**CONSTRUCTION OF THE DATA SET**

Here we set out the mathematical formulae to construct the data sets:

1. We fix the total number of subjects for each set at 10,000.
2. Simulate baseline values () for each of the biological parameters to form part of the future scoring system, bearing in mind that each parameter will have a determined number of decimals and that it should also lie within a determined range of values. The distributions of each of these parameters can be seen in Table 1.

Table 1: Distributions of the simulated parameters for the baseline situation, according to data from the Puras-GEVA study.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Formula | Decimal places | Range |
| Age (years) |  | 0 | 18-89 |
| SBP (mmHg) |  | 0 | 85-220 |
| HbA1c (%) |  | 1 | 3.9-10.9 |
| Atherogenic index |  | 2 | 1.92-10.84 |
| Male gender |  | 0 | 0-1 |
| Smoker |  | 0 | 0-1 |

Abbreviations: SBP, systolic blood pressure; HbA1c, glycated haemoglobin.

1. Simulate our main variable (time-to-cardiovascular disease):
   1. Determine the follow-up time (in days), which we have defined to have a distribution , establishing a range from 90 days (so that each subject has at least 2 measurements in the construction sample) and 4 years (end of study).
   2. Constant baseline hazard per 1 unit time (1 day): 0.0000125.
   3. Using the following Cox regression model, [systolic blood pressure (SBP), glycated haemoglobin (HbA1c)], simulate the probability of event .
   4. Simulate . If consider that the patient developed a cardiovascular disease; if not, consider no cardiovascular disease.
2. Simulate the repeated measures during the follow-up () for the patients in the construction sample. To do this we must consider that the data will be simulated for days until is greater than the follow-up time for each patient. For each parameter (SBP, HbA1c and atherogenic index) we shall determine the actual value at as the previous value plus a random quantity that will have two parts: the first part will depend on the variability of the biological parameter itself (inter) and the second part will depend on the variability of the parameter in each individual subject (intra). Table 2 shows the distributions used to obtain our simulations.

Table 2: Distributions of the simulated parameters for the follow-up, according to data from the Puras-GEVA study.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Inter | Intra | Range |
| SBP (mmHg) |  |  | 85-220 |
| HbA1c (%) |  |  | 3.9-10.9 |
| Atherogenic index |  |  | 1.92-10.84 |

This procedure is analogous in the statistical validation sample via simulation, except that the time will have a maximum of 5 years and the probability that the patient will fail to attend the visit, so we will not have the relevant values recorded in the database. Nevertheless, the value will be calculated in order to obtain the next simulation value in time. The probability of having a measurement was considered with a distribution with a range .