Why matching in test works – intuition:

It combines information for multiple strata’s with weighted average – it can be misleading when there is big difference between strata’s. But here I will demonstrate why it average the performance from different stratas.

For simplicity let’s assume we have 2 groups of eGFR : 1,2. For each group we will mark #cases as Ti for eGFR group i and #controls as Fi for eGFR group i. i ={1,2}.

We know that after matching:

========================================================================

Let’s focus on calculation of AUC which is also the probability that a random i.i.d case will be ranked higher then a random i.i.d control. We will mark as the AUC for eGFR group i.

**It doesn’t work for AUC and it can change the AUC to be any number between 0-1.**

**It does work for measurement like log loss/mse** or for example measuring treatment effect (average gain for each patient if treat and not treated). In some measurement matching is important, for example PPV in certain cutoff depends in the incidence rate, so is make sense to match by outcome probability before calculating the measure that depends in the outcome probability and combing them from 2 or more groups.

Let’s focus on mse for simplicity and assume the mse in the 2 distinct groups is similar (otherwise mixing them doesn’t make sense) or at least in the same direction (if you want to measure treatment effect). The advantage is bigger sample size which if combination of both groups (or multiple groups).

The are n samples that belongs to 2 distinct groups (low/high eGFR can be more than 2 distinct groups). G1, G2 are the group indices of each distinct group. We want to measure the MSE on the combined group. is the MSE on group i.

MSE:=

Which is “average” of the MSE on each group based on the group sizes. If the MSE on each group is similar – you gain a bigger sample size. If each group is affected by the incidence rate (the probability by be true case in the group, in domains without time dependency it is called prior), matching make sense to measure the effect on both/multiple groups.