1. Fixed effects small t (effectiveness, predictive power, consistency)

I have not found any relevant literature on this matter… Did you have any luck by chance?

1. NAN: non-random decision? (They might have symptoms and take a test, previous round was false positive, etc)

Given that we have quite a large data set, and the apprehensions we discussed (due to the missing values not being non-random), we chose to only consider individuals with consecutive tests, and avoid NANs in that way.

I do have an additional question about NANs. In the MISCAN-Colon model, you have all information, except the current haemoglobin value of an individual (which we eventually want to predict). Outside of MISCAN, we have all information, except the stage of an individual (healthy/adenoma/cancer), unless an individual has had a colonoscopy, in which case we know everything. The problem is, however, that the stage value outside of MISCAN is NAN for 99.6% of observations.

The trade-off here is as follows: we know that the current stage in MISCAN is a really influential predictor for the haemoglobin value, and therefore not including this variable *outside* of MISCAN would (almost) definitely result in an omitted variable bias. On the other hand, if we do choose to include this variable outside of MISCAN, we’d include a variable which is almost entirely imputed, and therefore we have to ask ourselves how much sense does it make to include such a variable?

I considered the option of not including the variable outside of MISCAN, while still including it in MISCAN, but this gives rise to the following problem. We use our analysis outside of MISCAN to determine which model we should implement into MISCAN (i.e., the model with the highest performance). We already have to make an assumption here that the relative model comparison of MISCAN is equal (or at least similar to) the relative model performance inside of MISCAN. However, I don’t think this assumption is fair when we use a different set of regressors outside of MISCAN compared to inside of MISCAN.

I also thought about changing the distribution of NAN vs not NAN in the data set, to make for a smaller percentage of imputed observations. However, even if we’d settle for 95/5 imputed/original, we’d be left with ±500k of 6m observations, so this’d be a huge sacrifice, and we’d still have a variable which is 95% imputed.

This also comes down to one of our discussion points of last time we talked:

1. How do your methods handle NAN? (RT methods can handle missing data quite well)

Short answer: they practically don’t to my current knowledge. So, it’s widely advised against to include any missing variables. SVRs can handle them a bit better, but still poorly (I elaborate a bit more on this in my proposal).

To summarize, we have three options. (1) Exclude current stage both outside of MISCAN and inside of MISCAN, which is undesirable because this would almost for sure deteriorate performance and is likely to induce a bias in the estimates. Or, (2) we include stage both outside and inside of MISCAN, with the knowledge that the variable is nearly entirely imputed. Or, (3) we sacrifice data points, and bring our data set to a more desirable distribution of NAN vs not NAN (my least favourite).

My supervisors at EMC stated that their preference was to include the imputed variable, but I simply don’t know how ‘econometrically valid’ this choice is. That said, to impute stage I will include two additional data sets (one simulated in MISCAN, and one from 2014), so we do have some extra information we can use for imputing.

1. For the purpose of the practical application of the data: How well can NN’s and SVR’s actually account for the zero inflation of the data? (Always predict zero obviously will decrease MSE)

I haven’t found much on the appropriateness of my models on zero inflated data. I only found one paper which uses neural networks on zero inflated data, where ANN outperformed Poisson regression, negative binomial regression (NBR), zero-inflated Poisson regression, and zero-inflated negative binomial regression.

<https://pdfs.semanticscholar.org/fe7d/4a41e18350b5db624b29c30fb2b3b7dece20.pdf?_ga=2.94970911.2025489364.1655123634-1923869308.1655123634>

For now, I intend to just try and see what happens. If any method fails, I will incorporate the tweedie XGBoost model we’ve discussed.

1. Including additional variables
   1. Number of previous rounds (participated in)

I include the current FIT round as one-hot-encoded binary variable. So FIT1, FIT2, FIT3, FIT4, where FIT4==1 and all other dummies ==0 in round 4. I think this is sufficient and including the explicit number of current rounds might be overkill. Besides, a variable which says the explicit number of current rounds would be categorical and would therefore have to be one-hot-encoded, which would bring us back to what I already included.

* 1. Transformations of variables / Interaction between variables

Reinoud wrote a paper on risk stratification using MISCAN and found that including the variables in table 2 were significant in likelihood ratio tets. I could include them to the model, but I do doubt that it’s worth it, given that these numbers were acquired on a different data set, with less rounds and less observations, and he notes that a similar AUC could be attained with less included variables. Also, Danica tried to include some extra variable interactions, but found little to no difference in performance.

Table

Description automatically generated

A picture containing table

Description automatically generated