1. Are there varying/unequal time intervals in your data? If yes, how do you plan to take this into account? Include previous time intervals as independent variables? Is the current time interval known for prediction of hemoglobin levels? i.e. can the time since the last measurement be used, or just the time since the last and second last measurement?

This is a really good point, thank you for bringing this up. The data is structured as follows:

* **Round 1**: every individual in the data set receives an invitation to participate every two years from the year that you are set to turn 55, every individual who accepted the invitation receives a FIT at home
  + Decline 🡪 NA
  + Accept invitation 🡪
    - Hb level > 47 µg Hb/g 🡪 Colonoscopy (individual leaves data set)
    - Hb level <= 47 µg Hb/g 🡪 Round 2
* **Round 2**: every individual without indication for cancer in round one (either due to low Hb or due to no-show), is asked to participate in the following round (2 years after round 1)
  + Decline 🡪 NA
  + Accept invitation 🡪
    - Hb level > 47 µg Hb/g 🡪 Colonoscopy (individual leaves data set)
    - Hb level <= 47 µg Hb/g 🡪 Round 3
* **Round 3**: every individual without indication for cancer in round one **and** two, is asked to participate in the following round (2 years after round 2)
  + Decline 🡪 NA
  + Accept invitation 🡪
    - Hb level > 47 µg Hb/g 🡪 Colonoscopy (individual leaves data set)
    - Hb level <= 47 µg Hb/g 🡪 Round 4
* **Round 4**: every individual without indication for cancer in round one, two **and** three, is asked to participate in the following round (2 years after round 3)
  + Decline 🡪 NA
  + Accept invitation 🡪
    - Hb level > 47 µg Hb/g 🡪 Colonoscopy (individual leaves data set)
    - Hb level <= 47 µg Hb/g 🡪 Round 5 (the data set does not contain information on this)

Thus, we can have the following cases:

* An individual never participates
* An individual only participates in one round (1 or 2 or 3 or 4)
* An individual participates in consecutive rounds (For example: 1,2 or 2,3,4 or 3,4)
* An individual participates in non-consecutive rounds (For example: 1,3 or 1,3,4 or 2,4 or 2,3,4)

The rounds in the data set are based on the invitation date, not based on the participation1. Thus, say individual A turns 55 in 2013, and receives an invitation but does not participate in this first round. She receives an invitation again in 2015, and chooses to participate this time. The data of this individual looks as follows:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sex | Birthyear | R1\_number | R1\_invitation\_date | R1\_participant\_status | R1\_date | R1\_Hb | R1\_age | R2\_number | R2\_invitation\_date | R2\_participant\_status | R2\_date | R2\_Hb | R1\_age |
| F | 1958 | 1 | 03/2013 | Non-participant | NA | NA | NA | 2 | 06/2015 | Participant | 07/2015 | 13 µg | 57 |

Thus, for each round we have information on: what round we are currently in (R\*\_number), when the invitation was sent (R\*\_invitation\_date), and whether the patient participates (R\*\_participant\_status). If the patient participates in said round, the data set also contains information on: date of test (R\*\_date), Hb value of test (R\*\_Hb), age at testing (R\*\_age).

For now, I’ve discussed with my supervisor that we’d start with only including individuals with consecutive *tests*. Thus, explicitly the following cases:

**Table 1:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Round 1 | R1\_Year | Round 2 | R2\_Year | Round 3 | R3\_Year | Round 4 | R4\_Year |
| 1 | X | 2000 |  |  |  |  |  |  |
| 2 | X | 2000 | X | 2002 |  |  |  |  |
| 3 | X | 2000 | X | 2002 | X | 2004 |  |  |
| 4 | X | 2000 | X | 2002 | X | 2004 | X | 2006 |
| 5 | *NAN* | *NAN* | X | 2002 |  |  |  |  |
| 6 | *NAN* | *NAN* | X | 2002 | X | 2004 |  |  |
| 7 | *NAN* | *NAN* | X | 2002 | X | 2004 | X | 2006 |
| 8 | *NAN* | *NAN* | *NAN* | *NAN* | X | 2004 |  |  |
| 9 | *NAN* | *NAN* | *NAN* | *NAN* | X | 2004 | X | 2006 |
| 10 | *NAN* | *NAN* | *NAN* | *NAN* | *NAN* | *NAN* | X | 2006 |

Which we recode to look like

**Table 2:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Test 1 | T1\_Year | Test 2 | T2\_Year | Test 3 | T3\_Year | Test 4 | R4\_Year |
| 1 | X | 2000 |  |  |  |  |  |  |
| 2 | X | 2000 | X | 2002 |  |  |  |  |
| 3 | X | 2000 | X | 2002 | X | 2004 |  |  |
| 4 | X | 2000 | X | 2002 | X | 2004 | X | 2006 |
| 5 | X | 2002 |  |  |  |  |  |  |
| 6 | X | 2002 | X | 2004 |  |  |  |  |
| 7 | X | 2002 | X | 2004 | X | 2006 |  |  |
| 8 | X | 2004 |  |  |  |  |  |  |
| 9 | X | 2004 | X | 2006 |  |  |  |  |
| 10 | X | 2006 |  |  |  |  |  |  |

This way, the time intervals between each observation within an individual is equal (2 years), with no NANs before the first entry (in contrast to the previous scenario where ID=5 to 10 have their first test in Round 2 or later).

I’ll have a meeting about this soon with my supervisors at EMC, and we will discuss other possibilities then2.

1. If time intervals were not varying/unequal, and if the data were approximately continuous, then standard time series forecasting methods or dynamic panel forecasting methods might be applicable. There are also new methods for the task of applying machine learning methods to multiple time series such as Feature Based Forecast Model Averaging.

FFMA is for determining the weights of the forecasts when you combine multiple machine learning algorithms, right? Are you suggesting that I might want to explore combining SVR and NN? I think this might become quite complex, as the forecasts from each individual model partly consist of a random effect, and I’d figure that it might be near impossible to assign weights to an inherently random factor right? TOO LITTLE OBSERVATIONS TO DO THIS ANYWAY, AVERAGING OF METHODS MIGHT STILL BE POTENTIALLY VALID

1. Can you make a histogram or density plot of the dependent variable? It would be useful to know how much zero inflation there is in the data, and what is the distribution of the nonzero observations. If there are mostly low discrete values, then a count data model is applicable. Otherwise a continuous outcome model (e.g. zero inflated or hurdle model) is applicable.

Yes, I can! I just need to make a few more choices in the data pre-processing (mainly with regards to the first point you made), which will affect the number of individuals that remain in the analysis, but once we made a decision on this, I’ll send you a

1. It would also be useful to know some summary statistics such as the averages of some covariates for observations with outcomes equal to zero and nonzero outcomes. This can potentially inform the choice of model.

Again, good point, and I’ll be sure to get this to you as soon as we’ve worked out some kinks in the data set!

1. It is also possible to include lags of other variables, not just the dependent variable.

This could be possible. I could include previous bloodtest result, although this is kind of inherent to previous Hb value, because the bloodtest result is based on whether the Hb value is above a certain threshold or not. Thus, I don’t think including this variable would add new information. What do you think? There really aren’t any other variables I could include, the data set is quite limited.

Also, I forgot to mention it in my proposal, but the stage variable (which indicates in which of 7 stages you’re in) is only available in case you have cancer, because a colposcopy has to be performed for this variable. So, this stage variable only occurs for individuals who have clinical cancer, and it only occurs once per individual, given that if an individual leaves the data set once cancer has been detected (they go into a surveillance program instead of the screening program). So it’s not possible to include a lagged version of this variable.

1. What is the number of observations? Is 2,493,999 the number of observations?

No, this number was based on a different data set. The data set I will be using contains 3,552,602 observations, with ID values ranging from 1- 3,710,672 (some IDs have apparently been deleted).

1. How do you plan to split the data into training and test data? How many time periods ahead do you plan to make predictions? (i.e. what is the forecast horizon?) Do you plan to train the models on the same data as the data used for predictions (predicting outcomes for time periods not yet observed), or do you plan to make predictions for entirely new time series? I assume the former if you plan to use estimates of the mixed effect.

I plan on creating three distinct groups of individuals. Group 1 is for training, group 2 is for validation, and group 3 is for testing. Groups 1 and 2 will probably be around +/- 300,000 individuals each (due to computation time), which leaves over 2,5 million individuals for the test group 3. Thus, each ID occurs only once over those three groups. I.e., if ID=5 is in group 1, this ID will not be in group 2 or group 3. With as forecast horizon the next test or round, depending on our decisions with data pre-processing.

1. Do you plan to use rolling or expanding windows, or just one training/test split for predictions at a particular point in time?

I intend to predict one period ahead, based on the last round of an individual. That is, if an individual’s last round is Round 3, we predict round 4.

1. You note that "mixed effects models are virtually unexplored for NN and SVM" while referencing several papers in sections 2.3.1 and 2.3.2. Is there something that I have misunderstood? Perhaps your point is that there are not many papers relevant to the total number of papers applying NN and SVM to prediction tasks.

Ah, I actually meant to take that sentence out. What I meant to say instead was that it is difficult to find papers that did *not* use mixed-effects, and modeled the correlations within patients explicitly in the ML model (such as what I intend to do, by including all the previous values, and the difference to catch trends). The problem is that the google search is biased as soon as you include ‘longitudinal’ with your machine learning method, because this will only show researches who did something outside of the ML model to account for this type of data. So, if I want to figure out how people modeled this type of data using ML without using additional algorithms, I’d just have to google NN in healthcare which leaves a bottomless pit of papers.

Basically, I couldn’t find any papers that made assessed how useful it is to include mixed-effects, compared to not including mixed-effects, when you attempt to explicitly model the correlation within individuals through the inclusion of lags. Sorry for the wordy explanation, I hope it’s clear what I’m trying to say.

1. The FIT number is on-hot encoded. There are other methods for encoding categorical data that are worth considering. See for example, the following paper and package:  
   <https://arxiv.org/abs/1908.09874v3>  
   <https://github.com/grf-labs/sufrep>

This is definitely interesting, but I don’t really know that it is worth using this method given that the number of groups in FIT number is only four, and we’d have to make assumptions, while one of the beauties of using ML methods is that they don’t rely on a priori assumptions. What do you think?

1. Other variations on SVR worth considering are Least Squares SVR, Relevance Vector Machines, and Bayesian SVR.

I would really like to. I’ve actually already written a few things down about LS-SVR, but the mixed-effects package available in Python (based on the Ngufor research) is only built to incorporate sklearn methods. Unfortunately, none of the methods you mention are available in sklearn. I will try and see what I can do though, once I start programming.

1. Section 4.2: What is the link function for this paper? I note that Zero-Inflated Negative Binomial models are no longer included in the proposal.  
     
   If the link function is approximately continuous and bounded from below by zero, there are other possible models, such as a hurdle model or zero inflated continuous outcome model, maybe with a log link conditional on being nonzero.

Indeed ZINB is no longer included, I only included the model to show what is being used now in practice, but given that I’ve changed directions in my research question, I didn’t feel the need anymore to describe this model in depth. It only made readers more confused, as I never planned to actually planned on incorporating this model, but the way I worded it made it seem that way.

I was actually planning on using the identity function, which is wat Danica van de Berg also did, constrained on non-zero outcomes. But maybe another function would be more suitable, do you have any advice on how I can determine what link function I should use?

1. Some other performance measures worth considering are Superior predictive ability tests, model confidence sets, other scoring rules (try to find some scoring rules for zero inflated outcome data), and graphs to check for calibration of predictions.

Consider including a separate comparison of predictions of zero-counts. See for example section 3.7 of the following document  
<https://cran.r-project.org/web/packages/pscl/vignettes/countreg.pdf>

Good points, I will look into this some more once I get closer to this point in my research.

**POSSIBILITIES:**

1. Only consider consecutive tests [what I’ve proposed in the document]
   1. Or, we do not recode the observations, and keep the original structure of Table 1, and include a dummy variable which indicates whether this is the first fit or not. However, I do not necessarily prefer this over what I’ve mentioned before, as the fact that someone only started to show up from round X onward holds no information about whether this individual has cancer or not. But, I could be wrong.
2. Similar to 1, instead of predicting the next *round*, we predict the next *test*. A test moment is only included if an individual shows up in a certain round, and will not be included at all otherwise. In this case, we can use all data and we would convert a table such as the following:

**Table 3:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Round 1 | R1\_Year | Round 2 | R2\_Year | Round 3 | R3\_Year | Round 4 | R4\_Year |
| 1 | X | 2000 |  |  |  |  |  |  |
| 2 | X | 2000 | *NAN* | *NAN* | X | 2004 |  |  |
| 3 | X | 2000 | X | 2002 | X | 2004 |  |  |
| 4 | X | 2000 | *NAN* | *NAN* | X | 2004 |  |  |
| 5 | X | 2000 | X | 2002 | X | 2004 | X | 2006 |
| 6 | *NAN* | *NAN* | X | 2002 | X | 2004 |  |  |
| 7 | *NAN* | *NAN* | X | 2002 | *NAN* | *NAN* | X | 2006 |
| 8 | *NAN* | *NAN* | *NAN* | *NAN* | X | 2004 |  |  |
| 9 | *NAN* | *NAN* | *NAN* | *NAN* | X | 2004 | X | 2006 |
| 10 | *NAN* | *NAN* | *NAN* | *NAN* | *NAN* | *NAN* | X | 2006 |

To this table:

**Table 4:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Test 1 | T1\_Year | Test 2 | T2\_Year | Test 3 | T3\_Year | Test 4 | T4\_Year |
| 1 | X | 2000 |  |  |  |  |  |  |
| 2 | X | 2000 | X | 2004 |  |  |  |  |
| 3 | X | 2000 | X | 2002 | X | 2004 |  |  |
| 4 | X | 2000 | X | 2004 |  |  |  |  |
| 5 | X | 2000 | X | 2002 | X | 2004 | X | 2006 |
| 6 | X | 2002 | X | 2004 |  |  |  |  |
| 7 | X | 2002 | X | 2006 |  |  |  |  |
| 8 | X | 2004 |  |  |  |  |  |  |
| 9 | X | 2004 | X | 2006 |  |  |  |  |
| 10 | X | 2006 |  |  |  |  |  |  |

Then, instead of predicting round 3 for ID=2 with a NAN for the previous round, we’d predict test 2 for ID=2, which uses test 1 as previous value.

A problem with this idea, however, is that the time intervals between test observations need not be the same. We could include an independent variable which shows the time between the previous tests and the test we’d like to predict, but I don’t know if that fully solves the problem.

1. Only consider full participation from round 1 onwards:
   1. 1
   2. 1,2
   3. 1,2,3
   4. 1,2,3,4 (fully healthy individuals)

This option leaves the least amount of data, but maintains the rounds structure of the data.

Impute missing values between rounds (underlined NANs in Table 3). This method maintains the rounds structure of the data, but it does include made-up values, which puts it at a disadvantage compared to the other methods.