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 then.

**POSSIBILITIES:**

1. Only consider consecutive tests
   1. With recoding [what I’ve proposed in the document]
   2. Without recoding the observations, to keep the original structure of Table 1, and instead 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. Specifically, what would ‘predicting one period ahead’ mean in this case?

1. Only consider full participation from round 1 onwards, so the following entries:
   * 1
   * 1,2
   * 1,2,3
   * 1,2,3,4 (fully healthy individuals)

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

1. 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.