk groups: The application of a novel modeling strategy in a large multiethnic cohort study. *American Journal of Preventive Cardiology*, *14*, 100493. <https://doi.org/10.1016/j.ajpc.2023.100493>

In all analyses, missing values were imputed using a random forests-based algorithm (Mayer 2021)7F 8 , assuming that the data were missing at random. Appendix E reports the imputed values for each variable.

* Missing data for the predictor values (anger, embitterment, cognitive reappraisal, expressive suppression, presence of and search for meaning in life) were imputed using the package ‘missRanger’, which uses a chaining random forests algorithm to impute mixed-type data sets. To impute the missing data, 100 trees were calculated and predictive mean matching was applied. The proportion of missing data was 2.8% or less on an item level.
* Sheetal, A., Jiang, Z., & Di Milia, L. (2023). Using machine learning to analyze longitudinal data: A tutorial guide and best-practice recommendations for social science researchers. Applied Psychology, 72(3), 1339–1364. <https://doi.org/10.1111/apps.12435>
  + The goal of our article was to highlight the advantages of machine learning when drawing on large longitudinal samples—that is, data that does not meet the assumptions of parametric statistics, data that has missing data not at random, and data that is not equally collected across data waves.
  + they can handle collinearity.
* Masip, G., Foraita, R., Silventoinen, K. et al. The temporal relationship between parental concern of overeating and childhood obesity considering genetic susceptibility: longitudinal results from the IDEFICS/I.Family study. Int J Behav Nutr Phys Act 18, 139 (2021). https://doi.org/10.1186/s12966-021-01205-9
  + The random forest imputation is based on 200 trees and was controlled for a variation of additional predictors. A huge advantage of using random forests is that they produce a single imputed dataset, they are adaptive to interactions and nonlinear relationships not needing to specify an imputation model, and they can handle mixed types of missing data.
  + As overall missing for these variables were limited (6.08%), imputation was performed using the R package “missRanger” (v.2.1.0). Imputed variables have been additionally checked for plausibility through density plots.
* Hasyyati, Atika Nashirah and Lumley, Thomas. ‘Imputation for Sub-sampling in Indonesian National Socioeconomic Survey’. 1 Jan. 2022 : 1207 – 1217.
  + he number of variables selected at every node randomly (mtry) and number of trees grown in each forest (num.trees) were evaluated until mtry == 500 and num.trees == 500. These facts are in line with [[16](https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji220085#ref016)] that increasing mtry has a limited effect on imputation error, but computation time is strongly increased. Additionally, their research also found that changing the number of trees in the forest has a stagnating influence on imputation error but a strong influence on computation time which is approximately linear in the number of trees [[16](https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji220085#ref016)].