Co-Morbidity in Bariatric Patients
A new approach in quantifying the severity

T.P.A. Beishuizen (0791613)

Biomedical Engineering - Computational Biology

Data Engineering - Information Systems

Eindhoven, University of Technology

Email: t.p.a.beishuizen@student.tue.nl

November 30, 2017

Contents

1   Introduction

The number of bariatric surgeries is increasing. worldwide. Although initially thought otherwise, this type of surgery has added benefits on top of losing weight. Among those benefits the remission of metabolic co-morbidities can be named. Due to binary labelling of those co-morbidities, valuable information is lost. On top of that the labelling is not clearly defined either. To obtain more and better results, this binary labelling could be replaced by a continuous severity score. Ruben Deneer conducted a research on trying to achieve a successful replacement.

The available data used for the research stemmed from the Catharina Hospital in Eindhoven (Subsection 1.1). This extensive data set was a combination of two data sets and consisted of 41 markers measured pre- and post-surgery for 2367 patients that underwent gastric sleeve or gastric bypass surgery. These 41 markers are divided in several sub-panels that each describe different processes in the body. Conditions for these co-morbidities that were tested for the severity score were type II diabetes mellitus (T2DM), hypertension and dyslipidemia. Extensive literature research was done to connect them with 41 markers to find possible relations (Subsection 1.2).

To better specify the research on the data, a main goal was created: *to use data mining techniques to develop score that can objectively quantify the severity of co-morbidities present in bariatric patients based on biomarkers, both before and after surgery.*[1] This means that it is not the goal to predict the outcome of the two types of surgery, but it is to quantify the improvement of the co-morbidities before and after the surgery.

Since the data is not available other than within the hospital, a second research is proposed to complement Ruben Deneer's. The creation of this research consists of four parts. The first part was creating a research question, hypothesis and an own analysis approach without the knowledge of Deneer's (section 2). Since no data is available, this can only be made globally. It should be detailed enough to compare it with Deneer's own approach, though. Secondly a summary is given what was done in Deneer's research (section 3). Thirdly the two are compared and remarks are given on the original approach (section 4). At last both are combined to create a final proposal for a possible follow-up research (section 5).

Over the course of this proposal report, many references are made to the thesis of Ruben Deneer. These will not always be included for redundancy. Only when things are directly quoted from the thesis, a reference will be placed. Examples would be tables or goals.

1.1   Data Sets

Two data sets were used in the study. The first one is called "The Dutch Audit for Treatment of Obesity" (DATO). This data set is a national database that houses all registrations and health statuses of pre- and post treatment bariatric surgery patients in the Netherlands. Several basic variables are noted, such as height, weight, BMI and date. Before surgery, the co-morbidities T2DM, hypertension and dyslipidemia were given a binary label of "Yes/No". After surgery they were given one of the following labels:

1. **Cured** No co-morbidity any more

2. **Improved** Less affected by co-morbidity

3. **Same** No change in co-morbidity status

4. **Worse** More affected by co-morbidity

5. **Denovo** Diagnosed co-morbidity while not present before surgery

6. **Not present** No co-morbidity present

The second data set came from a laboratory database, stored in health records. This extensive data set consisted of 3 clinical and 38 blood markers measured pre- and 6, 12 and 24 months post-surgery. The tests pre-surgery had some additional markers on top of the 41 ones. These markers

can be divided in the following categories: (Table 1.1) Complete blood count, liver function, kidney function, inflammation, lipid spectrum, coagulation, glucose metabolism, thyroid function and at last minerals and vitamins. The data sets of the patients that underwent bariatric surgery can be extracted from these.

One part of the research is to combine these two data sets. Some challenges arise when doing so. Such a challenge is obviously to find the right connection between them, using the survey and lab data of the same patient. These challenges must be solved before asking the question what markers can say about the severity of co-morbidities.

A second challenge would be to define what must be done with non-matching data. Pre-treatment for example more markers were used than post-treatment. These missing ones might be more useful for scoring the co-morbidity severity than the known markers. There also might be measurements that were missing or corrupted, whereas other markers might still be useful enough for a result. How to tackle this missing data challenge should be defined properly.

1.2   Co-morbidity Marker Relation

For some of the markers it is clear they affect the severity score of a co-morbidity. Co-morbidity is diagnosed from these markers and therefore a relation is imminent. For others more subtle relations are present, that in first sight would not be recognised. Both are important for assigning a co-morbidity severity score.

All three co-morbidities can be derived from specific markers. T2DM is known for higher glucose and insulin levels in the blood, while hardly finding any C-peptides. Ideally these would be used to find T2DM, however these values vary significantly over the day. Therefore help from other markers should be helpful. Hypertension is best measured with blood pressure measurements (BP). These however are not reliable for the nervous patients in the hospital and other markers must be found for help. Thirdly dyslipidemia obviously can be found using markers in the lipid spectrum (Table 1.1), however there may be other markers contributing to dyslipidemia detection, as well.

For T2DM, hypertension and to a lesser extend dyslipidemia, more knowledge on relations between them and markers would benefit the creation of a severity score. Literature research is done to understand more of those subtle relations and know that those exist when trying to create a morbidity score. The outcome of this literature search shows that quite some relations can be found between co-morbidities and the markers (Table 2). For most it is definitely worth to investigate any relation.

Table 1: The markers present in the bariatric laboratory data set [1]

	Before Surgery/Pre-Op/Screening	After Surgery/Post-Op/Follow-up
Complete blood count	hemoglobin	hemoglobin
	hematocrit	hematocrit
	erythrocytes	erythrocytes
	mean corpuscular hemoglobin	mean corpuscular hemoglobin
	mean corpuscular volume	mean corpuscular volume
	thrombocytes	thrombocytes
	leukocytes	leukocytes
Liver function	bilirubin	bilirubin
	aspartate aminotransferase	aspartate aminotransferase
	alanine aminotransferase	alanine aminotransferase
	lactate dehydrogenase	lactate dehydrogenase
	alkaline phosphatase	alkaline phosphatase
	gamma-glutamyltransferase	gamma-glutamyltransferase
Kidney function	urea	urea
	creatinine	creatinine
	potassium	potassium
	sodium	sodium
	calcium	calcium
	phosphate	phosphate
	albumin	albumin
Inflammation	C-reactive protein	C-reactive protein
Lipid spectrum	total cholesterol	total cholesterol
	high-density lipoprotein-cholesterol	high-density lipoprotein-cholesterol
	total/high-density cholesterol ratio	total/high-density cholesterol ratio
	low-density lipoprotein-cholesterol	low-density lipoprotein-cholesterol
	triglycerides	triglycerides
Coagulation	prothrombin time	prothrombin time
Glucose metabolism	hemoglobin A1c (IFCC)	hemoglobin A1c (IFCC)
	glucose	glucose
	insulin	-
	C-peptide	-
Thyroid function	parathyroid hormone	parathyroid hormone
	thyroid-stimulating hormone	-
	free T4	-
	cortisol	-
Minerals and vitamins	iron	iron
	ferritin	ferritin
	folic acid	folic acid
	zinc	-
	magnesium	-
	vitamin A	-
	vitamin B1	vitamin B1
	vitamin B6	vitamin B6
	25-OH vitamin D	25-OH vitamin D
	vitamin B12	vitamin B12

Table 2: Relations found in literature between the co-morbidities and markers.[1]. The number shows how many citations were found that state a relation exist, nothing means no relation. A 'D' means that the disease is diagnosed from those markers.

	Diabetes	Hypertension	Dyslipidemia
hemoglobin	3	1	
hematocrit	2	2	
erythrocytes	1	1	
MCH		1	
MCV		1	
thrombocytes			
leukocytes	4	3	1
bilirubin	3	1	1
ASAT	1	1	1
ALAT	4	1	1
LD	1		
Alkaline phospatae	2		
gamma GT	4	2	1
urea	1	3	
creatinine		2	1
potassium		2	
sodium		1	
calcium	2	1	1
phosphate			
albumin			
CRP	3	4	1
cholesterol		2	D
HDL-cholesterol		2	D
chol/HDL ratio			D
LDL-cholesterol		2	D
triglycerides		2	D
prothrombin time			
hemoglobin A1c (IFCC)	D		
glucose	D	2	
parathormone	1	2	
iron			
ferritin	3	1	1

2 First Proposal

To start of a research proposal, the first thing that needs to be done is to create a research question that fits the intended goal. This main question can then be divided in several sub questions for explaining the steps how to answer it. Next hypotheses should be made for the main and sub questions. Third and last the methods of how to answer the research questions are discussed. Since this is only a first proposal, the methods part will be limited, and extended in future proposals.

2.1 Research Questions

As stated earlier (Section 1) the goal was to develop a severity score for the co-morbidities using data mining. A research question for this goal would then be:

> *Can a severity score for co-morbidities present in bariatric patients based on biomarkers be developed?*

This main research question includes all aspects of the goal. It should be specified in several different sub-questions ot make it clearer. A first sub-question could be about the relation between co-morbidities and biomarkers, which is a major part on making this score.

1. *Which relations are present between co-morbidities and the biomarkers?*

A second sub-question should be about retrieving a score for the co-morbidities from the biomarkers. This score retrieval should use the outcome of the first sub-question to efficiently find the correct score.

2. *How can a severity score of co-morbidities be obtained from biomarkers?*

This second question mainly focuses on techniques to achieve a sufficient answer to the main research question. When both sub-questions are sufficiently answered, then the main one should definitely be, as well.

2.2 Hypothesis

The data set available for this research is quite extensive. The number of biomarkers is high, there are many relations between them, co-morbidities are known and a high number of patients has been tested for both. These pre-conditions are very beneficial to finding positive answers for the research questions.

Knowing many relations between the co-morbidities and biomarkers beforehand implies that a lot of them should be present in the data set. to the first sub-question which relations are present, the answer would be two-fold. First strong relations between the co-morbidities and on the other hand biomarkers that help diagnosing them are expected. This would for example be the case for dyslipidemia and biomarkers of the lipid spectrum. Weaker relations are expected for all found in literature (Table 2).

As for the second sub-question how a score can be defined from biomarkers, several different options are possible. When knowing there is a relation between biomarkers and the co-morbidity, by machine learning combined with regression a function can be made predicting a score outcome for the biomarkers. This score is then compared with the categorical labelling of co-morbidities to find out if they match. Through training the model and afterwards testing it, the score is validated.

With the hypotheses for both sub-questions the main research question can be answered, too. The hypothesis for this one would be that it is possible to find a severity score with the biomarkers. A model can made using the biomarkers as an input and retrieving a severity score as output.

2.3   Methods

Since the materials are all explained in Deneer's thesis, that part and most of the pre-processing is skipped in methods. One thing that needs to be discussed is the handling of markers that are measured pre-surgery, but not post-surgery. These could be removed as being redundant. They may, however show a significant difference between availability or absence of a co-morbidity. This is a vital result for possible further bariatric surgeries and co-morbidity scoring. Therefore at least some tests should be done with those pre-surgery biomarkers. For post-surgery measurements they can be removed from the tests or they can be given a dummy value as being unknown.

The first research question can be easily and efficiently answered with statistical analysis. This approach can be used to compare two groups of values, whether the co-morbidity is present or not. Since for this question a score is not yet needed, dividing the patients between having a co-morbidity or not suffices. The markers that show a significant difference in mean or variance should be noted and given extra care in future researching to create a score. Aside from comparing original measurements of the markers, combinations between them should be tested for statistical analysis as well. Some are already present, for example the *cholesterol / HDL-cholesterol ratio*. There may be more interesting combinations that could show a significant difference for either of the three co-morbidities.

Since the data needs some manipulation to find those markers that are significantly different for presence or absence of co-morbidities, a more advanced statistical program that can do those manipulation would be best to use. This means that the best choices would be R, MATLAB or Python. Python with SciPy was chosen of those three for also being a good choice for researching the second sub-question and because the actual users knows Python quite well. The outcome of this research can most likely immediately used in the second one for the next sub-question

The second research question is a bit more complex to answer. The quickest one is to use a way of machine learning to create this score. Some sort of regression is the first thing that comes to mind and simple support vector machines (SVMs) seem to be a good way to start. However obtaining a continuous score from categorical, or in the case of pre-surgery even binary, labels. A score interval needs to be set up and these binary or categorical values should be translated to a certain score on the interval. The way to score these values can be proposed, but most likely needs more insight to find out what is best. At first just a categorical learning system seems best. Both a selection of from sub-question resulting useful markers should be used as well as the complete set to find a good score.

At the start of researching the second sub-question a standard machine learning framework would suffice. This means that the aforementioned Python with the Scikit-learn package would be best to start with. When initial results indicate that a certain way of handling this data in another way would be more efficient, Python is used by many other machine learning programs, such as Theano, TensorFlow and both their packages. Therefore Python would be a great way to start the project of which the direction can change heavily.

3 Original research

Now that an initial proposal is made, the original research should be discussed. The research is divided into two parts. The first is all about the material and how this was preprocessed. The second part focused more on modelling using the data sets. Multiple approaches have been used to show different aspects of the data.

3.1 Preprocessing

As discussed earlier (subsection 1.1), the data was retrieved from two data sets, the DATO and lab set. These two were not perfectly lined up and therefore merging those required some specific rules. At first, any entries that were double on the same date were removed if having conflicting information. Entries that had surgery before 2012 were removed, too, due to not having a pre-screening test. Secondly all DATO and lab sets were paired up if they did not have more than 90 days between them. When double matches occurred, the closest ones were taken. Also only pairs were taken when they were on one of the four periods (pre-, 6, 12 or 24 post-surgery) and others were omitted.

After merging both data sets the lab data was prone to preprocessing. Ten biomarkers were dropped from the set, due to missing values. Five (Zinc, vitamin A, free T5, methylmalonic acid, LDL) were dropped due to severe missing data or explanations by other data. Five others (TSH, insulin, C-peptide, magnesium and cortisol) were dropped due to only being available in pre-surgery screening. CRP was dichomotized, due to being truncated and other variables were changed to their cut-off value when only a small portion of them wa truncated, too. Inconsistencies in DATO meant that the patient would be dropped. After all these markers were being removed, a total of 60 variables were present in the data set (figure 3.1) and a number of 2367 patients with a 2.35 of data sets per patient on average (figure 3.1). The number of variables used for modelling was reduced to 49. From those, a number of variables were removed again. Weight and height (incorporated in BMI), vitamins (were supplemented), iron (variation too high), year, type of and months before/after surgery (no value, or found elsewhere) were all removed. Furthermore Creatinine was replaced by glomerular filtration rate (GFR), the ASAT/ALAT ratio was added, the prothrombin time (PT) was replaced by the international normalized ratio (INR) and at last calcium was corrected by adding an albumin factor.

3.2 Modelling

After preprocessing the data, a series of steps should be taken to create a sufficient model for the issue. First an ordinal outcome definition was made for scoring the co-morbidities. Next, statistical learning models were created, based on logistic regression and all aspects of this model were discussed to find, validate and visualize a correct model.

To make one scale for all co-morbidities a new ordinal outcome was created. This scoring was made with three different categories, being: no, one or multiple co-morbidities present. The choice for this scoring was made with the idea that people with more co-morbidities also had a higher severity compared with patients that had less or none. The multiple co-morbidities category was a merge from having two and three, in order to make the categories more balanced in size. This way none has most patients (1329) and one and multiple still had a high enough number (569 and 469 respectively).

To create a model in this research, a classification method should be used. A well known one for that was logistic regression. Maximum likelihood estimation was used to estimate regression coefficients for the variables. Multicollinearity had to be limited as much as possible, as it was assumed to be absent for logistic regression. A correlation matrix was created to find multi-collinearity. Special interactions were included as well, for example the Body mass index (BMI) and presence for co-morbidities as being a prerequisite for the surgery. For ordinal outcomes of the logistic regression, both proportional odds (PO) and continuation ratio (CR) were used. 10-fold cross validation was done to validate the model and an assessment of model fit was done

Variable	Description	Type	Used in modeling
patient	Unique patient ID	character	No
OK.date	Date of surgery	date	No
period	Time period, i.e. Pre-op, 6 Months, 12 Months or 24 Months	factor	Yes: predictor
time	Months before or after surgery	integer	No
gender	Male or female	factor	Yes: predictor
age.at.surgery	Age at sugery	continuous	Yes: predictor
height	Height in meters	continuous	No
weight	Weight in kg	continuous	No
BMI	BMI in kg/m^2	continuous	Yes: predictor
surgery	Gastric sleeve or bypass	factor	No
hemoglobin	Hemoglobin in mmol/L	continuous	Yes: predictor
hematocrit	Hematocrit in L/L	continuous	Yes: predictor
erythrocytes	Erythrocytes in /pL	continuous	Yes: predictor
MCH	Mean corpuscular hemoglobin in fmol	continuous	Yes: predictor
MCV	Mean corpuscular volume in fL	continuous	Yes: predictor
thrombocytes	Thrombocytes in /nL	continuous	Yes: predictor
leukocytes	Leukocytes in /nL	continuous	Yes: predictor
glucose	Glucose in mmol/L	continuous	Yes: predictor
bilirubin	Bilirubin in μmol/L	continuous	Yes: predictor
ASAT	Aspartate aminotransferase in U/L	continuous	Yes: predictor
ALAT	Alanine aminotransferase in U/L	continuous	Yes: predictor
ASAT.ALATratio	Ratio between ASAT and ALAT	continuous	Yes: predictor
LD	Lactate dehydrogenase in U/L	continuous	Yes: predictor
ALP	Alkaline phosphatase in IU/L	continuous	Yes: predictor
GGT	Gamma Glutamyltransferase in U/L	continuous	Yes: predictor
urea	Urea in mmol/L	continuous	Yes: predictor
creatinine	Creatinine in μmol/L	continuous	Yes: predictor
potassium	Potassium in mmol/L	continuous	Yes: predictor
sodium	Sodium in mmol/L	continuous	Yes: predictor
calcium	Calcium in mmol/L	continuous	Yes: predictor
phosphate	Phosphate in mmol/L	continuous	Yes: predictor
albumin	Albumin in g/L	continuous	Yes: predictor
CRP	C-reactive protein <6 or $\geq$ 6 mg/L	factor	Yes: predictor
cholesterol	Total cholesterol in mmol/L	continuous	Yes: predictor
triglycerides	Triglycerides in mmol/L	continuous	Yes: predictor
HDL	High-density lipoprotein in mmol/L	continuous	Yes: predictor
cholHDLratio	Total cholesterol to hdl ratio	continuous	Yes: predictor
INR	International Normalized Ratio	continuous	Yes: predictor
HbA1c	Hemoglobin A1c in mmol/mol	continuous	Yes: predictor
PTH	Parathyroid hormone in pmol/L	continuous	Yes: predictor
iron	Iron in μmol/L	continuous	No
ferritin	Ferritin in μg/L	continuous	Yes: predictor
folate	Folate in nmol/L	continuous	Yes: predictor
vitB1	Vitamin B1 in nmol/L	continuous	No
vitB6	Vitamin B6 in nmol/L	continuous	No
vitD25OH	Vitamin D 25-hydroxy in nmol/L	continuous	No
vitB12	Vitamin B12 in pmol/L	continuous	No
CKD.EPI	CKD-EPI eGFR in ml/min/1.73m^2	continuous	Yes: predictor
dia.full	Diabetes label as recorded in DATO	factor	No
hyp.full	Hypertension label as recorded in DATO	factor	No
dys.full	Dyslipidemia label as recorded in DATO	factor	No
dia.bin	Diabetes label converted to binary	logical	Yes: outcome
hyp.bin	Hypertension label converted to binary	logical	Yes: outcome
dys.bin	Dyslipidemia label converted to binary	logical	Yes: outcome
no.of.comorbs	Number of co-morbidities (out of 3)	continuous	No
noneVSrest	Binary label indicating if there are no co-morbidities	logical	Yes: outcome
multipleVSrest	Binary label indicating if there are multiple co-morbidities	logical	Yes: outcome
oneVSrest	Binary label indicating if there is one co-morbidity	logical	Yes: outcome
comorb	Ordinal co-morbidity label	factor	Yes: outcome
classlabel	Label indicating which combination of co-morbidities are present	factor	No

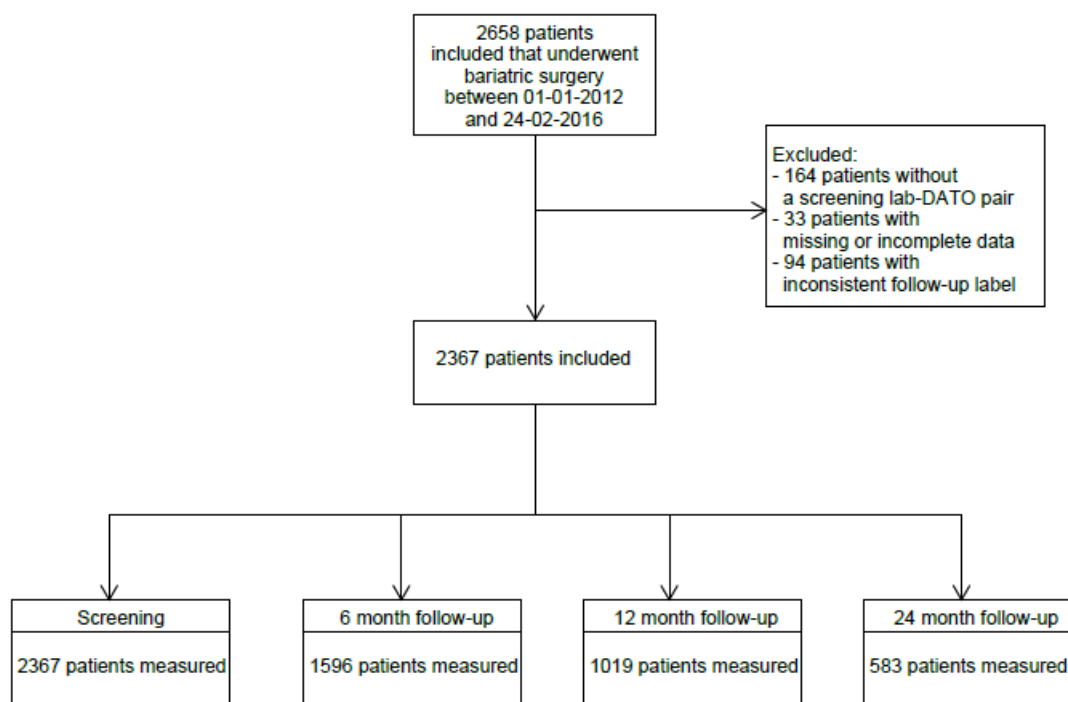Figure 1: All variables present after pre-processing [1]

Figure 2: All patients present after pre-processing [1]

for discrimination, calibration and overall performance. Lastly the model was visualized with a nomogram for further use.

4   Proposal Comparisons

A proposal was created after reading the introduction, background and material. The original research was also discussed and a summary was given. The new proposal should now be compared with the original research to find out when they match and when they do not. Several topics should be discussed that differ greatly between the original and new proposal. These differences should be addressed to create a final proposal for a new research. The first difference was about data omission. Many aspects of the original data set were debatable whether they should be removed or not. A second difference was the final scoring system. The newly proposed one was vastly different from the original one.

4.1   Data Omission

In preprocessing (subsection 3.1) a big number of data omissions was done with various reasons. While many of them were regarded as valid, a number of them could be seen as being a bit too rigorous. These questionable ones were discussed whether they should be left out.

1. **Lab set without a close DATO set** Several lab sets were removed because no DATO set was close enough. This seemed logical as they should match for complete results. However the DATO set had many variables that are not subject to much change over time. If a lab set was squeezed in between DATO sets that were very similar, an estimate for the intermediate DATO set at that time could easily be made.

2. **Predefined measuring moments** The DATO and lab sets were all measured on predefined moments. There were other times as well, though, were measurements took place. On those other times data was omitted. These omissions seemed too rigorous. These data sets were maybe not useful enough to show differences on those measuring moments, however these sets could be very helpful to create a model. These questionable omissions included pairs that were double at one moment and moments at 36 or 48 months.

3. **Markers only in pre-surgery measurements** Five of the biomarkers were only measured before surgery. The reason for this was not stated, however these markers were removed as about half of the data sets had these. When almost 50% of the data sets had these markers, there was enough available to check these variables on their relation with the co-morbidities. If all of them showed no significant difference in presence or absence of the co-morbidities, then they were not useful. If they would show a significant difference, they might be a good marker for further co-morbidity detection. Disregarding them only because half of the sets did not have the value seemed questionable.

4. **LDL-cholesterol** This marker from the lipid panel was removed due to its possible derivation from the same panel markers. The omission was discussed quite heavily in the thesis. Even though it was derived from other markers, the derivation was done in a very specific way. Maybe a description for one of the co-morbidities would benefit greatly of this specific derivation. Omitting LDL-cholesterol was less questionable than the other cases, however at least some tests could be done to check the significance.

4.2   Scoring System

At the project start a goal was made about creating a severity scoring system for co-morbidities. To achieve that from categorical data, as little data reduction as possible should be done to keep the score quality as high as possible. The proposed ordinal system however actually reduces information. First there is no distinction between two or three co-morbidities. The type of co-morbidity is removed, too, as well as its progress after surgery. This is a very counter intuitive way to correctly model a scoring system when these three decisions all could contribute to a more efficient scale.

A second remark about the scoring system is about the final product. The linear predictor score was not really an answer to the original goal. The goal was to create a severity score, whereas the linear predictor score in the nomogram is originated from its probability counterpart and not really showing severity. As discussed earlier, it was hard to create a severity score from categorical labelling, so maybe that goal might be too ambitious.

For this research it seemed better to keep the three co-morbidities separate as long as possible. For every co-morbidity separately a score could be made which in the end could be joined together. This score would be best based on a probability at the start, so how much chance there was to have a certain co-morbidity. After creating probability models for all of them, they could be combined in a joint probability model. While doing this, obviously collinearity needed to be taken into account. As recently discussed, a probability model was not according to the goal. A score was desired and not a probability. Since the data was not gathered in a way to find a score with the measured tasks, this was much harder. After possibly finding good results for a probability model, a next step could be to translate that model in a scoring system.

4.3   Modelling Tools

About every part of the original research was done in R. As discussed previously, especially the variable significance could be done efficiently in R, as it was one of the more advanced tools to do statistical analysis with. For the modelling part however there were more advanced programs or frameworks that could be used. One of the reasons to choose R was to create a "white-box" answer, that every step of the process could be explained. The disadvantage in choosing a white-box model, however, was that the result could be worse than for a more traditional machine learning approach. Those approaches could lead to a better result and were therefore definitely worth to investigate. In R however, machine learning options were extremely limited if even available. Other programs or frameworks could facilitate that better than R, such as Python with the Scikit-learn package, Theano or Tensorflow. This could also facilitate a better final visualization result for the scoring system than the nomogram, which was not very user-friendly and straightforward.

5 Final Proposal

The final proposal combined the first one and the original research for optimization. Questions from the first proposal were re-used, with some modifications and additions. The hypotheses were changed accordingly. The methods for answering the research questions were similar to the ones in the first proposal, however, they were extended after more information from the original research was added.

5.1 Research Questions

The main research question stayed the same as the one from the first proposal. This was the main goal of the research and therefore was a good all-encompassing question for it.

> *Can a severity score for co-morbidities present in bariatric patients based on biomarkers be developed?*

The first sub-question also stayed mostly the same. It was a bit unclear that there was relevant data in the DATO set. Therefore instead of saying biomarkers, the term "available variables" would be used. The word available was added due to not immediately omitting some of the variables, so their relation to co-morbidities would still investigated.

1. *Which relations are present between co-morbidities and the available variables?*

One thing the original research correctly investigated that was not thought of, was collinearity. This was already investigated for most variables, however the ones then removed might still have collinearity characteristics. Therefore a new sub-question was added to find those.

2. *Which collinearity relations are present within the available variables?*

The other sub-question was about adding a score to the co-morbidities originated from variables. This question was quite vague and with the help of the original research it was split up in three separate questions. The first one was about modelling the co-morbidities with variables.

3. *How are the co-morbidities best modelled with the available variables?*

The second part of the split sub-question was about the scoring part. Eventually a score should be given to the co-morbidities, which must be investigated thoroughly.

4. *How can scores best be added to the available co-morbidity models?*

The last part was about combining the co-morbidities together. Until now they were regarded separately, but the eventual result should have these scores combined. Therefore the third one of the split sub-question was about combining those.

5. *How can the co-morbidity models best be combined?*

This main question and the five sub-question should cover all aspects in this research. If all questions were answered one by one, a scoring model for the co-morbidities should be available for bariatric patients.

5.2 Hypotheses

Since three sub-questions are added. The new hypotheses should be added, as well. They will be addressed in order of appearance and afterwards a general hypothesis will be given for the main research question.

The first sub-question has the same hypotheses as in the earlier proposal. Strong relations should be found in the markers that the co-morbidities are diagnosed from and weak relations

should be found in the ones that are linked in literature. On top of that, since the original research answered this question partially, too, the following variables should show strong relations: triglycerides, HbA1c, CKD.EPI, age.at.surgery, HDL, cholesterol0-HDL-ratio, albumin, period, potassium, period * BMI, cholesterol and urea. All of these showed the best correlation.

The second sub-question has also been partially answered in the original research. Therefore strong relations should be found between: hematocrit and hemoglobin, MCV and MCH, ALAT and ASAT, HbA1c and glucose, hemoglobine and gender, erythrocytes and hematocrit. Aside from these mostly the omitted variables should show relations within their own panel. Insulin for example is expected to show a relation with glucose and C-peptide.

The third sub-question is partially answered in a specific direction by the original research. Logistic regression was extensively investigated, however other machine learning aspects were not used. The hypothesis is that most likely several machine learning methods should be sufficient to answer this question. With those machine learning methods a model can be derived for the co-morbidities to be found out of the available variables.

The fourth question is the also hardest. How can a score be added? The original research started first with an ordinal score, which is debatable for the information thrown away. When thinking ahead, the best idea for now is to start with a probability score. These probabilities usually also show the severity, as having a higher probability for a co-morbidity, the severity of it usually is bigger, too. After having a probability score further insights can be used to create a better score than probability. An example of such a score can be this:

> **Score proposal** At the start a score interval of 0% to 100% will be chosen, with 0%-50% being not having the disease and 50%-100% having it, with different severities. Binary cases will then be given 25% and 75% as absent and present co-morbidity respectively, knowing those as the values they should be closest to. The ordinal cases that showed improvement or deterioration will be given a score of 50%.

The last sub-question is bigger than only adding them together. Since the co-morbidities are also related to each other, these correlations have to be taken into account when creating a score. On the other side, when two co-morbidities are both present in a patient, they also have a bigger chance to be more severe, as they strengthen each other. Different additions or combinations can be tested to combine them.

The hypotheses of most sub-questions are fairly straightforward and with the original research are known to give results already. The aim is however to create better answers than the already available ones. For the first two questions this is definitely possible, the last three depend whether the results will be improving the known ones already, mainly because it is a new area for researching. Still, though, as there is already one present, the main research question can be answered with a yes.

5.3 Methods

This time the methods is changed a bit from the first proposal, as it would be better to line them up with the original research and show differences between those. So first we start with the preprocessing and every thing that is related to that. Secondly the modelling part will be discussed, together with the sub-questions that are answered in here.

In preprocessing almost every thing will be done the same as in the original research. Most differences will be in not removing certain parts of the data set. This means that more than one data set can be made for the different instances. All differences will be discussed.

1. **All paired DATO and lab set** All data that is available after pairing the DATO and lab set will be put in a data set. There will not be any distinction between data on predefined moments and not, because for modelling those do not matter at all. In case more results should be generated only on those predefined moments, a separate attribute set will be made to extract those.

2. **Variables omission** Not all variables will be omitted when a significant number of them is available. Relations between those variables and with the co-morbidities will be tested. When finding out that there are significant relations between them, possible further research will be done on both a set with and without them. Variables that fall into this category are the five that are only measured pre-surgery and LDL.

Several steps should be taken to achieve the final goal. First the variables needs to be compared with the co-morbidities (sub-question 1) and themselves (sub-question 2). Next a model has to be found (sub-question 3) for all co-morbidities as well as a score added to the model (sub-question 4). Lastly the three models need to be combined (sub-question 5) to create a final answer to the main research question.

To answer the first two sub-questions, statistical analysis must be done. Correlation matrices must be made to see whether two variables or one and a co-morbidity correlate significantly. Also the variables must be checked on their own population, whether they have a high variance and follow a statistical distribution. In case of high correlations between variables, a joint variable should be created if possible. At last, also combinations of variables should be created and tested for relations with co-morbidities. This all can efficiently be done in Python with the SciPy package.

A co-morbidity can be modelled using different approaches in machine learning. Several classification techniques are available such as support vector machines (SVM), nearest neighbour (NN), stochastic gradient descent (SGD). Research must be done which is most applicable and testing must be done which gives the best results. This is repeated three times for every co-morbidity. Within Python the Scikit-learn package has the basics to create proper results and if needed, programs such as Theano or TensorFlow can be used within Python for more extensive research. For testing, a similar approach as in the original result can be used, called 10 fold cross-validation. The final model and the visualisation depends on the type of machine learning used, however graphs that show clusters of morbidities for several variables seems a good way to approach that.

Scoring addition in machine learning approaches is a challenge. At first tests can be done with changing the methods from classification approach to their regression counterpart, for example from SVC (Support Vector Classification) to SVR (-Regression). This can have major for now unknown implications. As discussed earlier (subsection 5.3) setting up a score interval should be thought of before and also able to be implemented. The proposed score can be used as a base and be changed if advantageous.

The last challenge is to create a model that combines all three of the co-morbidities. To start this, all three separate models can be used to create the combination. If this does not give the desired results and makes scores much worse than initially thought, a similar ordinal system as used in the original approach can be used. Then sub-questions three and four are repeated only with co-morbidities combined, removing details from each different one.

References

[1] R. Deneer, "Scoring co-morbidity severity in bariatric patients based on biomarkers: a data mining approach," 2017.