**Normalization Method:**

The standard normalization method for this sort of task is to generate a static distribution for every possible cohort for every possible variable. For example, you might stratify cholesterol levels by gender, by ethnicity, and by age brackets that have been selected to best represent different stages of life with relation to cholesterol development. For the purposes of Lab100, this represents a very rigid and labor intensive way of doing things, while simultaneously having the possibility of not generating the most representative cohort. To this end, we developed a novel method of dynamically generating a cohort with which to compare your data to demographically similar individuals. Note that this system is separate from how we determine what is \*healthy\* for a given patient - a patient could be average for the population, but still be unhealthy if the population is, on average, unhealthy. Guidelines as to what is healthy or not are based off of medical literature and official guidelines.

The basic idea is that we want to compare the patient to everyone we have data for that is statistically similar to that patient. For categorical variables, this is easy - match gender to gender, and ethnicity to ethnicity. For quantitative variables, this is a bit harder. To find an appropriate cohort for the patient, we want to divide the population up into as many subdivisions as is feasible, and perform a Kolmogorov–Smirnov test between the subdivision of the subject and all other subdivisions, and form a cohort consisting of the contiguous group that is not statistically different for the chosen variable. In detail:

1. The knowledge base receives a request to get a population distribution for a variable for which we have distribution data, such as height.
2. The KB calls the normalization function, with the patient’s height and demographics as input. Note that demographics typically include ethnicity, race, and age. Currently, only the NHANES dataset is used to generate cohort based population statistics (as opposed to generating a normal distribution from known means and standard deviations). This means that if you do not use the above demographic data, you will not get a representative sample, and that certain demographics are severely underrepresented, and certain populations are overrepresented compared to the population proportions in the general population. Ideally, this dataset would expand to include a large number of individuals via lab100 that we could use to generate a sample that is potentially more representative of our patient population.
3. The normalization function start with a cohort of the entire available population.
4. It then removes all individuals who do not match the patient in one of the indicated categorical variables, like gender or ethnicity.
5. It then subdivides the population into subsets - default of 50 subsets, equally spaced. Ex 18-20, 21-23, 24-26.
6. Subsets must be a minimum size (default 50 individuals), so, starting with the subset at the top of the range, subsets are grouped together until they have a large enough range to represent a significant sample. For example, if age ranges 60-62 and 63-65 have only 30 individuals each, they would be combined into the age range 60-65 with 60 individuals, so that a statistical comparison of the subset with another subset would have enough statistical power. If a given sub population is very poorly represented in the data, this can result in very large ranges, which may indicate that you cannot accurately provide cohort data for the given demographic.
7. Each subset is compared with the subset that includes the patient. For example, if the patient is 40, the 39-41 age range would be compared with 42-45 and 36-38 age ranges using a Kolmogorov–Smirnov test.
8. The subsets which are statistically different (default pvalue = 0.1) are discarded. The resulting patient cohort is the group of all remaining subsets which are contiguous, which is to say that the range does not have any gaps from ranges that were discarded. Ie, if the patient is 40, and the age ranges 29-31, 36-38, 42-45, and 49-51 were found to not be statistically different, the sample range would only include individuals in the range 36-45.
9. The normalization returns the given range to the kb, where it will supply the appropriate values for generating estimated population statistics.