# Sample size estimation and power analysis

**Elucidating social patterning of chronic diseases using biological mechanisms in LACs settings: results from a nationally representative study and clinical cohorts in Colombia.**

Study: Investigating the link between SEP, biological mechanisms (allostatic load) and CVD, DM and multimorbidity using data from clinical cohorts.

The objective of a sample is to achieve that our estimates are weighted to represent both populations and increases statistical power. Determining the optimal sample size for a study assures an adequate power to detect statistical significance (Tenny & Abdelgawad, 2021). The procedures to estimate sample sizes to the paper 3 as described in detail as follows.

**Null hypothesis:** There is no association between lower SEP, adult allostatic load, and selected chronic conditions. The hypothesis test will be carried out with a reliability of 90%, alpha of 10% and power of the test, for a value in the parameter space of the alternate hypothesis of 0.36%, the 90%, that is a type II error of 10%.

**Observation and analysis units.**The study only has one unit of observation and analysis: individuals with biological sampling data in the CAC’s clinical cohorts.

**Determination of sample size.** This study will consider as a tracer indicator the prevalence of cardiometabolic diseases in the population of Colombia (The latest prevalence 2019 = 9,08% by High Cost-Account Fund’s report) to determine the sample size. Assuming an expected proportion of 0.07, a design effect of 1.2, a relative standard error of no more than 15% and a non-response percentage of 20% were used to achieve greater accuracy in estimating this indicator. In addition, we will follow the simple random sampling formulas proposed by Leslie Kish (Kish, 1995) to ensure the necessary sample size, is given by the following equation:



Where,

**** = Size of the population with sampling data for which the indicator is to be estimated at the minimum level of disaggregation. In this case, in a population of 9,993,810 individuals reported with one or more chronic conditions in Colombia.

**** = Minimum expected proportion of the indicator.

******

**** = Effect of sample clustering. This is defined as the relationship between the estimate of variation from a cluster sampling design and the estimate of variation from a simple random sampling (1,5 as standard value).

**** = Desired Relative Standard Error. Level of intended accuracy for the study. These values are small to the UK and Colombia population data. Therefore, the level standard is 0,01 (1%).

**** = Percentage of oversampling to ensure coverage of the study. In this case, an oversample of 20% is proposed (Cob = 1.20).

According to the preliminary application of this formula, we obtain a sample of 136,180 individuals.

However, during the study pilot, we will adopt the approach to repeat the sample-size estimates for each outcome and then select the largest number as the sample size required to answer all the research questions. This will be implemented with the purpose of adjusting the study design as needed. As a result, the individuals’ numbers included may be changed in the final study protocol.

Replication and adjustment materials of the study will be available here: https://github.com/juanrivillas/Early-life-inequalities-and-biological-aging-Colombia-

Corresponding author's contact details:

Juan Carlos Rivillas

Research Postgraduate

School of Public Health,

Imperial College London,

Sir Alexander Fleming Building,

South Kensington Campus

London SW7 2AZ

[j.rivillas-garcia20@imperial.ac.uk](mailto:j.rivillas-garcia20@imperial.ac.uk)