Hi Yan,

Huma has been working on a project to identify potentially undiagnosed cases of TSC (Tuberous Sclerosis Complex), a rare oncology disease. I’m attaching the approved scientific protocol. The data is linked HES and CPRD. We used various SVMs for the modelling. Huma has almost finished the project now.

We need someone to 1) QC her work and 2) generate more potentially undiagnosed cases (she has identified ~2-5 cases from 100,000 patients but we would like to process another 100,000 patients). I was wondering if someone from your team could help? We would use the sandbox code for PSA purposes.

Broadly speaking, the project would involve:

. 1) QC / use Huma’s code to generate an additional 100k patients (we have 1m in total but only have the processing power to handle 100k at a time). This would involve SAS using the London servers (the data cannot leave the servers).

. 2) QC / use Huma’s R code to create the patient-level feature vector.

. 3) QC / use Huma’s R code to run the SVMs to A) verify the results and B) identify additional cases from the new 100k patients.

We will provide further detail but this is a broad outline. I think it will probably take ~3 weeks for someone in your team to do. The code will not be ready for perhaps another couple of weeks (and then perhaps just the initial SAS bit), but I wanted to give you the heads-up about this project. Is this something you would be happy to take on?

John