

EINDHOVEN UNIVERSITY OF TECHNOLOGY

8Z230 MASTER THESIS COMPUTATIONAL BIOLOGY

DEPARTMENT OF BIOMEDICAL ENGINEERING

**Scoring co-morbidity severity in
bariatric patients based on
biomarkers: a data mining approach**

Graduation professor:

PROF. DR. P.A.J. HILBERS

Supervisors:

PROF. DR. IR. N.A.W. VAN RIEL

DRS. IR. S.L.M. VAN LOON

DR. A.-K. BOER

Advisors:

DR. S.W. NIENHUIJS

PROF. DR. V. SCHARNHORST

PROF. DR. E.R. VAN DEN HEUVEL

Author:

R. DENEER

0591954

May 31, 2017

Abstract

Objective

Worldwide there is an increase in the number of bariatric surgeries performed. The benefits of bariatric surgery extend beyond weight-loss and include improvement or remission of metabolic co-morbidities, especially in type II diabetes mellitus (T2DM). The traditional dichotomization (present/absent) of co-morbidities results in a loss of information and is not always clearly defined. In this study, data mining was used to develop a novel continuous co-morbidity severity score for bariatric patients.

Methods

Using data mining techniques an extensive set of 3 clinical and 38 blood markers measured pre- and 6, 12 and 24 months post-surgery are retrospectively analyzed for 2367 patients that underwent gastric sleeve or gastric bypass surgery. Since the co-morbidities of interest (T2DM, hypertension and dyslipidemia) were correlated, an ordinal outcome was defined that labeled patients as having none, one or multiple co-morbidities. Ordinal logistic regression models were fitted and 10-fold cross validated.

Results

Five markers were selected and used in the regression model to score the severity of co-morbidities in bariatric patients: age at surgery, HbA1c, triglycerides, CKD-EPI and potassium. The Somers' D_{xy} rank correlation between the score and the ordinal outcome was 0.634 (0.601 - 0.667). The score was also able to discriminate patients with one or multiple co-morbidities from patients without co-morbidities, the estimated area under the ROC curve (AUC) was 0.817 (0.801 - 0.833).

Conclusion

By mining data stored in clinical databases, a novel co-morbidity severity score for bariatric patients was developed. The developed score showed a strong positive correlation with the number of co-morbidities present in bariatric patients. With this score clinicians can monitor future patients with respect to overall improvement in co-morbidities and compare them to peers.

Contents

1	Introduction	6
2	Background	9
2.1	Bariatric surgery	9
2.1.1	Sleeve gastrectomy	9
2.1.2	Gastric bypass	10
2.2	Biomarkers	11
2.2.1	Complete blood count	13
2.2.2	Liver function	13
2.2.3	Kidney function	14
2.2.4	Inflammation	14
2.2.5	Lipid spectrum	14
2.2.6	Coagulation	15
2.2.7	Glucose metabolism	15
2.2.8	Thyroid function	15
2.2.9	Minerals and vitamins	16
2.3	Associations between biomarkers and co-morbidities	16
3	Materials and Methods	18
3.1	Pre-processing	18
3.1.1	The DATO dataset	19
3.1.2	Laboratory dataset	20
3.1.3	Merging datasets	20
3.1.4	Missingness	23
3.1.5	Data cleaning	25
3.1.6	Variable selection	27
3.1.7	Variable extraction	27
3.1.8	Data description	29
3.2	Modeling	31
3.2.1	Ordinal outcome definition	31
3.2.2	Statistical learning models	32
3.2.3	Logistic regression	34
3.2.4	Proportional odds and continuation ratio	34
3.2.5	Cross validation	35

3.2.6	Assessment of model fit	35
3.2.7	Model visualization	35
4	Results	38
4.1	Patient population	39
4.1.1	Prevalence of co-morbidities	39
4.1.2	Prevalence of ordinal co-morbidity labels	40
4.1.3	Predictor variables	42
4.2	Modeling	43
4.2.1	Wald statistics	43
4.2.2	Cross validation results	44
4.3	Final model visualization	47
5	Discussion	50
6	Conclusions and recommendations	54
	References	55
	Appendices	59
A	Summary of Literature review	60
B	Concise Statistical Description of Dataset after Pre-processing	62
C	Logistic Regression	70
C.1	Generalized Linear Models	70
C.2	Logistic regression	71
C.3	Example	72
C.4	Interaction terms	73
C.5	Maximum Likelihood Estimation	74
C.6	Penalized Maximum Likelihood Estimation	75
C.7	Hypothesis tests	76
C.7.1	Likelihood Ratio test	76
C.7.2	Wald test	77
C.7.3	Example	78
D	Ordinal Logistic Regression	80
D.1	Proportional Odds Model	81
D.2	Penalized Extended Continuation Ratio Model	81
D.3	Overfitting and limits on number of predictors	82
D.4	Penalization of extended CR model	83
E	Ordinality Assumption Plots	84

F	K-Fold Cross Validation	87
F.1	Cross-validation	87
F.2	Group k-fold	88
G	Performance Measures	90
G.1	Overall performance	90
G.1.1	R-squared	90
G.1.2	Brier score	90
G.2	Discriminative ability	91
G.2.1	Sensitivity	91
G.2.2	Specificity	91
G.2.3	Area under ROC curve	91
G.2.4	Somers' D_{xy} rank correlation	92
G.3	Calibration	93
G.3.1	Calibration plot	94
G.3.2	Calibration slope and intercept	95
G.4	Model parsimony	95
H	Difference in Patient Population between Periods	96
I	Influential Observations	98
J	Full Models	99
J.1	Diabetes Logistic Regression Model	99
J.2	Hypertension Logistic Regression Model	100
J.3	Dyslipidemia Logistic Regression Model	102
J.4	Proportional Odds Model	103
J.5	Penalized extended Continuation Ratio Model	104
	Appendix References	108

List of commonly used acronyms

T2DM	Type II diabetes mellitus
KDD	Knowledge Discovery in Databases
BMI	Body Mass Index
TWL	Total body weight loss
EWL	Excess weight loss
SG	Sleeve gastrectomy
RYGB	Roux-en-Y gastric bypass
MCV	Mean corpuscular volume
MCH	Mean corpuscular hemoglobin
ASAT	Aspartate aminotransferase
ALAT	Alanine aminotransferase
LD	Lactate dehydrogenase
ALP	Alkaline phosphatase
GGT	Gamma-glutamyltransferase
CRP	C-reactive protein
PT	Prothrombin time
HbA1c	Hemoglobin A1c
PTH	Parathyroid hormone
TSH	Thyroid-stimulating hormone
ft4	free T4
DATO	Dutch Audit for Treatment of Obesity

MMA Methylmalonic acid

GFR Glomerular filtration rate

CKD-EPI Chronic Kidney Disease Epidemiology Collaboration (equation)

INR International Normalized Ratio

MLE Maximum Likelihood Estimation

PO Proportional Odds

CR Continuation Ratio

AUC Area under the ROC curve

FUP Follow-up

LR Likelihood Ratio

d.f. Degrees of freedom

CI Confidence Interval

1 | Introduction

Data mining is becoming essential to health care due to the large volumes of data that are generated and stored. Much of this data is currently seen as a by-product of health-care delivery instead of a central asset from which data mining techniques can extract useful knowledge [1]. Although data mining is only a single step in the process of knowledge discovery in databases (KDD) as proposed by Fayyad, Piatetsky-Shapiro and Smyth in 1996 [2], KDD and data mining are used as synonyms [3]. "The goal of predictive data mining in clinical medicine is to derive models that can use patient specific information to predict the outcome of interest and to thereby support clinical decision-making" [4].

In this study data mining techniques are applied to clinical data stored in separate databases of patients that underwent bariatric surgery at the Catharina Hospital in Eindhoven. Unique to this data is that a comprehensive panel of metabolic biomarkers is repeatedly measured in all patients before and after surgery. The number of bariatric surgeries performed at the Catharina Hospital has increased more than threefold in the past five years, from 233 surgeries in 2010 to 711 in 2015. This increase is in part due to a shift from bariatric surgery to metabolic surgery.

It has become clear that bariatric surgery has a positive effect that extends beyond weight loss and includes improvement in diabetes, hypertension, dyslipidemia and reduction of overall mortality [5] [6] [7] [8]. To encompass the beneficial effects of bariatric surgery on metabolic illnesses, the term "metabolic surgery" was coined in 2007 at the Rome *Diabetes Surgery Summit*. The eligibility criteria for metabolic surgery also include patients with moderate obesity combined with type II diabetes mellitus (T2DM), whereas before only severely and morbidly obese patients were eligible for surgery¹. The shift in eligibility criteria is also causing a shift in the definition of success of surgery away from weight loss and towards resolution of co-morbidities [9]. While weight loss is a quantitative measure, often expressed in percent of total

¹moderately obese is a body mass index (BMI) > 30 kg/m², severely obese > 35 kg/m² and morbidly obese BMI > 40 kg/m²

body weight loss (%TWL) or excess weight loss (%EWL)², there is no universally accepted quantitative measure of co-morbidities. In most studies, co-morbidities are expressed as percentage of patients that achieve remission based on a co-morbidity classification scheme. However, as noted in the review by Franco et. al [10] there is no uniform definition of how patients are identified as having co-morbidities, moreover the definitions of "resolution" or "improvement" are also heterogeneous. Brethauer et. al. proposed standardized outcome reporting for bariatric surgery, which includes guidance as to how co-morbidities can be objectively classified [11]. This is, however, not yet universally adopted and while standardized outcome reporting for bariatric surgery is objective, it is still a *classification* and no *quantification*. If the systolic blood pressure (BP) is equal to or above 140 mm Hg a patient is classified as hypertensive by WHO guidelines, but is a patient with a BP of 139 mm Hg really different from a patient with a BP of 140? This dichotomization of co-morbidities results in a loss of information.

Therefore the **main goal** of this research is: *to use data mining techniques to develop a score that can objectively quantify the severity of co-morbidities present in bariatric patients based on biomarkers, both before and after surgery*. This score is developed to provide a personalized report in terms of improvement in co-morbidities of patients that undergo bariatric surgery.

Since all the laboratory results of biomarkers (including those contained in an extensive panel used for bariatric patients) are stored in a database, combining this database with a database that contains the registrations of co-morbidity status provides comprehensive and objective data on which to build the scoring model. Biomarkers are not only objective, but also the database in which biomarker results are stored is of high quality. Therefore, biomarkers (and several clinical markers such as age, height, weight, etc.) are used as independent variables. As dependent variables, the three most important co-morbidities are considered: T2DM, hypertension and dyslipidemia.

An objective quantification of co-morbidities in the form of a score serves multiple purposes in the clinic. Firstly, like %TWL and %EWL it allows the clinician to monitor improvement over time and comparison to peers. Recently van de Laar et al. developed percentile charts for weight loss to deliver a clear message on sufficient weight loss at any postoperative time, similar to growth charts for children [12]. Secondly, resolution of co-morbidities results in a reduction in medication used and therefore provides evidence of costs-effectiveness for health care insurers. Instead of using medication use as a proxy for presence of co-morbidities, a score provides a more objective picture of the

²%TWL and %EWL are calculated by:

$$\%TWL = \frac{\text{Initial weight} - \text{Post-op weight}}{\text{Initial weight}} \times 100$$

$$\%EWL = \frac{\text{Initial weight} - \text{Post-op weight}}{\text{Initial weight} - \text{Ideal weight}} \times 100$$

Initial weight = weight before surgery

Post-op weight = weight after surgery

Ideal weight = weight corresponding to a BMI of 25 kg/m²

improvements of a patient. Finally, in developing the score it will become clear which markers are associated with co-morbidities and provide guidance in interpretation of laboratory results.

Overweight and obesity can progress into the *metabolic syndrome*, this syndrome is a cluster of central obesity, T2DM, hypertension and dyslipidemia. Metabolic syndrome is said to be present when three or more out of five risk factors are present, these risk factors are: enlarged waist circumference, high triglycerides, high blood pressure, high blood glucose levels and reduced HDL-cholesterol. Since the metabolic syndrome is more of a continuum instead of a dichotomous variable, several scores have already been developed for the metabolic syndrome [13] [14] [15] [16]. In the scores that have been developed, an overall metabolic syndrome score is calculated by adding weights to the five risk factors. While this provides a continuous score instead of a dichotomization, the score is by definition based on these five risk factors. Therefore other markers that may be better predictors of co-morbidities are not taken into account. In this study, the goal is to use data mining techniques to quantify *co-morbidities* instead of the *metabolic syndrome* and to identify markers that are associated with co-morbidities. The DiaRem score is another score that was developed specifically for bariatric surgery [17]. It is able to predict the probability of T2DM remission following gastric bypass surgery. However, in this study the goal is not to predict the pre-operative probability of remission but to quantify the severity of co-morbidities and monitor improvement in a patient before and after surgery.

First, background information is provided by giving a general description of the two main types of bariatric surgery that are performed at the Catharina Hospital in Eindhoven: gastric sleeve and gastric bypass. In addition, a description is given for each of the biomarkers that are contained in the laboratory panel of the Catharina Hospital. In the materials and methods section the entire process of KDD is explained, from raw data to the final model. This section is divided into two parts: the data pre-processing and subsequent modeling. The pre-processing covers the steps taken to obtain, starting from the raw data, a dataset that is suitable for analysis and model building. The modeling section covers the building and validation of the co-morbidity scoring model. A selection of statistical learning models that are applicable to our goal and dataset is given. The results are also split into three sections: the descriptive statistics, the model performance and model visualization. In the descriptive statistics section, patient demographics and prevalence of co-morbidities are shown. The results from the model building process are shown in terms of discrimination, calibration and validation. Also shown is which markers are significantly associated with diabetes, hypertension and dyslipidemia separately. The chapter is concluded by visualizing the final model and applying the model to the data. In the chapter thereafter the found results are discussed and the study is concluded.

2 | Background

This chapter provides background information regarding the two main types of bariatric surgery performed at the Catharina Hospital and the biomarkers that are contained in the bariatric laboratory panel which is used to assess the metabolic health of patients undergoing bariatric surgery at the Catharina Hospital.

2.1 Bariatric surgery

The greater reduction in body weight and remission of T2DM in obese patients that undergo bariatric surgery compared to non-surgical approaches [18] has led to a world-wide increase in bariatric surgery [19]. The eligibility criteria for bariatric surgery are given in table 2.1. There are different types of bariatric surgery but this study is limited

BMI (kg/m ²)	Additional criteria
>40	none
>35	combined with a co-morbidity
>30	combined with type II diabetes

Table 2.1: Eligibility criteria for bariatric surgery. The combination of a BMI >30 kg/m² and T2DM is also called metabolic surgery.

to primary gastric sleeve and gastric bypass surgery which are performed at the Catharina Hospital in Eindhoven. Gastric sleeve is more formally called sleeve gastrectomy (SG) and gastric bypass is called Roux-en-Y gastric bypass (RYGB).

2.1.1 Sleeve gastrectomy

SG is basically removal of the greater curvature (left side) of the stomach after which a small sleeve remains, see figure 2.1. The new banana-shaped stomach contains about 20 to 25% of the original stomach volume. Initially SG was labeled as a restrictive procedure because the weight loss was attributed to a reduction in stomach size. However, it later became apparent that modifications of gastrointestinal hormones play a significant role. Markedly reduced ghrelin levels (which is also known as the "hunger

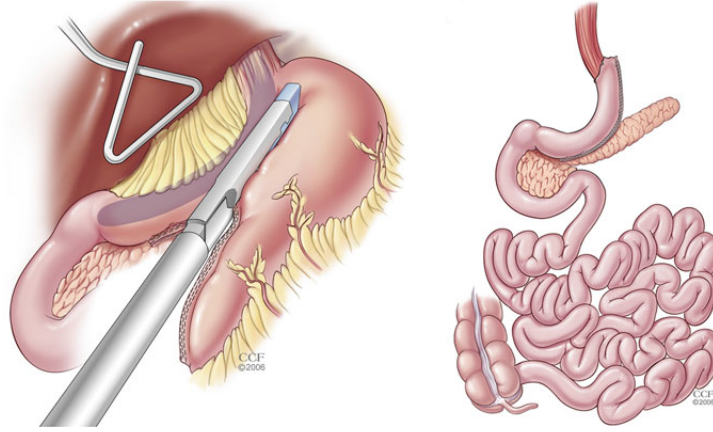


Figure 2.1: Sleeve gastrectomy, image courtesy of the Cleveland Clinical Center for medical art and photography.

hormone") in addition to increased peptide-YY levels are associated with greater appetite suppression and excess weight loss compared to RYGB [20]. SG has surpassed RYGB as the most performed bariatric surgery, 45.9% of all bariatric surgeries performed worldwide in 2014 were SG [19]. From a 2012 survey on laparoscopic sleeve gastrectomy which contained survey data from 46,133 SG performed by 130 surgeons worldwide, mean %EWL at year 1 was 59.3%; year 2, 59.0%; year 3, 54.7%; year 4, 52.3%; year 5, 52.4%; and year 6, 50.6% [21]. Weight loss seems to be maintained by SG as shown by Casella et al. who analyzed long term outcomes at 6 and 7 years for 182 patients that underwent SG, %EWL at 6 years was found to be 67.3% and at 7 years 65.7% [22]. SG is also known to resolve T2DM with remission rates reported around 80% [23] [24] [25] [26]. The benefits of SG on T2DM occur very early with a study showing beneficial effects within three days after surgery which seem to be related to hormonal changes [27]. SG is also shown to have a beneficial effect on other co-morbidities including hypertension [28] and dyslipidemia [29]. The long term effects of SG on resolution of co-morbidities are not well known.

2.1.2 Gastric bypass

In RYGB a small pouch is created from the upper stomach, see figure 2.2. The small intestine is divided between the upper and middle section (duodenum and jejunum), the middle section is then brought up and connected to the newly created stomach pouch. The stomach is bypassed by connecting the upper section of the small intestine to the middle section of the small intestine in a Y-configuration. Stomach acids and digestive juices can flow from the stomach to the small intestine through this connection. A systematic review and by Buchwald et al. showed that the mean %EWL for RYGB is 68.2% (95% CI: 61.5%-74.8%) [30]. Medium and long term follow-up was investigated by a systematic review from O'Brien et al. and showed that %EWL was highest in the



Figure 2.2: Roux-en-Y gastric bypass, image courtesy of the Cleveland Clinical Center for medical art and photography.

first year, 67.3%, and decreased to 52.5% at 10 years [31]. In similar fashion to SG, RYGB also has a beneficial effect on co-morbidities. In the same research by Buchwald et al. diabetes was resolved in 83.7% (95% CI: 77.3% to 90.1%) of patients that underwent RYGB, hypertension resolved in 67.5% (95% CI: 58.4% to 76.5%). Dyslipidemia is characterized by hyperlipidemia in the research by Buchwald et al. Overall hyperlipidemia resolved in 96.9% (95% CI: 93.6% to 100.0%) of patients. There were no significant differences between hypercholesterolemic and hypertriglyceridemic patients. Although Karamanakos et al. state that fasting ghrelin levels did not change significantly after RYGB [20], Cummings et al. state that RYGB is associated with markedly reduced ghrelin levels [32].

2.2 Biomarkers

To assess the metabolic health of patients before and after bariatric surgery, the Catharina Hospital in Eindhoven makes use of an extensive bariatric laboratory panel. This panel includes blood tests with markers related to complete blood count, kidney function, liver function, inflammation, lipid spectrum, coagulation, glucose metabolism, thyroid function, mineral and vitamin status (see table 2.2). This panel is extensive because bariatric surgery is an invasive procedure affecting multiple systems within the body, therefore the patient's nutrient status and organ functions have to be monitored before and after surgery. Note that some markers are only measured before surgery and that these are *not* fasting blood samples. Below a short description is given for each of

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

Table 2.2: List of markers contained in the bariatric laboratory panel as used by the Catharina Hospital.

the markers grouped by sub-panel.

2.2.1 Complete blood count

Hemoglobin is the oxygen carrying component of the erythrocytes (red blood cells) and the hemoglobin concentration measures the amount of hemoglobin in the blood.

Hematocrit is the volume percentage of erythrocytes in the blood.

Erythrocytes is the amount of red blood cells (erythrocytes) in the blood, also known as a red blood cell count.

Mean corpuscular volume (MCV) measures the mean or average size of individual erythrocytes. It is calculated from:

$$\text{MCV (fL)} = \frac{\text{hematocrit (L/L)}}{\text{erythrocytes (/pL)}} \times 10^3$$

Mean corpuscular hemoglobin (MCH) measures the amount of hemoglobin present in one erythrocyte. It is calculated from:

$$\text{MCH (fmol)} = \frac{\text{hemoglobin (mmol/L)}}{\text{erythrocytes (/pL)}}$$

Thrombocytes is the amount of platelets in the blood, also known as a platelet count.

Leukocytes is the amount of white blood cells (leukocytes) in the blood, also known as a white blood cell count. This test is used to detect infection or leukemia and to check immune system function.

2.2.2 Liver function

Bilirubin is a by-product of hemoglobin breakdown and travels to the liver where it is conjugated and eventually excreted in the small intestine.

Aspartate and alanine aminotransferase (ASAT and ALAT) are transaminases that the liver utilizes to catalyze reactions in which an amino group is transferred from a donor to an acceptor. In case of liver damage, these transaminases leak into the blood circulation and become elevated when measured. While ALAT is found predominantly in the liver, ASAT is also found in other organs. Therefore ALAT is a more specific indicator of liver inflammation than ASAT.

Lactate dehydrogenase (LD) is an ubiquitous enzyme that catalyzes the oxidation of NADH or reduction of NAD^+ , liver tissue is a rich source of LD.

Alkaline phosphatase (ALP) is an enzyme that removes phosphate groups from different types of molecules. It is present in the cells lining the biliary ducts of the liver. ALP levels can for instance be increased when there is an obstruction in the bile duct.

Gamma-glutamyltransferase (GGT) is another type of enzyme (transferase) produced primarily by the liver. It is involved in the transport of amino acids and peptides, as well catalyzation of glutathione (the main antioxidant in the body). GGT is more specific to the liver than ALP and can diagnose liver damage but not determine the cause.

2.2.3 Kidney function

Urea is a waste product of the urea cycle produced by the liver. Healthy kidneys filter urea from the bloodstream. If the blood urea concentration is too high, this can indicate that the glomerular filtration rate (GFR) is too low, which in turn suggests kidney failure.

Creatinine is a waste product that comes from muscle activity. Since the kidneys are responsible for removing creatinine from the blood, the creatinine concentration is inversely related to the glomerular filtration rate. Note that this is analogous to urea.

Potassium is an essential electrolyte in the human body that is typically regulated by the kidneys.

Sodium like potassium is an essential electrolyte and a part of salt that is filtered from the blood by kidneys. If kidney function is impaired, sodium levels can rise.

Calcium like potassium and sodium is an essential electrolyte. Most of the body's calcium is stored in bones, only the circulating calcium in the blood is measured.

Phosphate is an ion that contains the mineral phosphorus. Most of the phosphorus contained in phosphate is found in bones, the rest in tissues throughout the body. Kidneys help control the amount of phosphate in the blood.

Albumin is the most common protein found in the blood. It transports nutrients and hormones, provides proteins and prevents fluid from leaking out of blood vessels. A decreased albumin can indicate liver or kidney problems.

2.2.4 Inflammation

C-reactive protein (CRP) is an acute-phase protein made by the liver and released after injury, inflammation or the start of an infection. CRP is synthesized in response to factors released by macrophages and fat cells. Therefore CRP can be used to detect or monitor if there is significant inflammation.

2.2.5 Lipid spectrum

Cholesterol is the total cholesterol in the blood. It is given by the sum of LDL, HDL and VLDL-cholesterol, where VLDL is the carrying component for triglycerides.

High-density lipoprotein cholesterol (HDL-cholesterol), is one of the lipoproteins that transports cholesterol through the body. HDL is also known as "good" cholesterol because it transports cholesterol from the cells and tissue back to the liver.

total/high-density cholesterol ratio is the total-cholesterol divided by the HDL-cholesterol. This is better risk predictor for heart disease than total cholesterol or LDL-cholesterol.

Low-density lipoprotein cholesterol (LDL-cholesterol), is one of the lipoproteins that transports cholesterol through the body. LDL is also known as "bad" cholesterol because when blood contains more LDL than utilized by the body's cells, LDL cholesterol accumulates and can result in the formation of plaques.

Triglycerides are a different form of lipids than cholesterol which store unused calories and provide the body with energy when required.

2.2.6 Coagulation

Prothrombin time (PT) measures the time that it takes for blood to clot. Hence, if the PT is too long, it takes too long for the blood to form a clot. Prothrombin is part of the coagulation cascade.

2.2.7 Glucose metabolism

Hemoglobin A1c (HbA1c) is the amount of glycated hemoglobin, also called hemoglobin A1c. This form of hemoglobin is bound to glucose and reflects the average blood glucose levels over the past three months. HbA1c is a routine test for diabetes where a HbA1c > 6% (or > 42 mmol/mol) is considered as diabetic.

Glucose is the blood sugar level. Note that the blood sugar level naturally varies throughout the day and is influenced by many factors.

Insulin is a hormone produced in the pancreas and released by β -cells in the islets of Langerhans. Insulin promotes the uptake of glucose and, like glucose, varies throughout the day.

C-peptide is a useful marker for insulin production because pro-insulin splits into one molecule insulin and one molecule C-peptide. It may be useful to determine how much insulin a diabetic patient's pancreas is still producing (endogenous insulin).

2.2.8 Thyroid function

Parathyroid hormone (PTH) is produced by the parathyroid glands and its production is stimulated by low calcium levels. PTH results in release of calcium in the bloodstream from the bones and also resorption from the kidneys and excretion of phosphorus.

Thyroid-stimulating hormone (TSH) is produced by the pituitary gland and is used by the body to maintain stable amounts of thyroid hormones through a feedback loop.

free T4 (fT4) is used in conjunction with TSH to distinguish between different thyroid disorders.

Cortisol is made by the adrenal glands and its production is stimulated by the adrenocorticotrophic hormone (ACTH) that is produced by the pituitary glands. Therefore this test can show problems with the adrenal and pituitary glands. However, cortisol has a large variation throughout the day, so timing of the cortisol test is important.

2.2.9 Minerals and vitamins

Iron is the amount of iron in the blood.

Ferritin is the primary form of iron storage in cells and acts as a buffer against iron overload and iron deficiency.

Folic acid is also known as folate and cannot be produced by the body itself, but only supplied by the diet.

Zinc is an essential element with important functions throughout the body. Therefore a lack of zinc has numerous clinical manifestations.

Magnesium is often low if there is excessive excretion by the kidneys or impaired absorption in the intestine. Levels can also be high if excretion is impaired.

Vitamins A, B1, B6, 25-OH D and B12 are measured to detect deficiencies which are common after bariatric surgery as their absorption is limited [33].

2.3 Associations between biomarkers and co-morbidities

Both diabetes and dyslipidemia can be diagnosed from the results of the markers contained in the glucose metabolism and lipid spectrum sub-panels, respectively. However, not much is known about the association between the other markers and diabetes and dyslipidemia. Moreover, hypertension is normally diagnosed from a blood pressure (BP) measurement. While these BP measurements are performed in patients, clinicians reported that they were unreliable (BP measurements were mostly too high because patients were not in a relaxed state) and are therefore not included here. Since hypertension has to be determined indirectly through changes in markers, it is important to determine if there are any known associations between hypertension and any of the markers contained in the bariatric panel. From literature, evidence was found to conclude that hypertension can be observed from changes in several biomarkers (i.e. leukocytes, CRP, urea, GGT, etc.) contained in the bariatric lab panel. Also associations between markers not contained in the glucose metabolism and lipid spectrum were found for diabetes and dyslipidemia. See appendix A for a summary of

the found associations between the biomarkers of the bariatric lab panel and the three co-morbidities of interest.

3 | Materials and Methods

In this chapter the pre-processing of the data and subsequent model building, validation and selection is described.

3.1 Pre-processing

Patients that undergo bariatric surgery at the Catharina Hospital in Eindhoven are screened before surgery and have regularly scheduled follow-ups after surgery. The screening consists of both an examination by the clinician and a blood test. A follow-up consists of either an examination by the clinician, or an examination and a blood test. The protocol for bariatric patients at the Catharina Hospital in Eindhoven is summarized in figure 3.1. At each visit a summary of the findings by the clinician and the

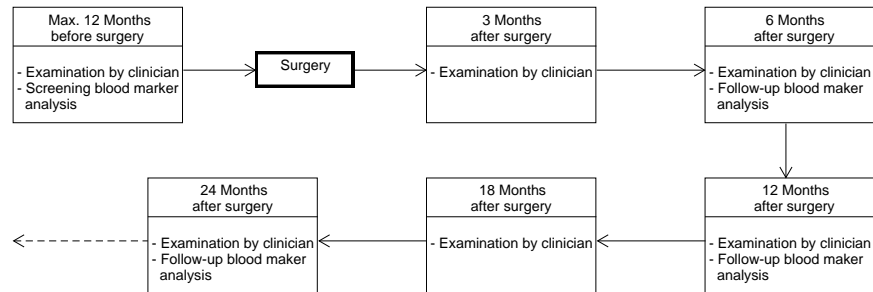


Figure 3.1: Flow diagram for repeated visits for bariatric patients.

results of the blood test (if performed) are stored in multiple databases. In this section it is described how data extracted from these databases were linked and how the merged dataset was subsequently analyzed for missingness and cleaned. All data of patients that underwent *primary* gastric sleeve or bypass surgery (no revision cases) at the Catharina Hospital in Eindhoven, in the period of 14-02-007 to 24-02-2016 were extracted from multiple databases.

3.1.1 The DATO dataset

The Dutch Audit for Treatment of Obesity (DATO) database is one of the databases where data is extracted from. The DATO database is a national database that contains registrations regarding the pre- and post-treatment health status of patients that underwent bariatric surgery in the Netherlands. Examination of patients is done by clinicians at the Catharina Hospital in Eindhoven and results are retrospectively entered in the DATO database. This registration also covers the status of co-morbidities among which diabetes, hypertension and dyslipidemia (aside from co-morbidities other information is registered such as a patient's height, weight, smoking status, drinking status, medication use, type of surgery, complications after surgery, etc.). Before surgery, diabetes, hypertension and dyslipidemia are labeled as either "Yes" or "No" depending on whether the co-morbidity is known to be present in the patient. After surgery the co-morbidities are labeled as "Cured", "Improved", "Same", "Worse", "Denovo" or "Not present". The criteria for these labels as formulated by DATO are given below.

Diabetes follow-up label criteria:

Cured HbA1c < 42 mmol/mol without medication or a 50% reduction in the use of anti-diabetics.

Improved more than 11 mmol/mol reduction in HbA1C with reduction in use of anti-diabetics (50% reduction or no longer requiring oral or insulin medication).

Same no remission or improvement as described above.

Worse HbA1c > 48 mmol/mol or restarting use of medication after partial or complete remission.

Denovo diabetes diagnosed and not present before surgery.

Not present diabetes is or was not present.

Hypertension follow-up label criteria:

Cured pre-hypertensive or normotensive (120-140/80-89 without medication)

Improved 50% reduction in use of hypotensive medications or reduction by one prescribed hypotensive medication or normotensive on same medication.

Same no remission or improvement as described above.

Worse increase or restarting use of medication after increase or re-occurrence of hypertension after complete remission.

Denovo hypertension diagnosed and not present before surgery.

Not present hypertension is or was not present.

Dyslipidemia follow-up label criteria:

Cured normalized LDL/HDL-ratio without medication.

Improved reduction in number of medications or dosage of lipid-lowering medications at equal LDL/HDL-ratio or equal number of medications at reduced LDL/HDL-ratio.

Same no remission or improvement as described above.

Worse increase or restarting use of medication after increase or re-occurrence of too high LDL/HDL-ratio.

Denovo dyslipidemia diagnosed and not present before surgery.

Not present dyslipidemia is or was not present.

3.1.2 Laboratory dataset

The second database where data is extracted from is the laboratory database, stored in the electronic health records. On 01-10-2011 the Catharina Hospital in Eindhoven introduced a bariatric laboratory panel containing a set of blood tests with (bio)markers (see section 2.2 for more details). Pre-operatively the panel consists of 47 markers and a subset of 42 markers is measured during follow-up (6, 12, 24, 36 months after surgery). All the lab measurements stored in electronic health records of patients that underwent bariatric surgery were extracted.

3.1.3 Merging datasets

After completing the database extractions as outlined above, two datasets are obtained containing information from different sources but from the same patients. These datasets were merged into a single dataset that combines, from both datasets, the relevant information for building our model. The process of merging both datasets is outlined below. For these steps a Java program created in NetBeans IDE 8.1 is used.

1. Imported all patients that underwent bariatric surgery from the DATO dataset.
2. Removed patients that underwent surgery before 01-01-2012.¹
3. Imported DATO entries and mapped them to patients by their unique patient ID.²
4. Imported lab measurements and grouped them in sets that were performed on the same patient and at the same date and time. Mapped these sets to patients by the unique patient ID.

¹Although the DATO database contains patients that underwent surgery as far back as 2007, we included only patients that underwent surgery after 01-01-2012 because the bariatric lab panel was introduced on 01-10-2011 and most patients have their screening 3 months prior to surgery.

²Some manual data entry errors became obvious during importing: 2 patients had two or more screenings on the same date, 3 patients had two or more surgeries on the same day and 76 patients had two or more follow-ups on the same day. Due to conflicting information these entries were not imported.

5. For each patient DATO entries were matched to *bariatric* lab sets (i.e. lab sets that contain the subset of 42 markers included in the bariatric panel) into lab-DATO-pairs, based on the date:

(a) For screening:

- i. Subtracted one year from the date of surgery.
- ii. Obtained all the bariatric lab sets and DATO screening entries starting from 1 year before surgery until the date of surgery.
- iii. From the set of lab sets and the set of DATO entries, tried all possible pairs between a lab set and DATO-entry. For each pair, the number of days between the lab set and DATO entry was calculated.
- iv. Allowed 90 days as a maximum time window between a lab set and DATO entry, pairs that had more than 90 days between the lab and DATO date were dropped.³
- v. From the remaining pairs chose the pair that had the least amount of days between the lab set and DATO entry as our screening lab-DATO-pair.⁴

(b) For each follow-up period:

- i. Determined when the scheduled follow-up date should occur, i.e. add 6, 12 or 24 months to the date of surgery.
- ii. Obtained all the lab sets and DATO entries within 3 months before and 3 months after the scheduled follow-up date.
- iii. Analogous to screening
- iv. Analogous to screening
- v. Analogous to screening

With this approach a lab-DATO-pair is obtained for each follow-up moment that the patient attended.

6. Removed patients that did not have a screening lab-DATO-pair. This pair had to be present because patients had to be compared to their pre-operative (baseline) status, both in terms of lab measurements and co-morbidity status.

Although there are follow-ups at 36, 48 and 60 months, only the 6, 12 and 24 month follow-ups are included because not enough data was available of the 36- and 48-month follow-up (see figure 3.2). It is also possible for patients to have multiple DATO entries and/or multiple lab sets within a follow-up time window. With this approach only one pair (i.e. the pair that is closest together in time) is included. Now that all lab-DATO-pairs are known, the databases can be merged into one single dataset that can be used to build a model on. The merged dataset is generated with the following procedure:

³The 90 day limit is chosen after consulting with clinicians who indicated that three months is a realistic time window for patients to complete their blood test and appointment with the clinician.

⁴This combination best represents the state at that time.

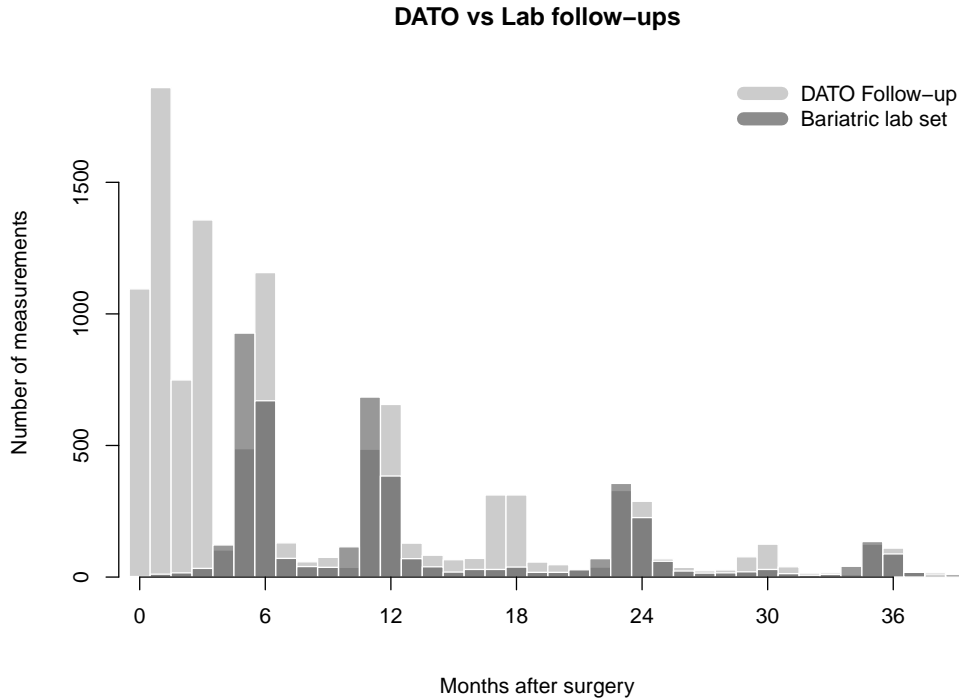


Figure 3.2: Number of bariatric lab sets and DATO follow-up entries versus months after surgery. The spread at the 6, 12 and 24 month follow-up periods indicates that some patients have their follow a few months before and some a few months after the scheduled date. Also observe that there are no lab follow-up measurements at 3, 18 and 30 months, this is in agreement with the protocol shown in figure 3.1.

1. Looped through all screening lab-DATO-pairs and printed on each row the patient ID, a selection of relevant DATO variables and the results of the lab measurements that are contained in the bariatric lab panel. The relevant DATO variables that were included are: type of surgery, date of surgery, gender, age at surgery, weight, height, diabetes status, hypertension status and dyslipidemia status.
2. Do the same for all lab-DATO-pairs found at 6, 12 and 24 months follow-up. The 18 and 36 month follow-ups were not included because no lab measurement is performed at these visits (see figures 3.2 and 3.1).
3. Combined the four datasets in a single dataset and sorted by patient ID so that all measurements were grouped by patient as shown in figure 3.3.

The final merged dataset contained a total of 6003 records and 2658 patients. Subsequent analysis is done in R version 3.3.1 [34].

DATO							Lab			
PatientID	Period	Sex	Age at surg.	Weight	Hypertension	...	Urea	Glucose	Cholesterol	...
2473337880	Pre-Op	F	44.7	143	Yes	...	4.8	5.5	5.90	...
2473337880	6 M	F	44.7	99	Improved	...	4.3	4.1	5.59	...
2473337880	12 M	F	44.7	95	Improved	...	5.3	5.4	5.71	...
7070366830	Pre-Op	F	36.7	138.2	No	...	2.9	5.3	5.31	...
46675430840	Pre-Op	M	60.7	178	Yes	...	4.5	7.3	4.85	...
...

Figure 3.3: Layout of merged dataset, each row represents a patient record.

3.1.4 Missingness

The missingness is examined by plotting the number of missing values for each variable as shown in figure 3.4.

Missingness in laboratory data

Where the missingness ranged from $\geq 50\%$ to 95%, these markers were dropped from the dataset. The following markers had a high number of missing values: vitA, zinc, fT4, MMA, insulin, C-peptide, magnesium, cortisol and TSH.

- Zinc and vitamin A were at some point in time removed from the bariatric panel and were not included in the follow-up bariatric panel.
- fT4 (free T4) is part of a thyroid function test and is only measured when TSH falls outside the reference range.
- The methylmalonic acid (MMA) test can detect vitamin B12 deficiencies when serum vitamin B12 levels are inconclusive (between 90 and 300 pmol/L) and is consequently only measured when vitamin B12 falls in this range.
- TSH, insulin, C-peptide, magnesium and cortisol are only measured during screening and not included in the follow-up bariatric panel.

LDL has a slightly higher percentage of missing values (1.8%) compared to the other clinical markers. After investigating the cases where LDL results were missing, these results were labeled as "cannot be determined". This turned out to be a result of the fact that LDL is not measured but calculated from the total cholesterol, HDL and triglycerides by the Friedewald equation [35]:

$$\text{LDL-cholesterol} = \text{total-cholesterol} - \text{HDL-cholesterol} - 0.45 \times \text{triglycerides}$$

However, the Friedewald equation is not valid for high triglycerides (above 400 mg/dL or 4.52 mmol/L [36]). If the Friedewald equation would be applied to hypertriglyceridemic patients it would result in a negative LDL value, instead the lab returns "cannot be determined". Although there are relatively few cases these could potentially be patients with co-morbidities that affect a model's predictive ability. This is a type of

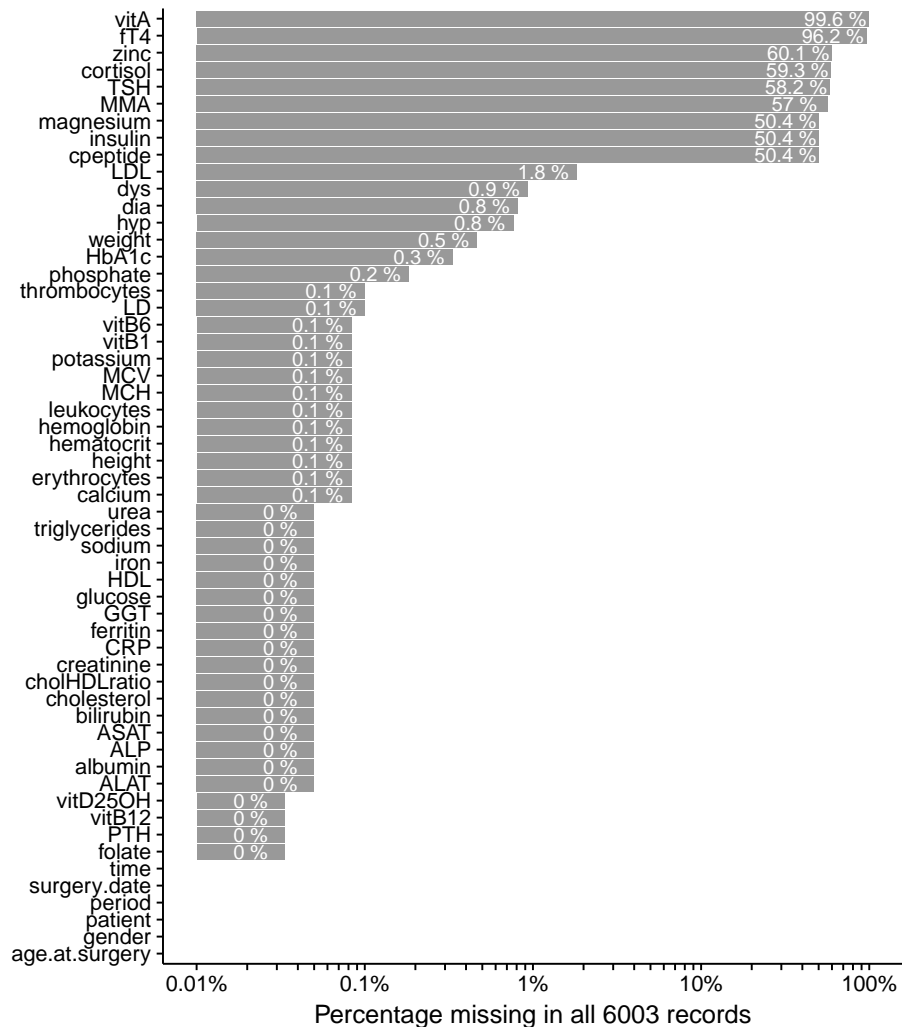


Figure 3.4: The majority of variables have only a few missing values ($\leq 2\%$), a subset of markers (vitA, zinc, fT4, MMA, insulin, cpeptide, magnesium, cortisol, TSH) have $\geq 50\%$ missing values.

systematic missigness where the missing value can be explained by other variables in the data. Therefore, instead of dropping these cases, LDL as a marker is dropped. The remaining clinical markers contain only few missing values which did not occur independently but are the result of a few cases in which the entire set of measurements was missing due to external circumstances e.g. because the sample was lost or destroyed. Imputing such a small number of missing values would not yield a significantly different result, therefore these records are dropped.

Missingness in DATO registrations

Aside from the lab measurements, the DATO registrations also contained missing data. If the diabetes, hypertension or dyslipidemia status were missing, that record was dropped from the dataset because the status labels define the outcome that the model will be fitted on (in machine learning this is called "supervised learning"). It is possible that data is missing at screening but that a complete follow-up is still present, in this case the patient is still excluded from analysis because the screening contains essential information. At screening patients are labeled as "Yes" or "No" with respect to the presence of a co-morbidity. At follow-up a patient can have one of six labels: "Cured", "Improved", "Same", "Worse", "Denovo" or "Not present" as described in section 3.1.3. If the screening label would be missing and the next label would be "Same" it would be impossible to determine whether the co-morbidity is still present or never was.

Missingness due to attrition

Since our data is in a longitudinal form that is most commonly seen in so called cohort studies⁵, there is also missingness in the form of patients that drop out after screening or a certain follow-up period. In cohort studies the phenomenon of drop-out is also called *attrition* and is important to examine because it can lead to attrition bias (or loss of follow-up) if the patients that drop out are different from patients that remain in the study [38]. A drop-out is defined as a patient that misses a follow-up appointment and does not return after this missed appointment. Note that a patient that missed their 6 and 12 month follow-up but did return for their 24 month follow-up is by this definition not a drop-out. These patients are sometimes referred to as *intermittent*. This definition is applied to the data and the results are plotted for each year of surgery in figure 3.5. In this data a patient had at least one visit (screening) and at most 4 visits (screening, 6, 12 and 24 month follow-up). On average a patient had 2.35 visits and a median of 2 visits. If only patients that had completed their 24-month follow-up were included (i.e. doing a *complete* case analysis) roughly 75% of patients would be excluded, therefore the choice is made to do an *available* case analysis where all the records that are present for each patient are included and retain patients that are only partially observed. Note that the high drop-out rate in the data is mostly a result of the fact that the data extracted contains a "snapshot" of the data that was available at that time. Patients that had underwent surgery one month prior to the extraction of the data could not have completed any follow-up and were counted as drop-out (see explanation below figure 3.5). Also outpatients are in this group, these are patients who undergo surgery at the Catharina Hospital but the follow-up occurs at a different hospital.

3.1.5 Data cleaning

Both the laboratory results and DATO registrations required cleaning. Data cleaning is the process of detecting corrupt or inaccurate records and correcting or removing these from the data.

⁵A cohort study is defined as: "a group of people with defined characteristics who are followed up to determine incidence of, or mortality from, some specific disease, all causes of death, or some other outcome." [37]

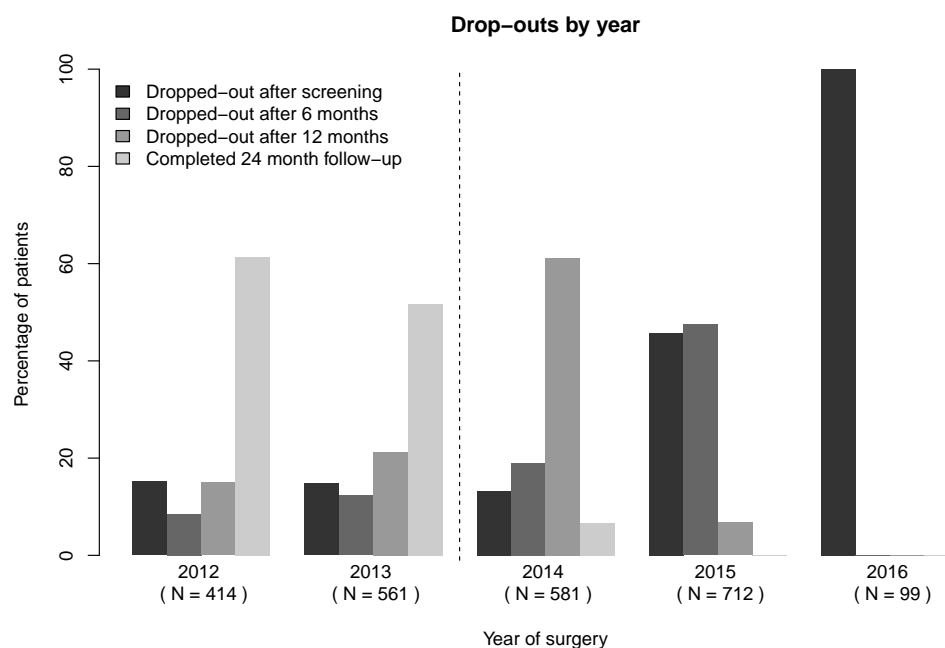


Figure 3.5: Shows how many patients drop-out by year. The dashed line indicates that patients from years 2012 and 2013 have had enough time to complete the 24 month follow-up. To explain: since data was extracted at 24-02-2016, patients that underwent surgery after 20-02-2014 did most likely not yet have a 24-month follow-up, patients that underwent surgery after 20-02-2015 did not yet have a 12- and 24-month follow up, etc. A representative drop-out pattern can only be concluded from the years 2012 and 2013.

Cleaning of laboratory data

In the clinical laboratory truncated results are sometimes reported. There are several markers that, when the result falls outside the measurement range or is not clinically relevant, are truncated. All markers that are truncated in the dataset can be seen in table 3.1. Note that while most markers are truncated only rarely, CRP was truncated in over 67% of the cases. CRP is an inflammation marker and values < 6 are considered normal (i.e. no inflammation present). In addition, the measurement uncertainty in low ranges becomes relatively large as the equipment is not calibrated in this measurement region. Only values ≥ 6 are reported untruncated. CRP is dichotomized in < 6 and ≥ 6 , although it was possible to create more categories (i.e. $\geq 6 - 10$ and ≥ 10), these categories would become too small to have any statistical power. For all other truncated lab measurements the $>$ or $<$ sign is simply removed since they occur so rarely (less than 5%) that it is not expected to affect the final result.

Cleaning of DATO registrations

The DATO labels are retrospectively entered by supporting staff and are susceptible to data entry errors, clinicians estimated these to be at around 5%. The weight of a patient

Marker	Truncated at	Number of occurrences (Total = 6003)
erythrocytes	<7.0	1
MCH	<1.10	1
ASAT	<5.0	1
	<50	2
ALAT	<5.0	1
CRP	<5.0	5
	<6.0	4011
folate	>45	252
vitamin B1	>1000	5
vitamin B6	>1000	21
	>300	5
vitamin D (25OH)	<10	16

Table 3.1: Markers that are truncated by the clinical lab.

is also manually entered and by inspection of the distribution of the weight values one obvious data entry was found: one patient had a weight of 632.2 kg at the 24 month follow-up. This weight was manually corrected to 63.2 kg. Another form of data entry error was observed by looking at subsequent labels for each patient; after a patient is cured of a co-morbidity (or the co-morbidity was never present) the next label cannot be "Improved". This may seem like a minor issue but if a patient has the label "Improved", in a formal sense the co-morbidity is still present. Since it cannot be determined which label is true (was the patient really cured at a previous measurement or was it merely an improvement that was mislabeled as cured?), patients with an inconsistent follow-up label are dropped.

3.1.6 Variable selection

In total there are 49 predictor variables, however not all of them are relevant to the outcome and a marker like LDL is not used because of the systematic missingness. Dropping irrelevant features enhances generalization and makes the final model easier to interpret. Moreover, it prevents "oversearching" and the discovery of spurious correlations during modeling. After consulting with clinicians and applying domain knowledge, some predictor variables were dropped, see table 3.2.

3.1.7 Variable extraction

In addition to dropping predictors new features are also created from the original predictors by using domain knowledge, this process is also referred to as *feature extraction*

Dropped predictor(s)	Motivation
weight, height	BMI incorporates both weight and height, height is correlated with gender
all vitamins	Vitamins are supplemented, folate kept as compliance marker §
LDL	Calculated by Friedewald equation, not valid for high triglycerides (see also section 3.1.4)
iron	Biological variation too high
year of surgery	Of no value to future predictions
type of surgery	Not yet known at screening
months before/after surgery	Categorized in "period" variable

Table 3.2: § Since folate is supplemented after surgery, folate deficiencies are found in patients that do not take supplements, i.e. are not compliant. Therefore folate serves as a proxy for compliance.

in machine learning:

- Replaced the measurement of serum creatinine by an estimation of glomerular filtration rate (GFR) as calculated by the CKD-EPI (Chronic Kidney Disease Epidemiology Collaboration) estimation equation [39]:

$$\text{CKD-EPI} = 141 \times \min(\text{Scr}/\kappa, 1)^\alpha \times \max(\text{Scr}/\kappa, 1)^{-1.209} \times 0.993^{\text{age}} \times \beta$$

$$\kappa = 61.9 \text{ if female}$$

$$\kappa = 79.6 \text{ if male}$$

$$\alpha = -0.329 \text{ if female}$$

$$\alpha = -0.411 \text{ if male}$$

$$\min = \text{The minimum of } \text{Scr}/\kappa \text{ or } 1$$

$$\max = \text{The maximum of } \text{Scr}/\kappa \text{ or } 1$$

$$\text{Scr} = \text{Serum creatinine in } \mu\text{mol/L}$$

$$\beta = 1.018 \text{ if female}$$

$$\beta = 1 \text{ if male}$$

GFR is clinically more meaningful because it is a direct measure of kidney function and CKD-EPI is less biased than the MDRD (Modification of Diet in Renal Disease) and Cockcroft-Gault equations [40]. Moreover, the interpretation of the serum creatinine level depends on age and gender.

- Introduced the ASAT/ALAT-ratio, this is simply the ratio between the concentrations of ASAT and ALAT. This ratio provides information about the cause of liver damage or hepatotoxicity [41].
- The prothrombin time (PT) is converted to the international normalized ratio (INR) with the following formula:

$$\text{INR} = \left(\frac{\text{PT}_{\text{meas}}}{\text{PT}_{\text{normal}}} \right)^{\text{ISI}}$$

PT_{meas} = Prothrombin time as measured in patient

$\text{PT}_{\text{normal}}$ = Mean normal PT (= 3,9 s)

ISI = International Sensitivity Index (= 0.935)

The prothrombin time of a normal individual will vary according to the type of analytical system employed. This is due to the variations between manufacturers that are used to perform the test. The INR was devised to standardize the results.

- Calcium in serum is bound to proteins, mainly albumin. As a result the measured calcium concentration in patients with low or high albumin levels may not accurately reflect the free calcium concentration that is physiologically important for cellular function. The following calcium correction formula is conventionally used for calculating a corrected or free calcium concentration [42]:

$$\text{corrected calcium} = \text{total calcium} + 0.02 \times (40 - \text{albumin})$$

$$\text{corrected calcium} = \text{free calcium in serum in mmol/L}$$

$$\text{total calcium} = \text{measured calcium in serum in mmol/L}$$

$$\text{albumin} = \text{albumin in serum in g/L}$$

The measured calcium concentration is replaced with the free/corrected calcium concentration.

3.1.8 Data description

164 patients were dropped because they did not have a screening lab-DATO-pair (see section 3.1.3), 33 patients were dropped because they had missing values in either their lab or DATO variables at screening (see section 3.1.4) and 94 patients were dropped because they had an inconsistent follow-up labels (see section 3.1.5). The result of preprocessing in terms of included and excluded patients is summarized in figure 3.6. Since the patient groups differ per period, the patient demographics are represented for each period in the results. The prevalence of all combinations of co-morbidities before surgery were plotted in an Euler diagram using the `eulerr` package to show the union and intersection of co-morbidities [43].

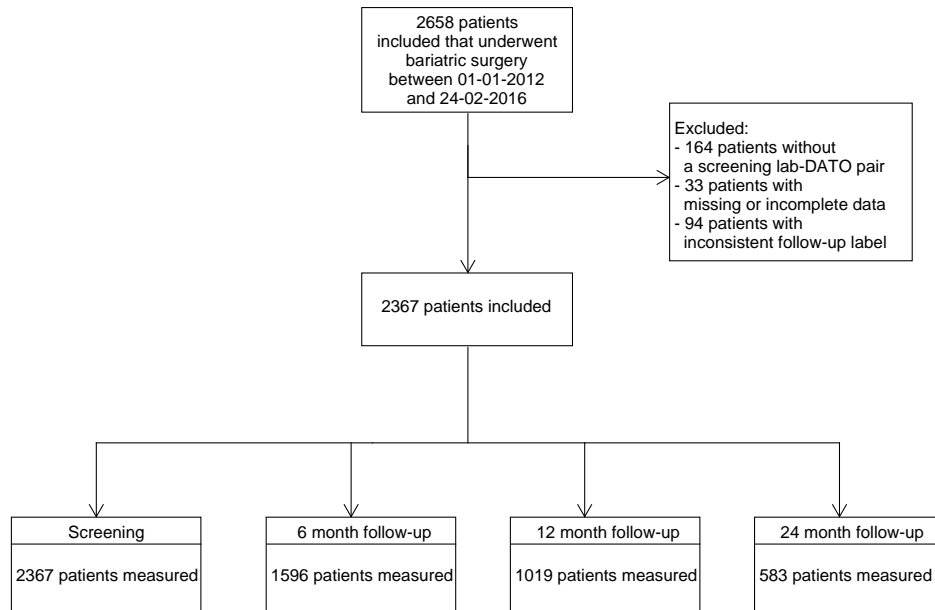


Figure 3.6: Number of patients included, stratified over each time-period. Missing or incomplete data refers to patients that have no complete measurement at any time period, or patients that have complete measurements at 6, 12 or 24 months but no complete screening measurement.

After variable selection and extraction (described in sections 3.2.4 and 3.1.7) a total of 38 predictor variables are included. The distributions of all the predictor variables and the statistical description of all the variables in the model are summarized by using the `describe` function in the `Hmisc` package [44].

3.2 Modeling

After the pre-processing is completed a dataset containing 5564 records of 2367 patients are obtained, this dataset is used for modeling. The steps that were followed are outlined in figure 3.7. The predictor variables are defined in the previous section, the other modeling steps are described below.

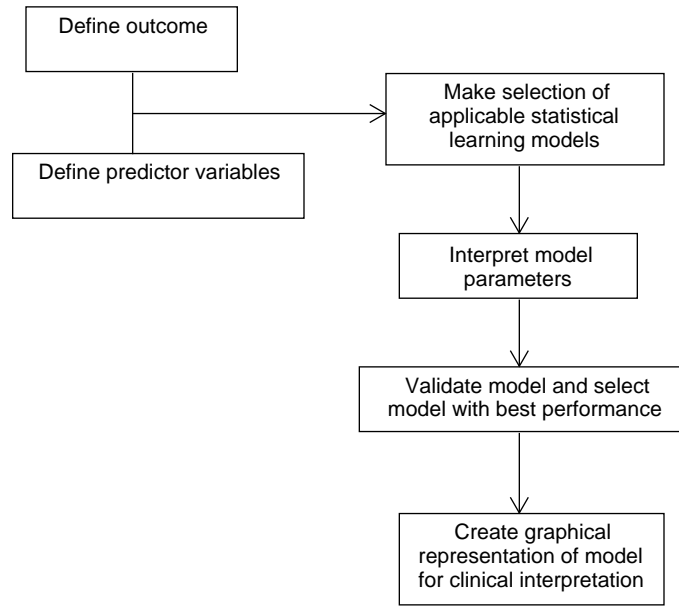


Figure 3.7: Model building and validation steps.

3.2.1 Ordinal outcome definition

The outcome in the dataset is given by the three binary (TRUE or FALSE) labels for diabetes, hypertension and dyslipidemia. In machine learning this is described as a multi-label classification problem [45]. A straightforward approach to solving a multi-label classification problem is the binary-relevance (BR) method [45]. In this method a classifier is trained independently for each label and given, a new unseen patient, each classifier predicts whether the respective co-morbidity is present. The BR method assumes label independence. Because our aim is to predict the true health state of bariatric patients, having three separate models for each of the three (possibly correlated) co-morbidities would not result in a single co-morbidity severity score. To obtain a single co-morbidity severity score an outcome based on the *number of co-morbidities present* is defined. Since our focus lies on diabetes, hypertension and dyslipidemia, a patient can be characterized as having 0, 1, 2 or 3 co-morbidities. The assumption is made that patients with more co-morbidities also have a higher co-morbidity severity

than patients with fewer or no co-morbidities. Therefore we defined the outcome on an *ordinal* scale. An ordinal variable is a type of categorical variable that has a natural ordering of categories, however, the distances between the categories is not known [46]. Patients with two and three co-morbidities are merged into a single "Multiple" category to create more balanced categories. The ordinal outcome definition is shown in table 3.3.

Outcome Level (Y)	Definition	N (at screening)
None	No co-morbidity	1329
One	Diabetes or hypertension or dyslipidemia.	569
Multiple	Diabetes & hypertension or diabetes & dyslipidemia or hypertension & dyslipidemia or diabetes & hypertension & dyslipidemia	469

Table 3.3: Ordinal outcome definition

Ordinal label assignment

Before the individual labels are converted to an ordinal label, the follow-up label "Same" is replaced by the actual previous label. For example if the previous label was "Cured" and the next label was "Same", "Same" is converted to "Cured", and if the previous label was "Yes" then the "Same" label is also converted to "Yes". Next an ordinal label assignment algorithm is applied as outlined in algorithm ?? . This algorithm assigned the ordinal outcome label to all the patient records.

3.2.2 Statistical learning models

After the predictors and ordinal outcome was determined, a statistical learning model is selected. Since a clinical prediction model is developed, the choice is restricted to white-box models, *predictive accuracy is as important as understandability of the process that leads to it*. Logistic regression models are the models of choice in many clinical data classification tasks [47]. Logistic regression belongs to the family of generalized linear models (GLM) introduced by Nelder & Wedderburn in 1972 [48]. Logistic regression models are flexible because they can incorporate continuous and categorical variables, non-linear transformations and interaction terms. Also the continuous variables do not need to be standardized before fitting. See appendix C for a general explanation of logistic regression and the process called maximum likelihood estimation (MLE) that is used to estimate the regression coefficients. All functions used in the model fitting procedure are part of the *rms* package by Harrell [49].

```

for each patient record do
  for each co-morbidity label do
    if label equals "Yes" or "Improved" or "Denovo" or "Worse" then
      | we count the co-morbidity as present;
    end
  end
  we have obtained the number of co-morbidities that are present;
  if this number equals 0 then
    | we assign the label "None";
  end
  if this number equals 1 then
    | we assign the label "One";
  end
  if this number is greater than 1 then
    | we assign the label "Multiple";
  end
end

```

Algorithm 1: Label conversion algorithm used to convert labels to ordinal outcome.

Correlations

One assumption of logistic regression is that the model should have little or no multicollinearity. Multicollinearity is a phenomenon in which predictor variables in a model are highly correlated. Since correlations are expected between some predictors e.g. ASAT and ALAT, hemoglobin and hematocrit, MCH and MCV, a heterogeneous correlation matrix is calculated to quantify the correlations between all predictor variables. This matrix consists of Pearson correlations between numeric variables, polyserial correlations between numeric and ordinal variables, and polychoric correlations between ordinal variables. This matrix is calculated for all measurements before surgery because independent cases are a requirement. The function `hetcor` in the `polycor` package was used for this [50]. This matrix is plotted with the `corrplot` package [51].

Interaction terms

When two predictors of interest interact, the relationship between each of the interacting predictors and the outcome depends on the value of the other interacting predictor.⁶ Because it is known that co-morbidity is an inclusion criteria for patients with a BMI < 35, it is expected that the prevalence of co-morbidities is higher in patients with

⁶Example: We define the disease status as the outcome and use age and gender as predictors. If only older males are susceptible to the disease, age and gender have an interacting effect. Age is not by itself related to disease status, nor is gender, but combined they are.

a BMI < 35 *before* surgery. After surgery, when patients loose weight, the category of patients with a BMI < 35 also consists of patients that lost weight but had no co-morbidities before surgery. Therefore an interaction between the period (i.e. before or 6-, 12-, or 24-months after surgery) and BMI is included.

3.2.3 Logistic regression

Before fitting an ordinal regression model on the ordinal outcome, three logistic regression models are fitted to each of the co-morbidities separately. While these models are not used to obtain the co-morbidity score, they allow for determination of which predictors are significantly associated with each of the co-morbidities individually. These associations are assessed with the Wald statistics (see appendix C "Hypothesis tests" for an explanation). Note that an assumption of logistic regression is violated in the sense that repeated measurements on the same patient are non-independent where independence is assumed. A misspecification can lead to incorrect variances and covariances, and hence confidence intervals and Wald tests. To adjust the variance-covariance matrix of the fitted logistic regression models to correct for repeated measures, the Huber-White method is used as implemented in the `robcov` function.

3.2.4 Proportional odds and continuation ratio

For an ordinal outcome two different types of ordinal logistic regression models are commonly used: proportional odds (PO) and continuation ratio (CR). The first of the models, the PO, is the most commonly used. Different models have different ways of incorporating the ordering of categories through the use of a different dichotomization of the ordinal outcome, we refer to appendix D for an explanation. Here both the PO and a penalized extended CR model are fitted. The penalized extended CR model has been described by Harrell [52] and allows for relaxation of the equal slopes assumption that is inherent to ordinal regression models. In addition it applies shrinkage through the use of penalized MLE. For an explanation of penalized MLE see appendix C.6. Before fitting the ordinal logistic regression models the proportional odds assumption is tested by plotting the means of X stratified by the levels of Y and visually inspecting if the means were in consistent order as described by Harrell [52] with the function `plot.x.mean.ordinaly`. Both the PO and the penalized extended CR model were fitted using the `lrm` function. The CR model was extended to allow for unequal slopes with the `cr.setup` function and penalization applied using the `pentrace` function. The optimal penalty was determined by using a grid search with different penalties for main and interactions effects.

Feature selection to obtain final, parsimonious, model

Manual feature selection was applied in section but there are still 38 predictors in the full model. Full models are not very parsimonious and may contain more predictors than we wish to collect in future patients⁷. The most common procedure for making

⁷A parsimonious model is a model that accomplishes a desired level of explanation or prediction with as few predictor variables as possible.

a model more parsimonious is to remove variables that are not significantly associated with the outcome, in machine learning this also called feature selection [53]. In logistic regression this is usually done with stepwise removal of variables based on P-values, i.e. remove all the variables with a P-value > 0.05 . However, Harrell and others have shown that such stepwise feature selection procedures are not recommended [52][54]. Therefore the "model approximation" procedure as described by Harrell [52] is used here. The linear predictor of the full model is approximated with an ordinary least squares regression until a R^2 of 0.95 is reached. The predictors that can approximate the full model with a R^2 of 0.95 are kept and other predictors are not included in the final model. This is done for both the PO and penalized extended CR model.

3.2.5 Cross validation

To assess how both models will perform on an unseen independent dataset, 10-fold cross validation is performed (see appendix F for a general explanation of cross validation). The data is randomly split in 10 roughly equally sized subsets based on *patients* and not *patient records* using the caret package [55]. Note that it is important to split data in patients and not patients records because measurements on the same patient are not independent (see appendix F for more details). In each fold (or iteration) the PO and penalized extended CR model are trained on 9/10th of the data and approximated to obtain a final (parsimonious) model. These final models are then used to predict the outcome on the remaining unseen 1/10th of the data (the independent test set). In each of the 10 folds the predictions of both models on the independent test set are saved and merged after the cross validation is complete. See figure 3.8 for a schematic overview of the cross validation procedure.

3.2.6 Assessment of model fit

Next the cross validated predictions are used to calculate a number of performance measures to validate both models. The performance measures can be subdivided in assessment of discrimination (area under the ROC curve for dichotomous outcome, Somers' D_{xy} for ordinal outcome), calibration (slope and intercept) and overall performance (R^2 and Brier score). We refer to appendix G for an explanation of the used performance measures. While the Somers' D_{xy} is used for the ordinal outcome, the estimated area under the ROC curve (AUC) can be calculated to assess the discriminative ability of the score for different binary responses. The AUC is calculated for different time periods and the three co-morbidities separately, see table 3.4.

3.2.7 Model visualization

After the performance of both models is assessed, one model is chosen. To obtain the final model that can be used in the clinic, the same procedure to fit the model as described above is applied, only now on the entire dataset. This model is then visualized in a nomogram for interpretation of individual predictor effects. The nomogram is plotted using the nomogram function in the rms package. In addition, patients are

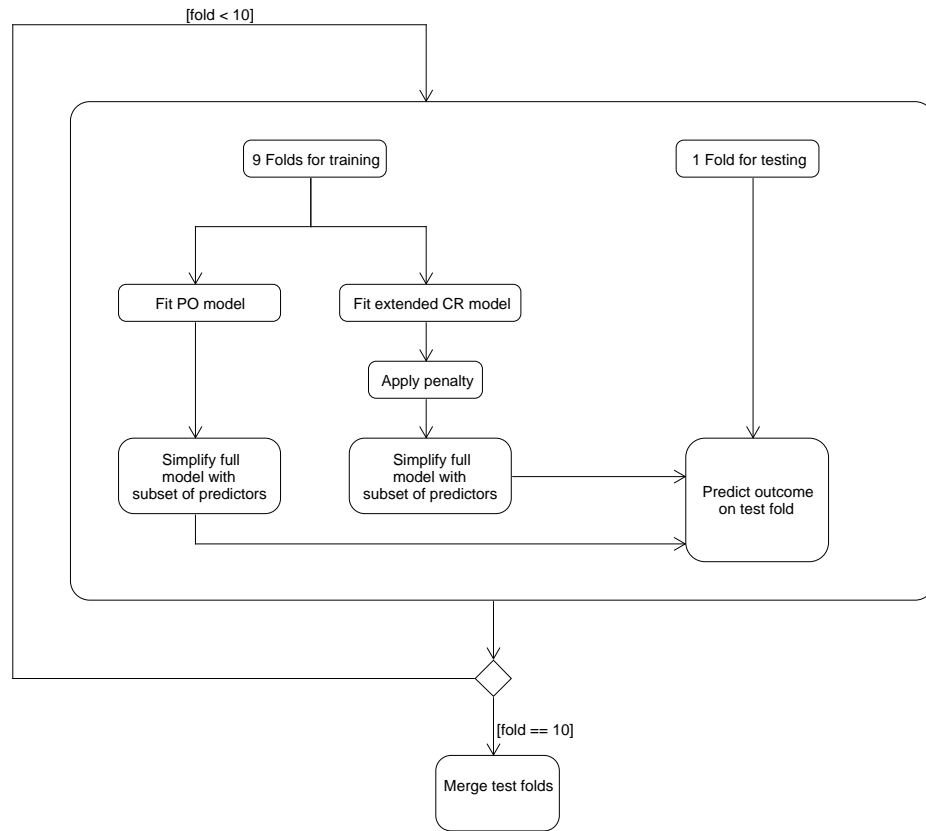


Figure 3.8: Diagram of 10-fold cross validation of PO and penalized extended CR model.

categorized as 0, 1, 2 or 3 co-morbidities before surgery and the mean of the linear predictor for each group was plotted against the time.

Response	Records used	Description	N
One or Multiple	All	TRUE if patient has one or multiple co-morbidities. FALSE if patient does not have any co-morbidities.	3645 1920
One	All	TRUE if patient has one co-morbidity. FALSE if patient has none or multiple co-morbidities.	1124 4441
Multiple	All	TRUE if patient has multiple co-morbidities. FALSE if patient has none or one co-morbidity.	796 4769
One or Multiple at Pre-Op	Before surgery	TRUE if patient has one or multiple co-morbidities. FALSE if patient does not have any co-morbidities.	1329 1038
One or Multiple at 6 Months	6 Months after surgery	TRUE if patient has one or multiple co-morbidities. FALSE if patient does not have any co-morbidities.	1110 486
One or Multiple at 12 Months	12 Months after surgery	TRUE if patient has one or multiple co-morbidities. FALSE if patient does not have any co-morbidities.	758 261
One or Multiple at 24 Months	24 Months after surgery	TRUE if patient has one or multiple co-morbidities. FALSE if patient does not have any co-morbidities.	448 135
Diabetes	All	TRUE if patient has diabetes. FALSE if patient does not have diabetes.	685 4880
Hypertension	All	TRUE if patient has hypertension. FALSE if patient does not have hypertension.	1500 4065
Dyslipidemia	All	TRUE if patient has dyslipidemia. FALSE if patient does not have dyslipidemia.	792 4773

Table 3.4: Different responses used to assess the discriminative ability of the score.

4 | Results

In this section the results of scoring the co-morbidity severity of bariatric patients are presented. First the descriptive statistics of the patient population and the prevalence of co-morbidities are shown. Then, as part of an exploratory data analysis before modeling, the predictor variables and their correlations are shown in a correlation matrix. With these predictors and the ordinal outcome labels the results of the modeling process are shown. This section is subdivided into two parts: First the Wald statistics are shown for three separate binary logistic regression models fitted to the binary outcomes of diabetes, hypertension and dyslipidemia, these are compared to the Wald statistics of two ordinal regression models. Secondly, the performance of both ordinal models are compared by their discriminative ability, calibration and overall performance in a 10-fold cross validation. The results are concluded by visualizing the best performing ordinal model in a nomogram and plotting the scores of our patient population.

4.1 Patient population

The patient demographics table can be seen in table 4.1. Because the patient population varies over time (see figure 3.5) the patient demographics table is split into different time periods.

	At screening (N= 2367)	6 month FUP (N = 1596)	12 month FUP (N= 1019)	24 month FUP (N = 583)
Gender female (%)	1867 (79)	1277 (80)	837 (82)	477 (82)
Age at surgery (years) \pm SD	42.2 \pm 11.2	42.0 \pm 11.1	42.0 \pm 11.0	42.0 \pm 11
Type of surgery (%)	Sleeve: 1270 (54) Bypass: 1097 (46)	Sleeve: 879 (56) Bypass: 717 (44)	Sleeve: 530 (52) Bypass: 489 (48)	Sleeve: 311(53) Bypass: 272 (47)
Height (m) \pm SD	1.69 \pm 0.09	1.69 \pm 0.09	1.69 \pm 0.09	1.69 \pm 0.08
Weight (kg) \pm SD	125.0 \pm 20.5	91.5 \pm 16.5	83.5 \pm 15.5	86.6 \pm 16.40
BMI (kg/m ²) \pm SD	43.6 \pm 5.6	32.2 \pm 4.8	29.2 \pm 4.7	29.9 \pm 5.1
EWL (%) \pm SD	–	64.9% \pm 17.8%	80.2% \pm 20.7%	76.5% \pm 23.0%
Diabetes (%)	Yes: 411 (17) No: 1956 (83)	Yes: 166 (10) No: 1430 (90)	Yes: 73 (7) No: 946 (93)	Yes: 35 (6) No: 548 (94)
Hypertension (%)	Yes: 826 (35) No: 1541 (65)	Yes: 375 (23) No: 1221 (77)	Yes: 196 (19) No: 823 (81)	Yes: 103 (18) No: 480 (82)
Dyslipidemia (%)	Yes: 432 (18) No: 1935 (82)	Yes: 189 (12) No: 1407 (88)	Yes: 108 (11) No: 911 (89)	Yes: 63 (11) No: 520 (89)

Table 4.1: Patient demographics.

4.1.1 Prevalence of co-morbidities

The union and intersections of co-morbidities before surgery are shown in an Euler diagram in figure 4.1. A barplot depicting the prevalence of all the combinations of co-morbidities before and 6, 12 and 24 months after surgery (FUP) is plotted in figure 4.2. Before surgery 56.1% of bariatric patients have no co-morbidity (i.e no diabetes, hypertension or dyslipidemia). Of all the possible combinations of co-morbidities, hypertension (without any other co-morbidities) has the highest prevalence (17.8%), followed by patients that have a combination of all three co-morbidities (6.8%). The combination of diabetes and dyslipidemia is the most uncommon (2.7%). Of all 411 patients that have diabetes before surgery, 80% also have an additional co-morbidity such as hypertension and/or dyslipidemia. Of all 826 patients that have hypertension only 49% has an additional co-morbidity and of all 432 patients that have dyslipidemia before surgery, 85% has an additional co-morbidity. 6 months after surgery the percentage of patients without a co-morbidity increases to 69.5%, a reduction is observed in all combinations of co-morbidities with the strongest reduction in patients with only

hypertension (-4.3%) and the smallest reduction in patients with only dyslipidemia (-0.2%). 12 months after surgery the percentage of patients without a co-morbidity increases slightly to 74.4%, all combinations of co-morbidities show a reduction except for dyslipidemia. 24 months after surgery the total percentage of patients without a co-morbidity is 76.8% (+2.4% increase from the 12 month FUP), overall co-morbidities stay constant with respect to the 12 month FUP. Note that as described in section 3.1.4 the patient population varies over time, as a result there are significantly more females in the 12-month FUP (χ^2 test for equality of proportions P-value = 0.001) and significantly fewer patients with hypertension before surgery in the 24-month FUP population (P = 0.012), see appendix H for full details. Remission rates for diabetes are 36.4% after 6 months, 55.8% after 12 months and 63.5% after 24 months. Remission rates for hypertension are 31.3% after 6 months, 44.8% after 12 months and 44.3% after 24 months. Remission rates for dyslipidemia are 31.0% after 6 months, 38.0% after 12 months and 41.0% after 24 months.

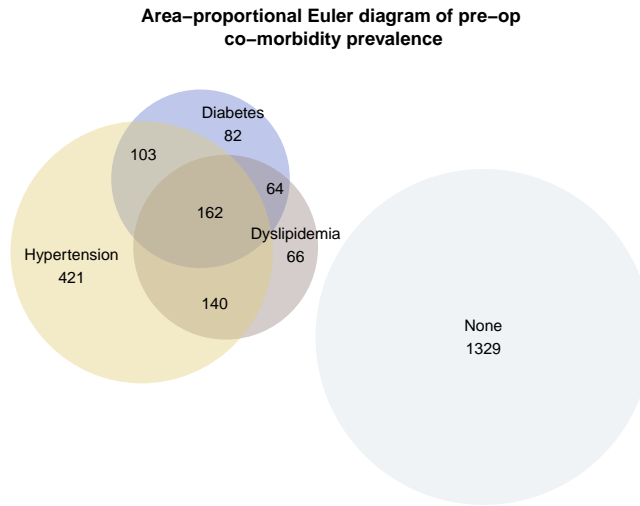


Figure 4.1: Euler diagram of co-occurrence of co-morbidities in bariatric patients before surgery.

4.1.2 Prevalence of ordinal co-morbidity labels

After applying the ordinal co-morbidity labels, "None", "One" or "Multiple" to the entire patient population, the prevalence of these labels is shown in figure 4.3. Note that there are significantly less patients (P = 0.002) with one co-morbidity and significantly more patients (P = 0.009) without co-morbidities before surgery in the 24 month FUP patient population (as determined by a χ^2 test for equality of two proportions, see appendix H). Of the patients that are included in the 24 month FUP (N = 583), 60.9% did not have a co-morbidity before surgery, of the patients that are not included in the

24 month FUP (N = 1784), 54.6% did not have a co-morbidity before surgery (see appendix H for full details).

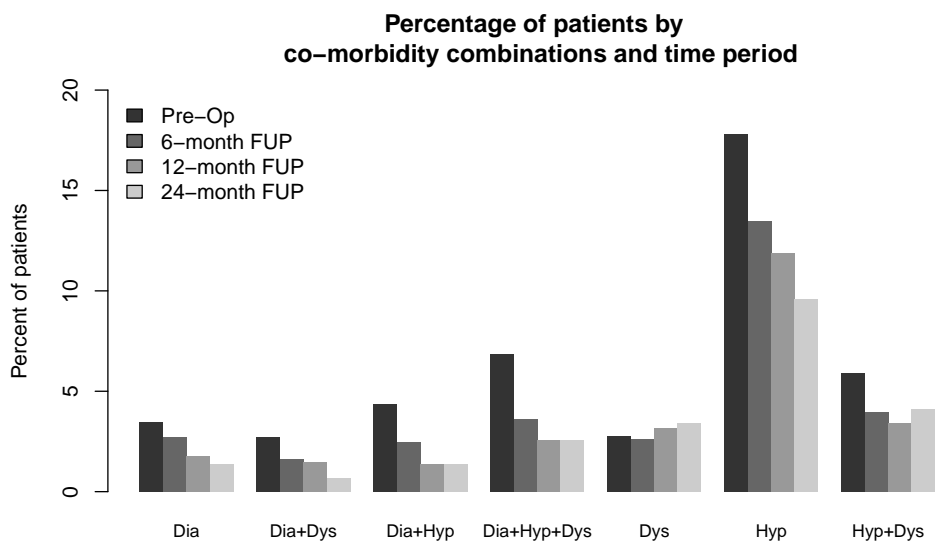


Figure 4.2: Barplot of percentage of patients by combination of co-morbidity and time period, patients without co-morbidities not shown.

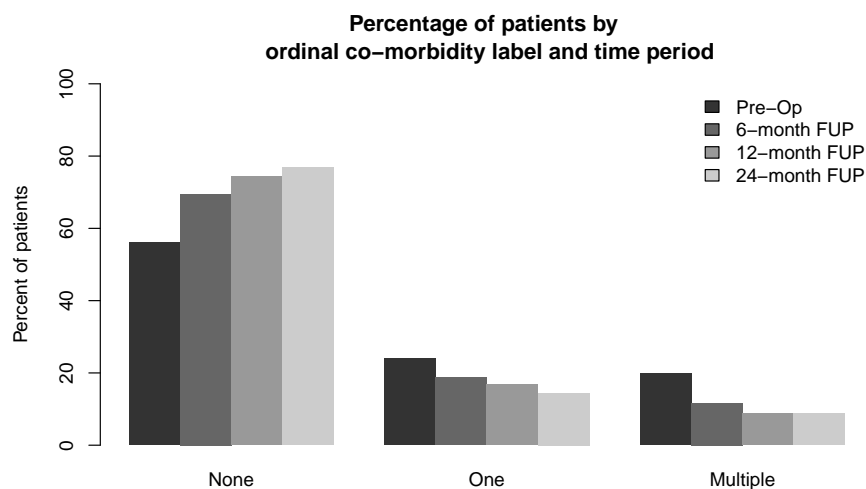


Figure 4.3: Barplot of percentage of patients by ordinal co-morbidity label and time period.

4.1.3 Predictor variables

A concise statistical description of all the predictor variables can be seen in appendix B. All predictors except age have outliers with an absolute Z-score > 3 . Seven predictors even have extreme outliers with Z-scores greater than 20: BMI, ASAT, ALAT, ALP, GGT, triglycerides and PTH. The correlation matrix for all predictors before surgery is plotted in figure 4.4. Strong correlations ($|r| > 0.7$) are observed between: hematocrit and hemoglobin (Pearson $r = 0.93$), MCV and MCH (Pearson $r = 0.89$), ALAT and ASAT (Pearson $r = 0.86$), HbA1c and glucose (Pearson $r = 0.78$), hemoglobin and gender (Biserial $r = -0.74$), erythrocytes and hematocrit (Pearson $r = 0.72$). These strong correlations imply that multicollinearity is present and can make regression coefficients concerning these predictors unstable and difficult to interpret.

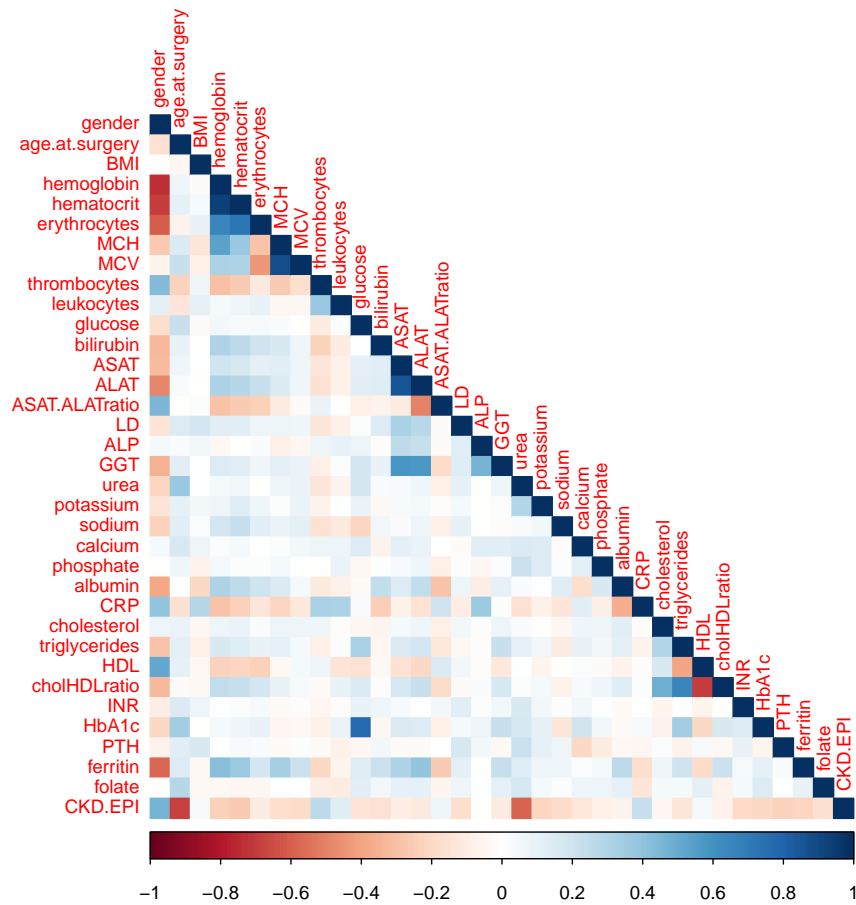


Figure 4.4: Heterogeneous correlation matrix of predictors before surgery. Pearson correlations between numeric variables, biserial correlations between numeric and binary variables, and polychoric correlations between ordinal variables. Insignificant correlations are blank.

4.2 Modeling

4.2.1 Wald statistics

The Wald statistics of three separate logistic regression models fitted to the co-morbidity labels of diabetes, hypertension and dyslipidemia are shown in figure 4.5. The diabetes logistic regression model has a χ^2 likelihood ratio (LR) of 2135, for hypertension the LR χ^2 is 1450 and for dyslipidemia LR χ^2 is 1049. All three models have 41 d.f. The estimated AUC is 0.94, 0.81 and 0.82 for respectively diabetes, hypertension and dyslipidemia.

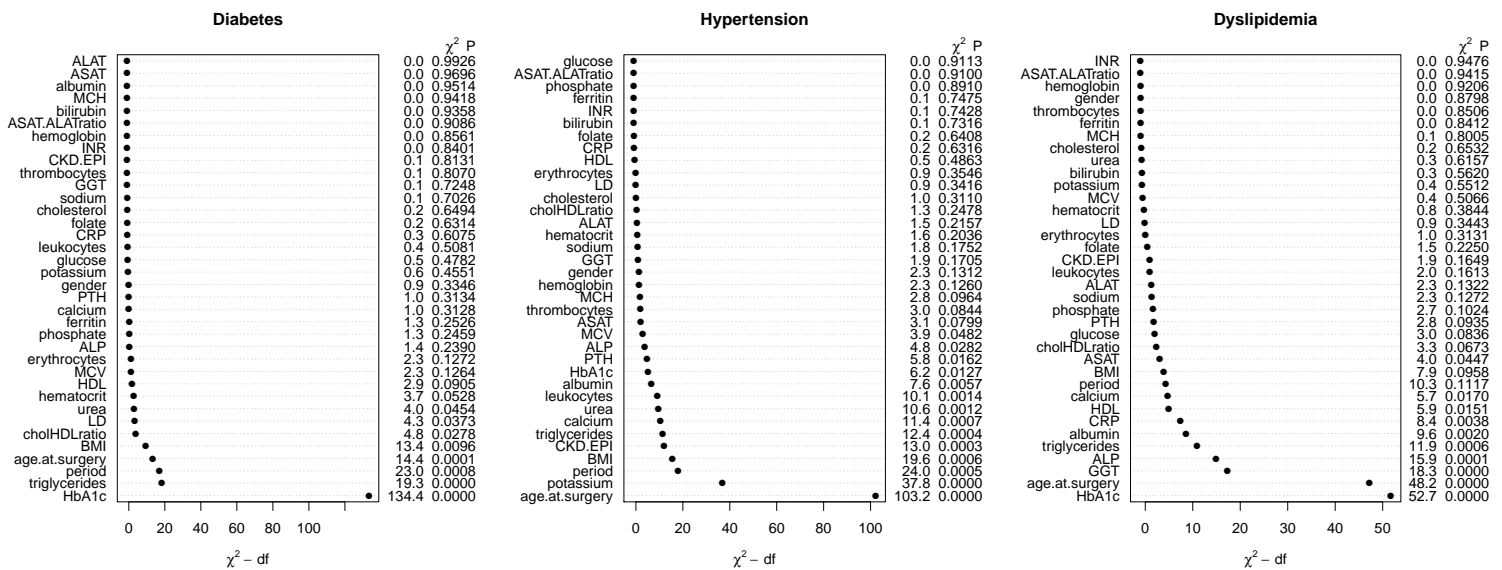


Figure 4.5: Significance of predictors derived from Wald statistics in three separate binary logistic regression models for diabetes, hypertension and dyslipidemia. The significance is judged by the partial Wald χ^2 minus the predictor d.f. The Wald χ^2 values for each line include contributions from all higher-order effects, interaction effects have been removed from the plot

The Wald statistics of the PO and penalized extended CR model fitted to all the patients records are shown in figure 4.6. The full PO model has a χ^2 likelihood ratio of 2386 with 41 d.f. The full penalized extended CR model has a χ^2 likelihood ratio of 2544 with effective d.f. 72.35. Although the penalized extended CR model has a greater likelihood ratio, it also uses more d.f. to allow for unequal slopes between the categories. A likelihood ratio test confirms that the penalized extended CR model fits the data significantly better ($P < 0.001$). HbA1c and age at surgery have the highest Wald χ^2 by a large margin in both models. These are followed by a set containing: BMI, period (before/after surgery), triglycerides and potassium. Note that there is an interaction effect period \times BMI.

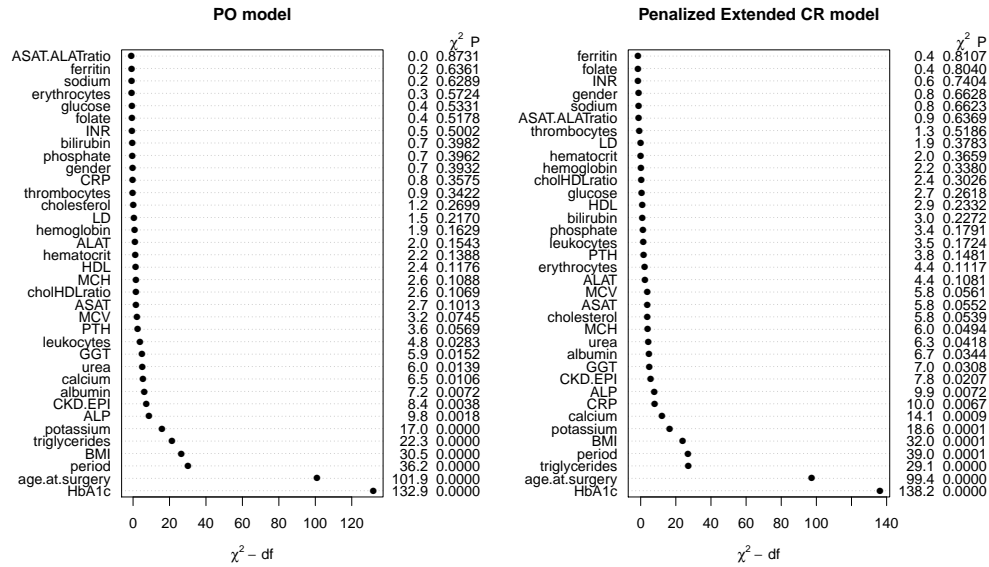


Figure 4.6: Significance of predictors in full PO model and penalized extended CR model. The significance is judged by the partial Wald χ^2 minus the predictor d.f. The Wald χ^2 values for each line include contributions from all higher-order effects, interaction effects have been removed from the plot.

4.2.2 Cross validation results

While the Wald and LR statistics show a significant association between the predictors and the outcome this does not imply that the model has good discriminative ability or is well calibrated. To objectively evaluate and compare the performance of both models the results below are based on a 10-fold cross validation. In each fold the penalty λ for the penalized extended CR ratio is determined with a grid search. On average the penalty for main effect terms was 0.93 and for interaction terms 4.9. These penalties are relatively small, indicating that not much penalization was applied.

Discriminative ability

The AUC for various binary responses and the linear predictor of both models is shown in table 4.2. Because HbA1c turned out to be the predictor that was most significantly associated with the outcome, HbA1c is also included to demonstrate the gain in AUC when using the developed co-morbidity score. While HbA1c has excellent discriminative ability in diabetes, both the ordinal models perform significantly better in discriminating patients without co-morbidities from patients with one or multiple co-morbidities. There are no significant differences in terms of discriminative ability between both ordinal models. The ordinal models are best at discriminating between patients with multiple co-morbidities and patients with one or no co-morbidities. The

ordinal models show better discriminative ability in patients with diabetes versus patients with hypertension or dyslipidemia. In a sub analysis not shown here, the AUC for the PO model for patients with *only* hypertension (and no other co-morbidities) was 0.63 (0.617 - 0.644), for patients with *only* diabetes this AUC was 0.71 (0.764 - 0.820) and for patients with *only* dyslipidemia 0.56 (0.493 - 0.625).

Co-mordidity(ies)	Proportional Odds AUC (95 % CI)	Penalized extended Continuation Ratio AUC (95 % CI)	HbA1c only AUC (95 % CI)
One or multiple	0.815 (0.798 - 0.832)	0.817 (0.801 - 0.833)	0.740 (0.724 - 0.755)
One	0.652 (0.637 - 0.666)	0.658 (0.645 - 0.672)	0.580 (0.563 - 0.596)
Multiple	0.880 (0.861 - 0.900)	0.876 (0.854 - 0.897)	0.836 (0.815 - 0.857)
One or multiple pre-op	0.826 (0.809 - 0.843)	0.828 (0.812 - 0.845)	0.753 (0.741 - 0.766)
One or multiple at 6-month FUP	0.786 (0.759 - 0.812)	0.790 (0.763 - 0.816)	0.698 (0.672 - 0.723)
One or multiple at 12-month FUP	0.778 (0.745 - 0.812)	0.779 (0.744 - 0.815)	0.678 (0.624 - 0.732)
One or multiple at 24-month FUP	0.782 (0.709 - 0.855)	0.789 (0.712 - 0.865)	0.705 (0.616 - 0.794)
Diabetes	0.904 (0.887 - 0.921)	0.902 (0.882 - 0.921)	0.920 (0.900 - 0.939)
Hypertension	0.786 (0.769 - 0.803)	0.788 (0.773 - 0.803)	0.697 (0.68 - 0.713)
Dyslipidemia	0.804 (0.781 - 0.828)	0.802 (0.776 - 0.829)	0.746 (0.719 - 0.773)

Table 4.2: The AUCs are calculated by taking the mean over the 10-folds. The confidence interval (CI) by mean $- 2 \times \text{SE}$ and $2 \times \text{SE} + \text{mean}$.

Calibration

The calibration of both models is compared in table 4.3. Again there are no significant differences between the PO and penalized extended CR model. While the slope may be a good measure of calibration, this does not convey the nature of the mis-calibration, i.e. are low probabilities predicted too high or high probabilities too low? The calibration curves are plotted in figure 4.7. Visual inspection of the curves is in concordance with the slope and intercept values in table 4.3, the slope is near 1 meaning the predicted probabilities are close to the actual probabilities.

	Proportional Odds	Penalized extended Continuation Ratio
Somers' D_{xy}	0.630 (0.596 - 0.664)	0.634 (0.601 - 0.667)
AUC	0.815 (0.798 - 0.832)	0.817 (0.801 - 0.833)
R^2	0.37 (0.33 - 0.40)	0.37 (0.34 - 0.41)
Brier	0.159 (0.154 - 0.164)	0.159 (0.154 - 0.163)
Slope	0.97 (0.87 - 1.07)	0.99 (0.90 - 1.08)
Intercept	-0.032 (-0.21 - 0.15)	-0.010 (-0.20 - 0.18)

Table 4.3: Discrimination, overall and calibration performance measures.

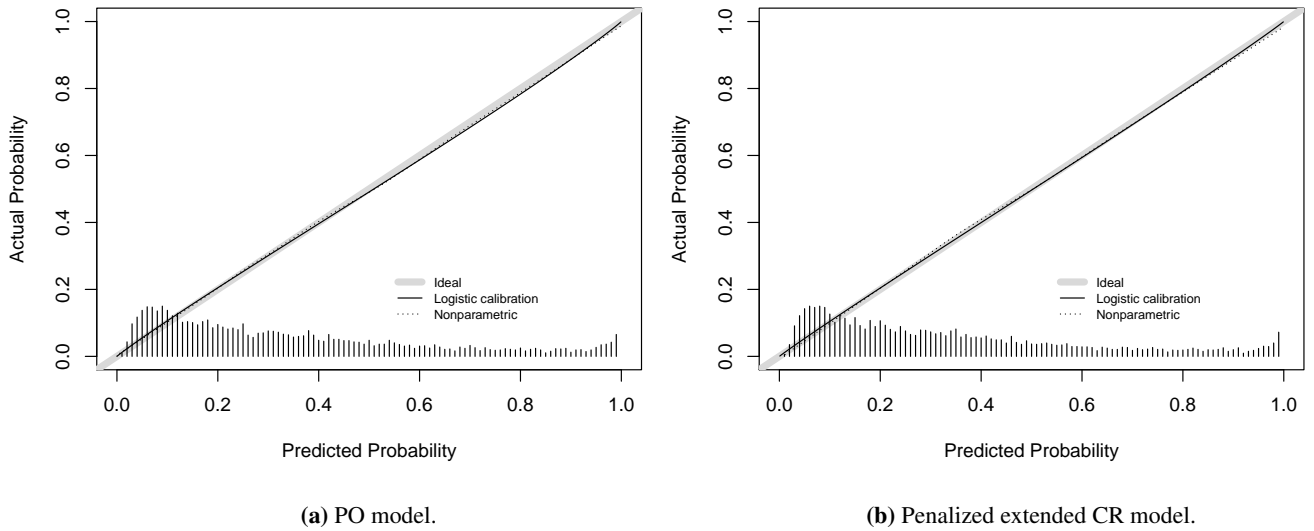


Figure 4.7: Calibration plots of both ordinal models obtained from 10-fold cross validation. The rug plot shows the distribution of predicted probabilities of the entire dataset.

Feature selection

A form of feature selection is applied with Harrell's model approximation procedure, the final models contain a subset of the original 38 predictors. Since folds are randomly sampled from the entire dataset, feature selection also has a variability in terms of which predictors it selects, depending which patients are included in each fold. In figure 4.8 is shown how many times each predictor is included in the final model for all of the 10 folds (i.e. which features are most often selected). The PO model contains

a median of 8 predictors while the CR model contains a median of 6. The feature selection in the penalized extended CR model seems more stable with 5 predictors that are always in the final model: triglycerides, potassium, HbA1c and CKD-EPI. For the PO model the feature selection seems more unstable. Because both models perform similarly in terms of discrimination and calibration we choose the most parsimonious model as our model to be used in the clinic: the penalized extended CR model.

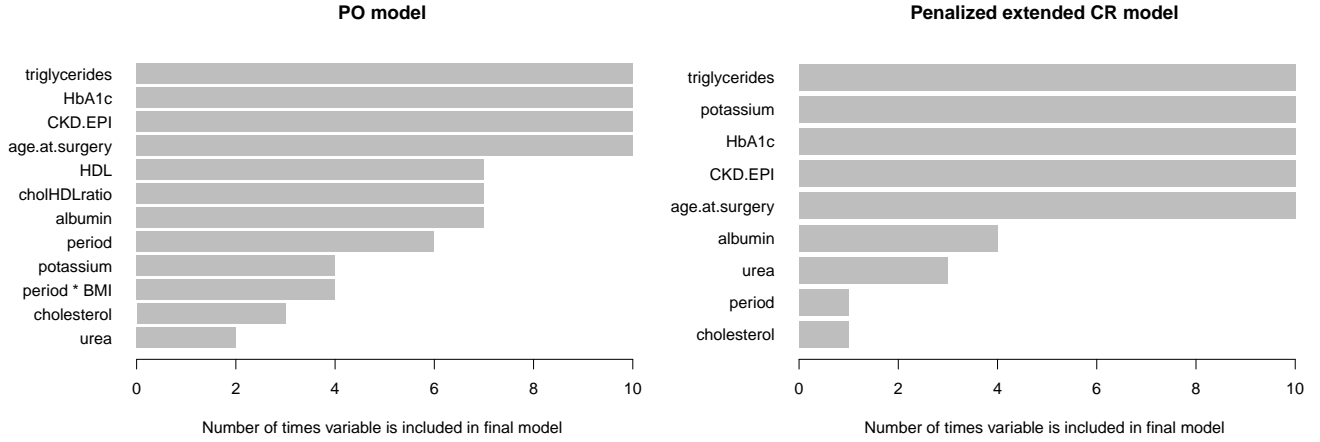


Figure 4.8: Number of times a predictor is included in the final model totaled over the 10 folds.

4.3 Final model visualization

A PO and penalized extended CR model performed similarly in terms of discrimination and calibration. The penalized extended CR model was chosen as a final model and the co-morbidity severity score in this model is calculated from four predictors: HbA1c, age at surgery, triglycerides, CKD-EPI and potassium. The final penalized extended CR model fitted on all the patient records is given by the equation:

$$\text{Prob}\{y = 1\} = \frac{1}{1 + \exp(-X\hat{\beta})}, \text{ where}$$

$$X\hat{\beta} = 2.894 + 0.104 \text{ HbA1c} + 0.0554 \text{ age.at.surgery} + 0.328 \text{ triglycerides} - 0.0230 \text{ CKD.EPI} - 0.541 \text{ potassium}$$

I.e. to calculate the probability that a patients has one or more co-morbidities, $\text{Prob}\{y = 1\}$, the measured HbA1c in mmol/mol, age at surgery in years, triglycerides in mmol/L, CKD-EPI in mL/min/1.73 m² and potassium in mmol/L are entered in the

equation above. This equation can be visualized in a nomogram for easier interpretation, see figure 4.9. While the model returns a probability of having one or multiple

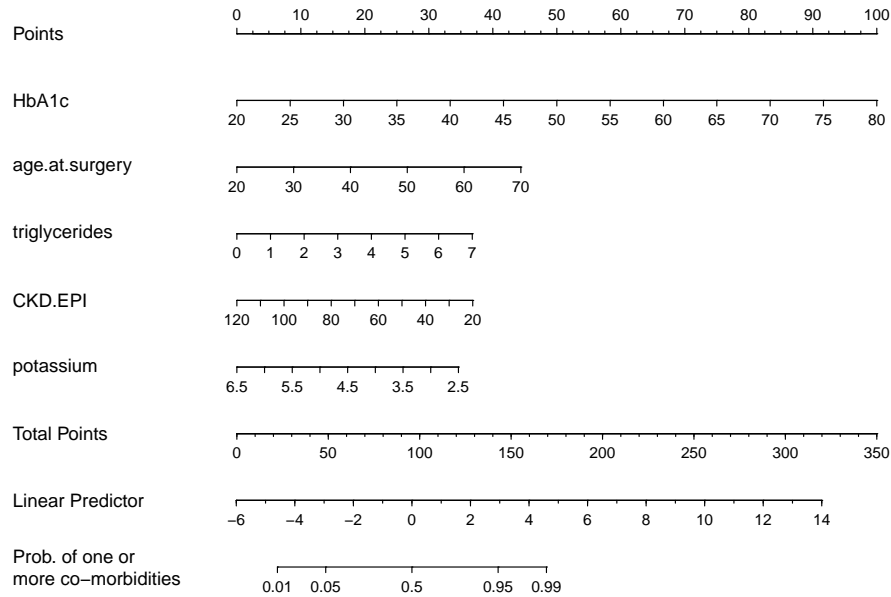


Figure 4.9: Nomogram for predicting the probability that a bariatric patient has one or more co-morbidities. For each marker the number of points can be determined by drawing a line to the top (e.g. a HbA1c of 40 mmol/mol corresponds to about 33 points, an age of 30 to 9 points, etc.), these points are then added to obtain the total points. The total points correspond to a linear predictor value (e.g. 50 points to a linear predictor of -3) and from the linear predictor the probability of one or multiple co-morbidities follows.

co-morbidities, the goal is to obtain a co-morbidity severity *score* and no *probability*. The linear predictor is therefore used directly as the score. To guide the interpretation of the score (i.e. what value can be considered a high score?), it is visualized during screening in figure 4.10 and grouped by the ordinal outcome. Note that by definition a score of 0 corresponds to a probability of 0.5 of having one or multiple co-morbidities, therefore a score < 0 implies less severity of co-morbidities whereas a score > 0 implies higher severity of co-morbidities. E.g. during screening if the score is > 2 there are 15 patients without co-morbidities and 212 patient with one or multiple co-morbidities. If the score is < -2 there are 335 patients without co-morbidities and 33 patients with one or multiple co-morbidities. To show the effect of surgery on the score, patients are grouped by the number of co-morbidities that are present before surgery and the mean score is then plotted over time, see figure 4.11.

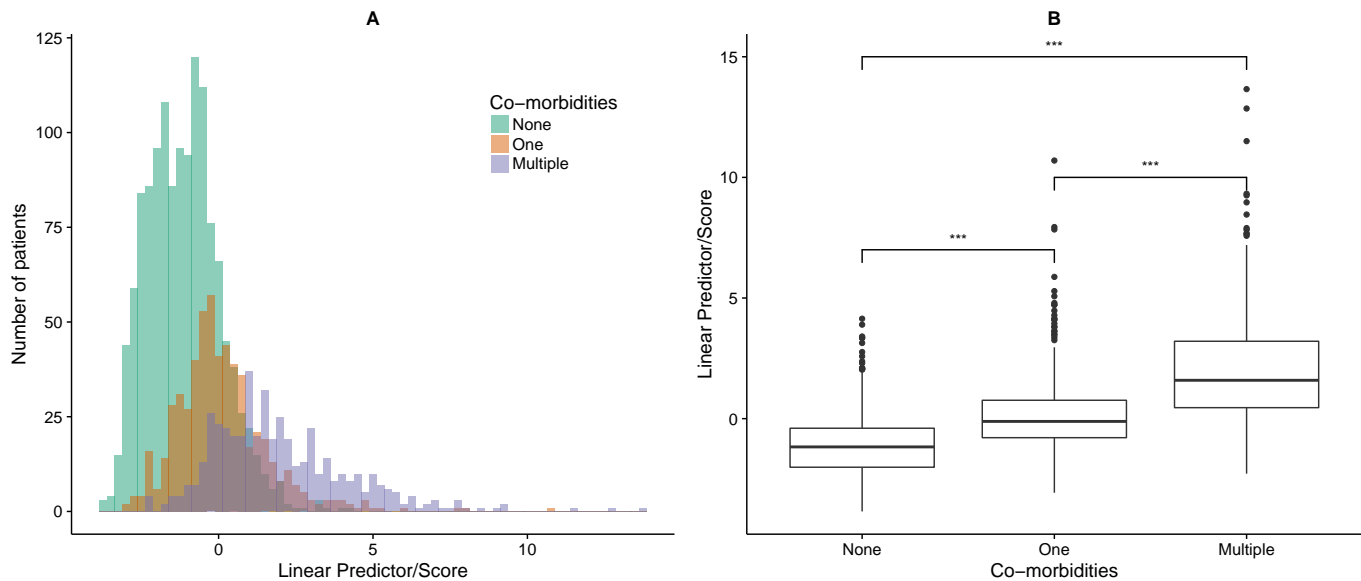


Figure 4.10: (A) Shows the distribution of the score (or linear predictor) during screening. (B) The population means of the score are significantly different (t-test P-value < 0.001) across different ordinal outcomes.

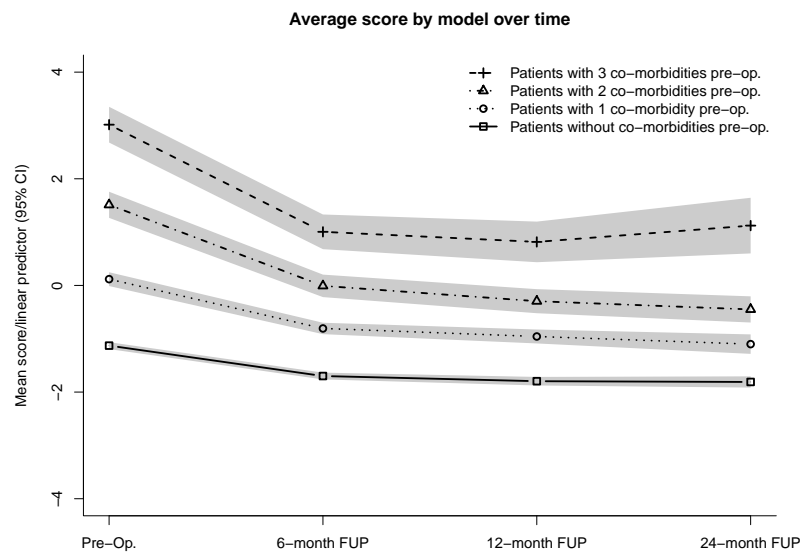


Figure 4.11: Mean score of patients grouped by number of co-morbidities and plotted over time.

5 | Discussion

The goal of this research was to develop a score that could objectively quantify the severity of co-morbidities in bariatric patients based on clinical markers, both before and after surgery. After retrospectively analyzing 3 clinical and 38 blood markers from 2367 patients that underwent bariatric surgery a novel co-morbidity score was developed that quantifies co-morbidity severity based on an ordinal regression model. From age at surgery, HbA1c, triglycerides, CKD-EPI and potassium a co-morbidity score is calculated that showed a strong positive correlation with the number of co-morbidities present in patients (Somers' $D_{xy} = 0.634$). The score also showed good discriminative ability between patients with and without co-morbidities (AUC = 0.817). The discriminative ability is highest before surgery and only slightly lower (AUC = 0.779) after surgery. This score is usable to track improvement in co-morbidities of bariatric patients for at least 2 years after surgery. In addition, this research has shown that HbA1c has potential as a single marker for co-morbidity severity in bariatric patients. Although HbA1c is used to diagnose diabetes, it is shown that co-morbidities in bariatric patients often co-occur. Therefore HbA1c also has predictive value for hypertension and dyslipidemia. While metabolic syndrome scores [13] [14] [15] [16] provide a continuous score of metabolic syndrome severity, these require a fasting glucose sample, measurement of the blood pressure and waist circumference by a clinician. The co-morbidity severity score developed here only requires a (non-fasting) blood sample and the patient's age at surgery.

Since the longitudinal data that was used for analysis did not come from a cohort study but was extracted using data mining techniques, higher drop-out rates than normally found in cohort studies were observed (see section 3.1.4 and in particular figure 3.5). Going from one time period to the next (i.e. from screening to 6-month FUP, 6- to 12-month FUP, etc.) the patient population is reduced by roughly one-third. This introduces bias in two ways. First we expect there to be some form of attrition bias, this is bias that is introduced when those who drop out of the study are systematically different from those who remain in the study [38]. Clinicians indicated that patients tend to drop-out more frequently if they regain weight after surgery, therefore the data is biased towards the non-regainers. Secondly the model is fitted on data available of all patients, of this data 43% is screening data, 29% is 6-month FUP data, 18% is 12-month FUP data and only 10% is 24-month FUP data. Therefore the model will be biased towards screening data. This could also be an explanation for the fact that the predictive ability is higher before than after surgery.

The remission rates for co-morbidities diabetes, hypertension and dyslipidemia are lower than those reported in the systematic review of Buchwald et. al [30] and Yip et. al [24]. While Buchwald et. al only contains RYGB surgeries, Yip et. al reported T2DM remission rates of 76% and 68% at 12 months for RYGB and SG respectively, here we found 55.8% remission of T2DM after 12 months. However, as Brethauer et. al noted, it is difficult to compare remission rates across studies because definitions of co-morbidity improvement or remission are inconsistent [11]. Therefore it is possible that the definition of resolution of co-morbidities is more stringent in DATO than in the studies included by Buchwald and Yip. This could also be the case for %EWL, the %EWL found here is higher than reported in the systematic reviews by Buchwald and O'Brien [30] [31]. O'Brien stated 67% EWL for years 1 and 2 after RYGB, here we found 80% and 77% EWL after years 1 and 2 for RYGB and SG combined. Since the definitions of ideal weight and initial weight are not uniform, the %EWL can vary between studies [11]. Moreover, the attrition bias can play a part. When regainers drop-out, the %EWL will be overestimated.

All correlations that were observed between clinical markers are explainable from knowledge from clinical chemistry. The strong correlations between hemoglobin and erythrocytes can be explained by the fact that hemoglobin is a component of erythrocytes, therefore an increase in erythrocytes leads to an increase of hemoglobin. Hematocrit is the volume percentage of erythrocytes so if the volume remains constant, hematocrit increases when erythrocytes increase. Similarly, hematocrit increases when hemoglobin increases and since MCH and MCV are calculated from respectively hemoglobin and hematocrit, an increase in MCH leads to an increase in MCV. The correlation between hemoglobin and gender is known from the reference values, women have lower reference limits for hemoglobin concentration than men. The correlation between ASAT and ALAT is present because in health the ASAT/ALAT-ratio is equal to 1 (equilibrium). The correlation between glucose and HbA1c is trivial, since HbA1c reflects the average blood glucose levels over the past three months.

While this research emphasizes predictive modeling and not interpretation of individual effects of markers, these are not mutually exclusive. Therefore a (possible) explanation from domain knowledge and the binary logistic regressions of diabetes, hypertension and dyslipidemia is given for each of the markers that are included in the final model. From the binary logistic regression of diabetes; HbA1c, triglycerides and age at surgery are significantly associated with the outcome and included in the final (ordinal) model. This finding is expected since the diagnosis for diabetes is, by and large, based on the measured HbA1c levels (see section 3.1.1). From the logistic regression results one could conclude that glucose is not significantly associated with diabetes (Wald test p-value = 0.48). However, we know from figure 4.4 that glucose is highly correlated with HbA1c and should therefore also be associated with diabetes. This can be explained by the phenomenon of multicollinearity¹ where variables

¹In multivariate logistic regression when predictor variables are highly correlated, the regression coefficients may change erratically in response to small changes in the model or data [52]. When removing HbA1c from the model, the standard error of glucose decreases and the associated P-value from the Wald test becomes significant. Note that while multicollinearity affects standard errors and interpretation of individual regression coefficients, it does not affect the predictions made by the model on new data that has the same degree of multicollinearity as the original data [52].

that are significant in univariate analysis become insignificant in multivariate analysis. Triglycerides are expected to be higher in patients that also suffer from dyslipidemia in combination with diabetes. Age is also an expected factor (T2DM was formerly called adult-onset diabetes), prevalence of diabetes increases with age [56]. From the binary logistic regression of hypertension; age at surgery, CKD-EPI and potassium are significantly associated with the outcome and included in the final model. As blood vessels become less elastic with age, blood pressure also increases with age. Since the level for diagnosing high blood pressure does not change with age, the probability of hypertension increases. Hypertension can be a leading cause of chronic kidney disease (CKD) or a complication of CKD. Chronic hypertension can cause *hypertensive kidney disease* in which the chronic high blood pressure causes damage to kidney tissue, which results in a reduced GFR. High blood pressure can also be a complication of *renovascular hypertension*. Due to reduced blood flow to the kidneys (e.g. caused by stenosis) the kidneys mistakingly detect a low blood pressure and release hormones that promote sodium and water retention, which in turn causes blood pressure to rise. Unlike sodium, potassium has a vasodilatory effect. While excessive sodium intake increases blood pressure by increasing the intra-vascular fluid volume and in turn cardiac output, potassium restores this effect. However, this could also be an effect of diuretics that are used to treat hypertensive patients. Diuretics reduce hypertension because they lower blood pressure by eliminating sodium and water through diuresis (diuresis is the increased production of urine). Some diuretics also cause an elimination of potassium through urine, this could explain lower potassium in hypertensive patients. When potassium was tested in a univariate model to predict hypertension, potassium was not significantly associated with the outcome (data not shown). However, when CKD-EPI was added with potassium in a multivariate model to predict hypertension, potassium was significantly associated with the outcome. This is the opposite case of HbA1c and glucose in diabetes: where glucose is only significant in univariate analysis but not in multivariate analysis, potassium is not significant in univariate analysis but is significant in multivariate analysis. From the binary logistic regression of dyslipidemia; HbA1c, age at surgery and triglycerides are in the final model. These markers are all expected since dyslipidemia is by definition an elevation of triglycerides, HbA1c is used to diagnose diabetes which can also be present in patients with dyslipidemia and dyslipidemia prevalence increases with age [57].

Medication use was not taken into account since the DATO registration of medication use was incomplete for follow-up and not very detailed. Note that the majority of patients suffering from co-morbidities take medication which in turn have an effect on blood marker results. However, while one would expect that patients on medication to treat diabetes would have normal HbA1c levels, this was not the case. Even though the majority of patients suffering from diabetes take medication which would bring down HbA1c levels, mean HbA1c levels were still elevated (data not shown). This has implications for the predictions made by the model. If a patient is severely diabetic, hypertensive and dyslipidemic but does adhere to the prescribed medication use and thus has no abnormal blood marker results, the patient gets a low co-morbidity severity score while in fact the patient has multiple co-morbidities. The model is based on population-averages and if a patient deviates from the average bariatric patient (i.e. in terms of adhering to prescribed medication) the model is not valid.

Another important aspect of building a model on averages are influential observations. While there is no single definition of an influential observation, several clinical markers have extreme values as explained in the results. An extreme value is not by definition an influential observation, if a patient that is labeled with diabetes has an extreme HbA1c level of 120 mmol/mol this observation is not an influential observation in the context of the regression model. However, if a patient that is not labeled with diabetes has an HbA1c of 120 mmol/mol this is an influential observation that affects model performance. In the latter case the patient is most likely mislabeled, as explained in section 3.1.5 DATO labels are not 100% accurate and are most likely the biggest source of error. The influential observations that affect the model are shown in appendix I. We chose not to remove these observations as this would provide an overly optimistic view of model performance. Moreover, not all influential observations arise from mislabelings, some are due to other co-morbidities or diseases.

Logistic regression assumes a linear relationship between the log odds of the dependent variable and the independent variables. To assess whether any gain would be made in terms of predictive ability, this assumption was relaxed by using non-linear restricted cubic splines as described by Harrell [52]. However, no significant improvements were seen in predictive ability. Logistic regression also performed similar to another popular machine learning technique called Random Forests [58]. This indicates that no significant concessions were done in terms of predictive ability by choosing a white-box logistic regression model.

To internally validate the model on unseen data, all performance statistics are calculated from 10-fold cross validation. There are two ways to calculate the AUC from 10-fold cross validation:

Merged AUC merge the folds into one dataset and calculate one overall AUC.

Averaged AUC calculate the AUC for each fold and take the average of these AUCs.

Here we used the averaged AUC from the 10 folds as Forman and Scholz recommend after having performed several simulations [59]. In machine learning the use of repeated cross validation is also advocated. In repeated cross validation the random splitting of the data into N-folds is done repeatedly (e.g. 50 or 100 times) and for each of these splits the N-fold cross validation is performed. So instead of obtaining 10 AUCs from 10-fold cross validation, one obtains with 50 repeats of 10-fold cross validation: $50 \times 10 = 500$ AUCs. However, while this reduces the variance (because there are more samples to average over) this increases the bias as shown by Vanwinckelen and Blockeel [60]. Therefore we did not perform repeated cross validation.

6 | Conclusions and recommendations

The co-morbidity severity score that is developed here allows for an objective assessment of a bariatric patient's metabolic health state with respect to diabetes, hypertension and dyslipidemia. This provides not only the ability for a comparison between patients but also monitoring individual patient improvement over time. While the score is internally validated, the actual value to patients and clinicians can only be determined after it is implemented in the clinic. In a broader perspective this study shows that by using data mining techniques to retrospectively analyze data that is seen as a by-product in health-care, knowledge can be gained. This knowledge can be presented in any desired form (descriptive statistics, figures, prediction models, etc.). Important to consider is that the quality of the knowledge that is extracted, is a direct result of the quality of the data itself but also the pre-processing ("garbage in, garbage out"). Missingness in clinical data (or any data) should always be investigated since systematic missingness can lead to bias. In a supervised learning context, the quality of the final model is also determined to a great extent by the correctness of the outcome labels that the model is trained on. Therefore it is stressed that objective criteria have to be formulated for the outcome of interest and these criteria should be strictly adhered to, to obtain a useful clinical prediction model. Finally, consider the impact of multicollinearity. While it does not affect the predictive ability of the model and is therefore of little interest in a machine learning context. It is of great importance if conclusions are to be drawn on the basis of individual regression coefficients and P-values.

Recommendations for further research would be to implement the model in the clinic and to assess the usability of the score by clinicians and its impact on patients. Should this prove successful, an external validation of the model that was obtained here can provide evidence for use in other clinics as well. To achieve this, age and the four biomarkers that are included in the model would have to be measured in patients that undergo bariatric surgery in a different hospital. The model must first be calibrated to account for differences in measurement results and after calibration be validated on new data.

References

1. Murdoch, T. B. & Detsky, A. S. The inevitable application of big data to health care. *Jama* **309**, 1351–1352 (2013).
2. Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. From data mining to knowledge discovery in databases. *AI magazine* **17**, 37 (1996).
3. Piatetsky-Shapiro, G. Data mining and knowledge discovery 1996 to 2005: overcoming the hype and moving from “university” to “business” and “analytics”. *Data Mining and Knowledge Discovery* **15**, 99–105 (2007).
4. Bellazzi, R. & Zupan, B. Predictive data mining in clinical medicine: current issues and guidelines. *International journal of medical informatics* **77**, 81–97 (2008).
5. Rubino, F. *et al.* Bariatric, metabolic, and diabetes surgery: what’s in a name? *Annals of surgery* **259**, 117–122 (2014).
6. Buchwald, H. *et al.* Weight and type 2 diabetes after bariatric surgery: systematic review and meta-analysis. *The American journal of medicine* **122**, 248–256 (2009).
7. Madsbad, S., Dirksen, C. & Holst, J. J. Mechanisms of changes in glucose metabolism and bodyweight after bariatric surgery. *The lancet Diabetes & endocrinology* **2**, 152–164 (2014).
8. Schauer, P. R. *et al.* Bariatric surgery versus intensive medical therapy for diabetes—3-year outcomes. *New England Journal of Medicine* **370**, 2002–2013 (2014).
9. Frühbeck, G. Bariatric and metabolic surgery: a shift in eligibility and success criteria. *Nature Reviews Endocrinology* **11**, 465–477 (2015).
10. Franco, J. V. A., Ruiz, P. A., Palermo, M. & Gagner, M. A review of studies comparing three laparoscopic procedures in bariatric surgery: sleeve gastrectomy, Roux-en-Y gastric bypass and adjustable gastric banding. *Obesity surgery* **21**, 1458–1468 (2011).
11. Brethauer, S. A. *et al.* Standardized outcomes reporting in metabolic and bariatric surgery. *Obesity surgery* **25**, 587 (2015).
12. Van de Laar, A. W., de Brauw, M., Bruin, S. C. & Acherman, Y. I. Weight-Independent Percentile Chart of 2880 Gastric Bypass Patients: a New Look at Bariatric Weight Loss Results. *Obesity surgery* **26**, 2891–2898 (2016).

13. Wiley, J. F. & Carrington, M. J. A metabolic syndrome severity score: A tool to quantify cardio-metabolic risk factors. *Preventive medicine* **88**, 189–195 (2016).
14. Soldatovic, I., Vukovic, R., Culafic, D., Gajic, M. & Dimitrijevic-Sreckovic, V. siMS Score: Simple Method for Quantifying Metabolic Syndrome. *PloS one* **11**, e0146143 (2016).
15. Lee, A. M., Gurka, M. J. & DeBoer, M. D. A MetS severity score to estimate risk in adolescents and adults: current evidence and future potential. *Expert review of cardiovascular therapy* **14**, 411 (2016).
16. Wijndaele, K. *et al.* A continuous metabolic syndrome risk score. *Diabetes Care* **29**, 2329–2329 (2006).
17. Still, C. D. *et al.* A probability score for preoperative prediction of type 2 diabetes remission following RYGB surgery. *The lancet. Diabetes & endocrinology* **2**, 38 (2014).
18. Gloy, V. L. *et al.* Bariatric surgery versus non-surgical treatment for obesity: a systematic review and meta-analysis of randomised controlled trials. *Bmj* **347**, f5934 (2013).
19. Angrisani, L., Formisano, G., Santonicola, A., Hasani, A. & Vitiello, A. in *Bariatric and Metabolic Surgery* 19–24 (Springer, 2017).
20. Karamanakos, S. N., Vagenas, K., Kalfarentzos, F. & Alexandrides, T. K. Weight loss, appetite suppression, and changes in fasting and postprandial ghrelin and peptide-YY levels after Roux-en-Y gastric bypass and sleeve gastrectomy: a prospective, double blind study. *Annals of surgery* **247**, 401–407 (2008).
21. Gagner, M., Deitel, M., Erickson, A. L. & Crosby, R. D. Survey on laparoscopic sleeve gastrectomy (LSG) at the Fourth International Consensus Summit on Sleeve Gastrectomy. *Obesity surgery* **23** (2013).
22. Casella, G. *et al.* Long-term results after laparoscopic sleeve gastrectomy in a large monocentric series. *Surgery for Obesity and Related Diseases* **12**, 757–762 (2016).
23. Vidal, J. *et al.* Type 2 diabetes mellitus and the metabolic syndrome following sleeve gastrectomy in severely obese subjects. *Obesity surgery* **18**, 1077 (2008).
24. Yip, S., Plank, L. D. & Murphy, R. Gastric bypass and sleeve gastrectomy for type 2 diabetes: a systematic review and meta-analysis of outcomes. *Obesity surgery* **23**, 1994–2003 (2013).
25. Jiménez, A. *et al.* Long-term effects of sleeve gastrectomy and Roux-en-Y gastric bypass surgery on type 2 diabetes mellitus in morbidly obese subjects. *Annals of surgery* **256**, 1023–1029 (2012).
26. Abbatini, F. *et al.* Long-term effects of laparoscopic sleeve gastrectomy, gastric bypass, and adjustable gastric banding on type 2 diabetes. *Surgical endoscopy* **24**, 1005–1010 (2010).
27. Basso, N. *et al.* First-phase insulin secretion, insulin sensitivity, ghrelin, GLP-1, and PYY changes 72 h after sleeve gastrectomy in obese diabetic patients: the gastric hypothesis. *Surgical endoscopy* **25**, 3540 (2011).

28. Sarkhosh, K., Birch, D. W., Shi, X., Gill, R. S. & Karmali, S. The impact of sleeve gastrectomy on hypertension: a systematic review. *Obesity surgery* **22**, 832–837 (2012).
29. Perathoner, A. *et al.* Significant weight loss and rapid resolution of diabetes and dyslipidemia during short-term follow-up after laparoscopic sleeve gastrectomy. *Obesity surgery* **23**, 1966–1972 (2013).
30. Buchwald, H. *et al.* Bariatric surgery: a systematic review and meta-analysis. *Jama* **292**, 1724–1737 (2004).
31. O’Brien, P. E., McPhail, T., Chaston, T. B. & Dixon, J. B. Systematic review of medium-term weight loss after bariatric operations. *Obesity surgery* **16**, 1032–1040 (2006).
32. Cummings, D. E. *et al.* Plasma ghrelin levels after diet-induced weight loss or gastric bypass surgery. *New England Journal of Medicine* **346**, 1623–1630 (2002).
33. Bloomberg, R. D., Fleishman, A., Nalle, J. E., Herron, D. M. & Kini, S. Nutritional deficiencies following bariatric surgery: what have we learned? *Obesity surgery* **15**, 145–154 (2005).
34. R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2016). <https://www.R-project.org/>.
35. Friedewald, W. T., Levy, R. I. & Fredrickson, D. S. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clinical chemistry* **18**, 499–502 (1972).
36. Warnick, G. R. & Wood, P. D. National Cholesterol Education Program recommendations for measurement of high-density lipoprotein cholesterol: executive summary. The National Cholesterol Education Program Working Group on Lipoprotein Measurement. *Clinical chemistry* **41**, 1427–1433 (1995).
37. Morabia, A. *A history of epidemiologic methods and concepts* (Birkhäuser, 2013).
38. Salkind, N. J. Encyclopedia of Measurement and Statistics. **1**, 57–60 (2006).
39. Levey, A. S. *et al.* A new equation to estimate glomerular filtration rate. *Annals of internal medicine* **150**, 604–612 (2009).
40. Michels, W. M. *et al.* Performance of the Cockcroft-Gault, MDRD, and new CKD-EPI formulas in relation to GFR, age, and body size. *Clinical Journal of the American Society of Nephrology* **5**, 1003–1009 (2010).
41. De Ritis, F., Coltorti, M. & Giusti, G. An enzymic test for the diagnosis of viral hepatitis: the transaminase serum activities. *Clinica Chimica Acta* **2**, 70–74 (1957).
42. Payne, R., Carver, M. E. & Morgan, D. Interpretation of serum total calcium: effects of adjustment for albumin concentration on frequency of abnormal values and on detection of change in the individual. *Journal of clinical pathology* **32**, 56–60 (1979).

43. Larsson, J. *eulerr: Area-Proportional Euler Diagrams* R package version 1.0.0 (2016). <https://cran.r-project.org/package=eulerr>.
44. Harrell Jr, F. E., with contributions from Charles Dupont & many others. *Hmisc: Harrell Miscellaneous* R package version 4.0-2 (2016). <https://CRAN.R-project.org/package=Hmisc>.
45. Tsoumakas, G., Katakis, I. & Vlahavas, I. in *Data mining and knowledge discovery handbook* 667–685 (Springer, 2009).
46. Agresti, A. *Analysis of ordinal categorical data* (John Wiley & Sons, 2010).
47. Steyerberg, E. *Clinical prediction models: a practical approach to development, validation, and updating* (Springer Science & Business Media, 2008).
48. Nelder, J. A. & Wedderburn, R. W. M. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)* **135**, 370–384 (1972).
49. Harrell Jr, F. E. *rms: Regression Modeling Strategies* R package version 5.1-0 (2017). <https://CRAN.R-project.org/package=rms>.
50. Fox, J. *polycor: Polychoric and Polyserial Correlations* R package version 0.7-9 (2016). <https://CRAN.R-project.org/package=polycor>.
51. Wei, T. & Simko, V. *corrplot: Visualization of a Correlation Matrix* R package version 0.77 (2016). <https://CRAN.R-project.org/package=corrplot>.
52. Harrell, F. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis* (Springer, 2015).
53. Guyon, I. & Elisseeff, A. An introduction to variable and feature selection. *Journal of machine learning research* **3**, 1157–1182 (2003).
54. Flom, P. L. & Cassell, D. L. Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use. *NorthEast SAS Users Group (NESUG): Statistics and Data Analysis* (2007).
55. Kuhn, M. *caret: Classification and Regression Training* R package version 6.0-73 (2016). <https://CRAN.R-project.org/package=caret>.
56. Volksgezondheidszorg.info. Prevalentie diabetes naar leeftijd en geslacht <https://www.volksgezondheidszorg.info/onderwerp/diabetes-mellitus/cijfers-context/huidige-situatie>.
57. Volksgezondheidszorg.info. Ongunstig cholesterolgehalte (hypercholesterolemie) <https://www.volksgezondheidszorg.info/onderwerp/cholesterol/cijfers-context/huidige-situatie>.
58. Breiman, L. Random forests. *Machine learning* **45**, 5–32 (2001).
59. Forman, G. & Scholz, M. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *ACM SIGKDD Explorations Newsletter* **12**, 49–57 (2010).
60. Vanwinckelen, G. & Blockeel, H. On estimating model accuracy with repeated cross-validation in *BeneLearn 2012: Proceedings of the 21st Belgian-Dutch Conference on Machine Learning* (2012), 39–44.

Appendices

A | Summary of Literature review

To assess whether any of the clinical markers are known from previous research to be associated with diabetes, hypertension or dyslipidemia a search was performed on Google Scholar. Each combination of marker and co-morbidity were entered as keywords and results were analyzed to see if an association between the respective marker and co-morbidity was found. This could be either a positive or negative association. Different synonyms for clinical markers were also tried (e.g. parathormone and parathyroid hormone, white blood cells and leukocytes). Note that only the markers contained in the follow-up panel are researched, also vitamins were not taken into consideration since they are supplemented. See table [A.1](#).

	Diabetes	Hypertension	Dyslipidemia
hemoglobin	[1] [2] [3]	[4]	
hematocrit	[3] [2]	[5] [4]	
erythrocytes	[2]	[4]	
MCH		[4]	
MCV		[4]	
thrombocytes			
leukocytes	[2] [6] [7] [8]	[9] [10] [4]	[11]
bilirubin	[12] [13] [14]	[15]	[16]
ASAT	[17]	[18]	[18]
ALAT	[19] [20] [21] [17]	[18]	[18]
LD	[22]		
alkaline phosphatase	[17] [23]		
gamma GT	[21] [17] [24] [25]	[26] [25]	[25]
urea	[27]	[28] [29] [4]	
creatinine		[30] [31]	[32]
potassium		[33] [34]	
sodium		[33]	
calcium	[35] [36]	[36]	[36]
phosphate			
albumin			
CRP	[37] [38] [39]	[38] [40] [41] [4]	[38]
cholesterol		[42] [4]	§
HDL-cholesterol		[42] [4]	§
chol/HDL ratio			§
LDL-cholesterol		[42] [4]	§
triglycerides		[42] [4]	§
prothrombin time			
hemoglobin A1c (IFCC)	†		
glucose	†	[43] [4]	
parathormone	[44]	[45] [46]	
iron			
ferritin	[47] [48] [49]	[50]	[49]

Table A.1: References found for associations between co-morbidities and markers.

§ dyslipidemia is diagnosed from these markers.

† diabetes is diagnosed from these markers.

B | Concise Statistical Description of Dataset after Pre-processing

df.all.correct

60 Variables5565 Observations

patient

n	missing	distinct
5565	0	2367

lowest : 2473337880 7070366830 7722366810 7874337800 8326936780
highest: 993871469810 996174337880 996570336880 998272337880 998475431890

surgery.date

n	missing	distinct
5565	0	586

lowest : 2012-01-09 2012-01-10 2012-01-16 2012-01-18 2012-01-23
highest: 2016-02-15 2016-02-17 2016-02-18 2016-02-23 2016-02-24

period

n	missing	distinct
5565	0	4

Value Pre 6 M 12 M 24 M
Frequency 2367 1596 1019 583
Proportion 0.425 0.287 0.183 0.105

time

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
5565	0	30	0.991	4.878	9.235	-6	-4	-2	5	11	22	23

lowest : -11 -10 -9 -8 -7, highest: 22 23 24 25 26

gender

n	missing	distinct
5565	0	2

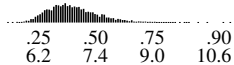
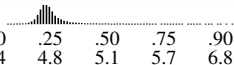
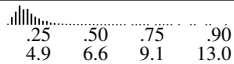
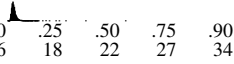
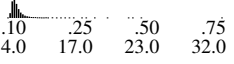
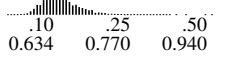

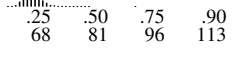
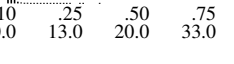
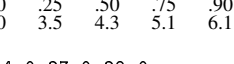
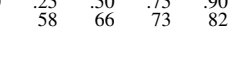
Value M F
Frequency 1107 4458
Proportion 0.199 0.801

age.at.surgery

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
5565	0	465	1	42.12	12.69	23.4	26.3	33.6	43.0	50.2	56.5	59.8

lowest : 18.1 18.2 18.4 18.5 18.6, highest: 67.6 67.7 68.0 68.5 68.9

height													
	n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
	5565	0	55	0.999	1.69	0.09776	1.56	1.59	1.63	1.68	1.74	1.81	1.84
lowest : 1.37 1.39 1.45 1.46 1.47, highest: 1.93 1.94 1.95 1.96 2.00													
weight													
	n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
	5565	0	919	1	103.7	29.05	68.0	73.0	83.4	100.4	120.8	137.7	149.0
lowest : 48.0 48.3 51.0 51.7 52.0, highest: 215.0 217.0 222.0 227.1 230.0													
BMI													
	n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
	5565	0	419	1	36.21	9.3	24.9	26.3	29.4	35.6	41.9	46.6	50.4
lowest : 15.7 18.5 19.6 20.2 20.3, highest: 66.9 67.7 69.0 69.3 72.3													
surgery													
	n	missing	distinct										
	5565	0	2										
Value	sleeve bypass												
Frequency	2990 2575												
Proportion	0.537 0.463												
hemoglobin													
	n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
	5565	0	63	0.998	8.576	0.8204	7.5	7.7	8.1	8.5	9.0	9.5	9.8
lowest : 4.0 5.0 5.2 5.3 5.5, highest: 11.1 11.3 11.4 11.8 12.2													
hematocrit													
	n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
	5565	0	31	0.989	0.415	0.0345	0.37	0.38	0.40	0.41	0.43	0.46	0.47
lowest : 0.20 0.28 0.29 0.30 0.31, highest: 0.53 0.54 0.56 0.58 0.59													
erythrocytes													
	n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
	5565	0	36	0.993	4.685	0.4369	4.1	4.2	4.4	4.7	4.9	5.2	5.4
lowest : 2.2 3.1 3.3 3.4 3.5, highest: 6.2 6.3 6.6 6.7 7.0													
MCH													
	n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
	5565	0	95	0.999	1.834	0.123	1.64	1.70	1.77	1.84	1.91	1.96	2.00
lowest : 1.10 1.22 1.24 1.25 1.26, highest: 2.17 2.18 2.19 2.20 2.25													
MCV													
	n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
	5565	0	45	0.995	88.75	5.029	81	83	86	89	92	94	96
lowest : 56 61 63 64 66, highest: 103 104 105 106 109													
thrombocytes													
	n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
	5565	0	58	0.997	267.6	70.03	180	190	220	260	300	350	380
lowest : 33 53 58 64 79, highest: 550 560 620 650 690													

leukocytes													
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95	
5565	0	147	1	7.756	2.458	4.7	5.2	6.2	7.4	9.0	10.6	11.8	
lowest : 2.0 2.4 2.8 2.9 3.0, highest: 18.6 19.5 20.9 21.5 27.1													
glucose													
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95	
5565	0	141	0.998	5.528	1.347	4.2	4.4	4.8	5.1	5.7	6.8	8.2	
lowest : 2.4 2.6 2.7 2.8 3.0, highest: 19.1 19.4 19.6 19.9 21.8													
bilirubin													
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95	
5565	0	123	1	7.768	4.406	3.3	3.8	4.9	6.6	9.1	13.0	16.0	
lowest : 1.7 1.8 1.9 2.0 2.1, highest: 50.0 53.0 54.0 58.0 75.0													
ASAT													
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95	
5565	0	97	0.997	24.57	10.27	14	16	18	22	27	34	42	
lowest : 5.0 6.4 7.3 8.3 8.4, highest: 160.0 180.0 240.0 360.0 870.0													
ALAT													
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95	
5565	0	140	0.999	28.31	17.36	12.0	14.0	17.0	23.0	32.0	47.0	62.8	
lowest : 5.0 5.6 5.7 6.0 6.2, highest: 290.0 300.0 320.0 510.0 670.0													
ASAT.ALATratio													
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95	
5565	0	202	1	0.9824	0.3351	0.570	0.634	0.770	0.940	1.140	1.370	1.530	
lowest : 0.14 0.16 0.19 0.20 0.23, highest: 2.89 3.19 3.20 3.29 4.56													
LD													
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95	
5565	0	227	1	180.7	36.16	134	143	158	177	199	222	238	
lowest : 69 83 89 91 92, highest: 380 394 439 443 578													
ALP													
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95	
5565	0	165	1	84.07	25.78	51	57	68	81	96	113	126	
lowest : 24 25 26 27 28, highest: 209 220 232 350 744													
GGT													
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95	
5565	0	199	0.999	30.2	26.31	8.0	10.0	13.0	20.0	33.0	57.6	82.0	
lowest : 4 5 6 7 8, highest: 389 445 453 553 1700													
urea													
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95	
5565	0	93	0.999	4.472	1.468	2.7	3.0	3.5	4.3	5.1	6.1	6.9	
lowest : 1.4 1.5 1.6 1.7 1.8, highest: 20.0 22.0 24.0 27.0 28.0													
creatinine													
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95	
5565	0	124	0.999	67.9	15.2	50	53	58	66	73	82	90	

potassium												
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
5565	0	27	0.987	4.006	0.3197	3.6	3.7	3.8	4.0	4.2	4.4	4.5

sodium	n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
	5565	0	19	0.978	140.8	2.366	.137	.138	.139	.141	.142	.144	.144

calcium												
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
5565	0	96	0.998	2.279	0.09108	2.15	2.18	2.22	2.28	2.33	2.38	2.42

phosphate												
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
5565	0	119	1	1.074	0.1897	0.79	0.86	0.96	1.07	1.19	1.28	1.35

albumin														
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95		
5565	0	24	0.985	44.19	2.899	.40	.41	.42	.44	.46	.47	.48		

CRP	n	missing	distinct
	5565	0	2

cholesterol												
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
5565	0	525	1	4.745	1.111	3.26	3.52	4.04	4.67	5.38	6.02	6.48

n		missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
5565		0	137	0.998	1.542	0.9283	0.64	0.73	0.93	1.30	1.80	2.60	3.20

HDL												
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
5565	0	223	1	1.297	0.3886	0.81	0.89	1.04	1.25	1.50	1.76	1.93

65

cholHDLratio												
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
5565	0	88	0.999	3.918	1.449	2.2	2.5	3.0	3.7	4.6	5.6	6.4
lowest : 1.4 1.5 1.6 1.7 1.8, highest: 11.0 12.0 13.0 15.0 19.0												
INR												
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
5565	0	132	0.996	1.022	0.126	0.91	0.92	0.95	0.99	1.03	1.07	1.11
lowest : 0.81 0.82 0.83 0.84 0.85, highest: 4.30 4.56 4.79 4.82 5.19												
HbA1c												
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
5565	0	88	0.996	38.87	8.463	31	32	34	37	40	48	58
lowest : 11 22 24 25 26, highest: 112 115 120 121 124												
PTH												
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
5565	0	106	1	6.367	2.899	3.1	3.6	4.4	5.8	7.6	9.7	11.0
lowest : 0.15 0.83 0.87 1.10 1.60, highest: 29.00 41.00 54.00 78.00 120.00												
iron												
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
5565	0	111	0.997	14.61	6.224	6.8	8.0	10.0	14.0	18.0	22.0	25.0
lowest : 2.3 2.7 2.8 2.9 3.0, highest: 44.0 46.0 49.0 51.0 54.0												
ferritin												
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
5565	0	522	1	122.1	111.9	15	23	45	88	156	253	336
lowest : 3 4 5 6 7, highest: 1361 1474 1491 1499 1580												
folate												
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
5565	0	84	0.999	21.18	11.24	8.9	10.0	13.0	19.0	27.0	37.0	44.0
lowest : 4.6 4.7 5.1 5.4 5.5, highest: 41.0 42.0 43.0 44.0 45.0												
vitB1												
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
5565	0	385	1	135.8	35.44	89.0	98.6	114.0	133.0	154.0	174.0	187.0
lowest : 42.0 44.0 47.7 48.0 48.9, highest: 304.0 334.0 354.0 753.0 1000.0												
vitB6												
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
5565	0	135	0.999	102.6	60.51	46	52	64	84	110	160	200
lowest : 11 20 22 28 29, highest: 820 870 890 980 1000												
vitD25OH												
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
5565	0	101	1	57.48	29.72	19	24	37	56	75	92	100
lowest : 10 11 12 13 14, highest: 160 170 180 190 200												
vitB12												
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
5565	0	109	0.999	313	161.5	150	170	210	270	360	470	580

lowest : 74 76 78 79 80, highest: 1100 1200 1300 1400 1500

CKD.EPI	n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
	5565	0	121	1	112.5	16.95	86	96	104	113	123	131	136

lowest : 5 6 7 13 15, highest: 151 153 155 156 158

dia.full	n	missing	distinct										
	5565	0	7										
Value			No	Yes	Cured	Improved	Same	Worse	Not present				
Frequency			1956	411	223	217	54	7	2697				
Proportion			0.351	0.074	0.040	0.039	0.010	0.001	0.485				

hyp.full	n	missing	distinct										
	5565	0	8										
Value			No	Yes	Cured	Improved	Same	Worse	Denovo				
Frequency			1541	826	295	385	308	8	1				
Proportion			0.277	0.148	0.053	0.069	0.055	0.001	0.000				
Value			Not present										
Frequency			2201										
Proportion			0.396										

dys.full	n	missing	distinct										
	5565	0	7										
Value			No	Yes	Cured	Improved	Same	Worse	Not present				
Frequency			1935	432	123	132	240	11	2692				
Proportion			0.348	0.078	0.022	0.024	0.043	0.002	0.484				

dia.bin	n	missing	distinct										
	5565	0	2										
Value			FALSE	TRUE									
Frequency			4880	685									
Proportion			0.877	0.123									

hyp.bin	n	missing	distinct										
	5565	0	2										
Value			FALSE	TRUE									
Frequency			4065	1500									
Proportion			0.73	0.27									

dys.bin	n	missing	distinct										
	5565	0	2										
Value			FALSE	TRUE									
Frequency			4773	792									
Proportion			0.858	0.142									

no.of.comorbs	n	missing	distinct	Info	Mean	Gmd							
	5565	0	4	0.71	0.535	0.7867							

Value	0	1	2	3
Frequency	3645	1124	535	261
Proportion	0.655	0.202	0.096	0.047

noneVSrest

n	missing	distinct
5565	0	2

Value	FALSE	TRUE
Frequency	1920	3645
Proportion	0.345	0.655

multipleVSrest

n	missing	distinct
5565	0	2

Value	FALSE	TRUE
Frequency	4769	796
Proportion	0.857	0.143

oneVSrest

n	missing	distinct
5565	0	2

Value	FALSE	TRUE
Frequency	4441	1124
Proportion	0.798	0.202

comorb

n	missing	distinct
5565	0	3

Value	None	One	Multiple
Frequency	3645	1124	796
Proportion	0.655	0.202	0.143

classlabel

n	missing	distinct
5565	0	8

Value	Dia	Dia+Dys	Dia+Hyp	Dia+Hyp+Dys	Dys	Hyp	Hyp+Dys
Frequency	151	109	164	261	160	813	262
Proportion	0.027	0.020	0.029	0.047	0.029	0.146	0.047

Value	None
Frequency	3645
Proportion	0.655

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 μ mol/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 μ mol/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 μ mol/L	continuous	No
ferritin	Ferritin in μ g/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

Table B.1: Description of variables in dataset after pre-processing.

C | Logistic Regression

This explanation is based on Harrell and Steyerberg, for more details see the references [51] and [52]. Data of bariatric patients at screening is used to illustrate the concepts explained below.

C.1 Generalized Linear Models

The generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allows regression on different types of outcome data such as binary data, count data, probability data, etc. The general linear regression model (GLM) as formulated by John Nelder and Robert Wedderburn [53] is given by:

$$C(Y|X) = g(X\beta)$$

Where Y denotes the outcome, $X = X_1, X_2, \dots, X_p$ the predictor variables and $\beta = \beta_1, \beta_2, \dots, \beta_p$ the list of regression components that have to be determined. $C(Y|X)$ denotes a property of the distribution of Y given X . For example:

$$C(Y|X) = E(Y|X) = X\beta$$

which is the ordinary linear regression. $C(Y|X)$ can also be the probability that $Y = 1$ given X . In this case the binary logistic regression model applies (explained further below):

$$C(Y|X) = P\{Y = 1|X\} = (1 + \exp(-X\beta))^{-1}$$

This model can be made linear in the regression coefficients β by a transformation in the property $C(Y|X)$:

$$h(C(Y|X)) = X\beta$$

The expression $X\beta$ is also referred to as the *linear predictor* and $h(\dots)$ is called the *link function*. To summarize, a GLM consists of three components:

- A probability distribution from the exponential family. The exponential family contains a large number of probability distributions such as the normal, binomial, Poisson and gamma distributions.
- A linear predictor $X\beta$
- A link function h such that $C(Y|X) = h^{-1}(X\beta)$

C.2 Logistic regression

As stated above, logistic regression models are a specific type of GLM which applies when the outcome variable is binomially distributed. If we were to fit an ordinary linear regression model:

$$E(Y|X) = X\beta$$

to a binary outcome we would also obtain values for $E(Y|X)$ above 1 and below 0 (see figure C.1).

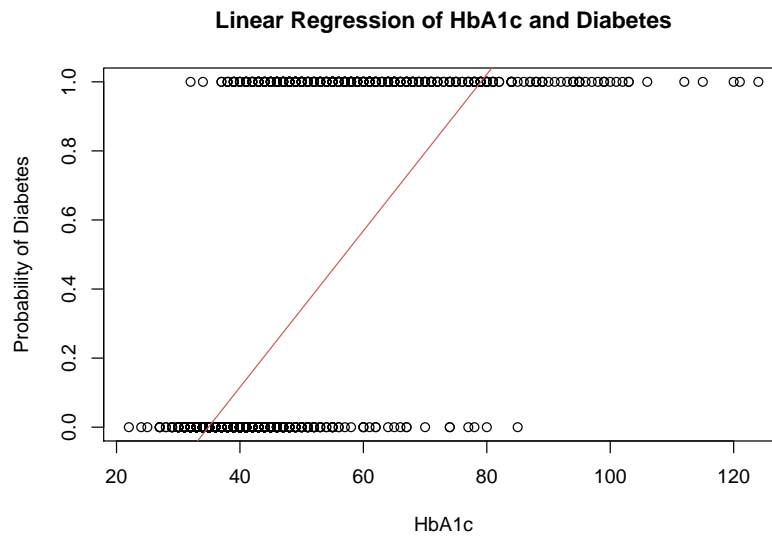


Figure C.1: Ordinary least squares applied to a binomial outcome of diabetes yes (= 1) and no (= 0). Note that with linear regression the probability can exceed 1 or fall below 0 while in the logistic regression it always falls between 0 and 1.

Therefore the preferred model for the analysis of binary responses is the logistic regression model:

$$P\{Y = 1|X\} = (1 + \exp(-X\beta))^{-1}$$

The function:

$$P = (1 + \exp(-x))^{-1}$$

is called the logistic function and transforms any value x between 0 and 1, note that when $x = 0$, $P = 0.5$. To express x in terms of P we use the *inverse* logistic function, called the logit:

$$x = \log[P/(1 - P)]$$

$$x = \text{logit}\{Y = 1\}$$

The logit serves as a link function between the linear predictor and the binary outcome Y . Therefore the logistic regression model can be written in terms of the logit to make

it linear in $X\beta$:

$$\underbrace{\text{logit}}_{\text{link fun.}} \{Y = 1|X\} = \underbrace{\beta_0 + \beta_1 X_1}_{\text{linear predictor}}$$

Where β_0 is also called the intercept term and is equal to the log odds of $Y = 1$ when $X = 0$ and β_1 the log odds of the $Y = 1$ after a one unit increase in X_1 .

C.3 Example

Instead of fitting a ordinary linear regression model to HbA1c and the binary outcome of diabetes (as in figure C.1) we will fit a logistic regression model. The result can be seen in figure C.2. To calculate the probability we take the inverse logit (or logistic

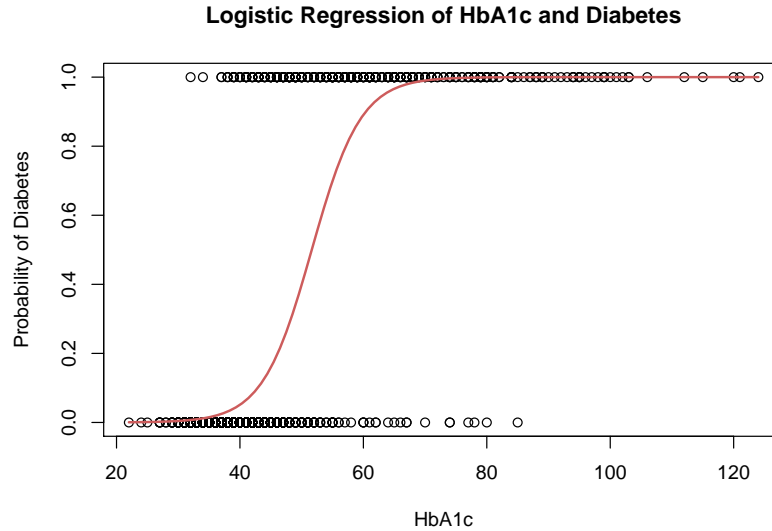


Figure C.2: Logistic regression equation given by: $\text{logit}\{\text{Diabetes} = 1|X\} = -12.961 + 0.251 \times \text{HbA1c}$

function) of the linear predictor:

$$\begin{aligned} \text{logit}\{\text{Diabetes} = 1|\text{HbA1c}\} &= -12.961 + 0.251 \times \text{HbA1c} \\ P\{Y = \text{Diabetes}|\text{HbA1c}\} &= \text{logit}^{-1}(-12.961 + 0.251 \times \text{HbA1c}) \end{aligned}$$

The linear predictor (LP $= -12.961 + 0.251 \times \text{HbA1c}$) is plotted against the probability in figure C.3. A LP value of 0 always corresponds to a probability of 50% and LP values of -4 correspond to low probabilities $< 2\%$ and LP values of 4 correspond to high probabilities $> 98\%$.

From this model we can conclude that $\beta_1 = 0.251$ (β_0 is the intercept), so an increase of 1 mmol/mol means a increase of 0.251 in the log odds of having diabetes. By taking the exponential of this number we obtain the odds ratio: $\exp(0.251) = 1.29$.

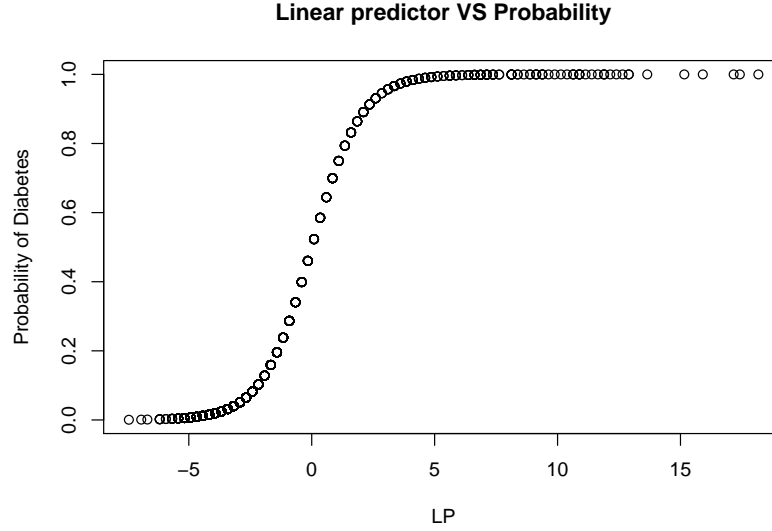


Figure C.3: The $LP = -12.961 + 0.251 \times \text{HbA1c}$ and the probability is calculated from the inverse logit of the LP.

An odds-ratio > 1 means the predictor is associated with a higher odds of the outcome. Note that the presence of a positive odds-ratio for an outcome does not necessarily indicate that this association is statistically significant. Confidence intervals and p-values must be given. When we do a multivariate logistic regression the coefficient β_i is an "adjusted" effect, it is conditional on the values of other predictors in the model. The effect of a one unit increase in x_i is associated with the log-odds of β_i under the condition that all the other predictors are kept constant. If we fit a multivariate logistic regression of HbA1c and age we obtain:

$$\text{logit}\{\text{Diabetes} = 1 | \text{HbA1c}, \text{age}\} = -14.4751 + 0.2364 \times \text{HbA1c} + 0.0465 \times \text{age}$$

If we would increase the age one year (*and* keep HbA1c constant) we increase the log odds of having diabetes by 0.00465.

C.4 Interaction terms

Both models that we considered above are so called *main effects* models, that is, no interaction terms are included:

$$\begin{aligned} \text{logit}(\text{Diabetes} = 1 | X) &= \beta_0 + \beta_1 \text{HbA1c} \\ \text{logit}(\text{Diabetes} = 1 | X) &= \beta_0 + \beta_1 \text{HbA1c} + \beta_2 \text{age} \end{aligned}$$

Interaction terms are added to a model when the effect of X_1 on Y depends on the level of X_2 . For example if we expect older males to have a risk of hypertension, we include

an interaction between age and gender:

$$\text{logit}(\text{Hypertension} = 1|X) = \beta_1 \text{age} + \beta_2 \text{gender} + \beta_3 \text{gender*age}$$

Some interaction terms that are commonly seen in clinical settings are:

- Interaction between treatment and severity of disease.
- Interactions involving age and risk factors.
- Interaction between month and markers, since some markers have seasonal effects.
- Etc. see Steyerberg [52]

C.5 Maximum Likelihood Estimation

The logistic regression coefficients β are usually fitted using maximum likelihood estimation [51]. The likelihood refers to the probability of the data given the model. For convenience in numerical estimation the natural logarithm of the likelihood ($\log L$) is often used. The $\log L$ is calculated by summing the distances between the natural log of the predicted probability P of the binary outcome to the actually observed outcome Y over all the all the measurements:

$$\log L = \sum \underbrace{Y}_{\text{positives}} \times \log(P) + \underbrace{(1-Y)}_{\text{negatives}} \times \log(1-P)$$

Y refers to the outcome and P the predicted probability for each measurement. A perfectly fitting model would have a $\log L$ of zero, this means that for each $Y = 1$ the predicted probability $P = 1$ and for each $Y = 0$, $P = 0$. Of course this is rarely the case in a medical setting. The value for P is unknown but we can solve this to obtain the value of P this is *most likely* to have occurred. This value for P is the *maximum likelihood estimate* (MLE) of the population probability and from this value for P the β coefficients follow. Figure C.4 right demonstrates visually what the MLE does. The MLE value of P is that value of P that maximizes the $\log L$, the maximum value can be found at the point at which the slope of $\log L$ is 0. Setting the first derivative of the $\log L$ function to zero and solving this equation with an iterative algorithm such as Newton-Raphson yields the MLE value of P . The negative of the second derivative of the $\log L$ evaluated at the MLE is the observed *Fisher information matrix*, the estimated standard errors of the MLE are the square roots of the diagonal elements of the inverse of the observed Fisher information matrix (also called the variance-covariance matrix). The second derivative indicates the extent to which the log likelihood is peaked rather than flat. This makes the interpretation in terms of standard errors intuitively. The standard errors are important to determine how precisely the model estimates the coefficient's unknown value. For example, dividing the coefficient by its standard error results in a z-score. If the z-score associated with this z-statistic is less than a certain alpha value (usually 0.05) we can conclude that the coefficient is significantly different from zero.

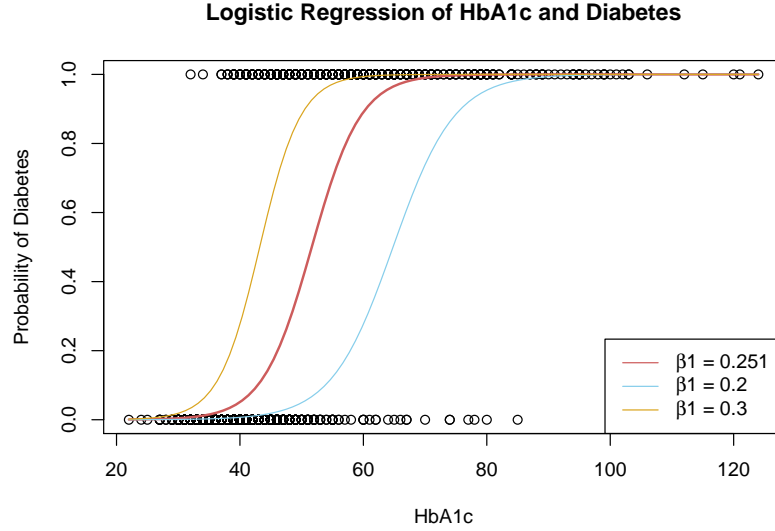


Figure C.4: Effect of changing the β coefficient on the logistic regression curve. $\beta = 0.251$ is the value for β found by the MLE, a lower value for β (0.2) underestimates the probability and a higher value for β (0.3) overestimates the probability of diabetes.

C.6 Penalized Maximum Likelihood Estimation

Using the MLE we obtain the best possible fit to the data at hand, but this can also result in fitting noise in the data. Shrinkage results in regression coefficient estimates that are lower in mean squared error and therefore closer to the true unknown parameter values [54]. There is trade-off however, with reducing the variance the bias of the regression coefficients increases, see figure C.5. Ridge regression is one approach to shrinkage [54] but here we use penalized maximum likelihood estimation [51]. Instead of maximizing the log likelihood a penalty term is added:

$$\text{PML} = \log L - \underbrace{\frac{1}{2} \lambda \sum (s_i \beta_i)^2}_{\text{penalty term}}$$

The PML function is now maximized by MLE, where L is the maximum likelihood of the fitted model, λ a penalty factor, β the estimated regression coefficients for each predictor i in the model and s_i a scaling factor to make $s_i \beta_i$ unit-less. For convenience reasons the scaling factor s_i can be chosen as the standard deviation of each predictor i . The PML can be generalized to:

$$\text{PML} = \log L - \frac{1}{2} \lambda \beta' P \beta$$

Where β is the vector of regression coefficients (β' the transpose) and P a penalty matrix. The default penalty matrix P has the variance of the continuous predictors in

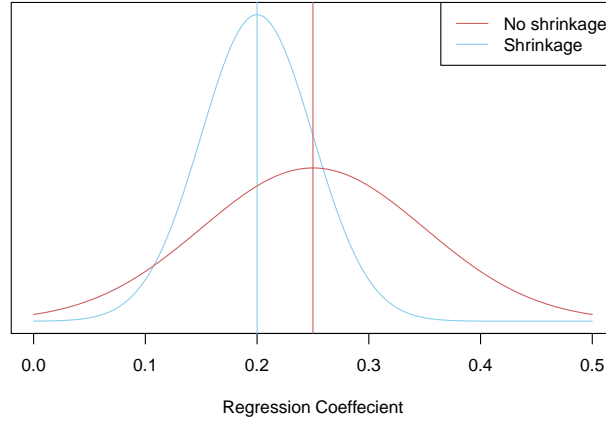


Figure C.5: After shrinkage, regression coefficients have lower variance but higher bias. If we assume that the true value of the regression coefficient equals 0.25, then after shrinkage we obtain a lower (biased) value but with a smaller variance.

its diagonal elements so that the penalty to the log likelihood is unitless. For categorical predictors the penalty is also made independent of the choice of the reference cell. The choice for the penalty factor λ can be made by cross-validation on some statistic such as AUC. An example in figure C.6 using a different penalty function called the LASSO [54] shows the influence of λ on the AUC and number of predictors included in the model. Instead of choosing a cross validation to determine the penalty factor λ , simulation studies have shown that a modified AIC is quicker and performs better at finding a good value of λ . If a variety of λ are tried (one can choose different λ values for main effects and for interactions) the λ that maximizes the AIC is likely to maximize predictive accuracy on a future dataset [51].

C.7 Hypothesis tests

C.7.1 Likelihood Ratio test

The performance of a logistic regression model can be evaluated by comparing the found $\log L$ with a null model $\log L_0$:

$$\log L_0 = \sum Y \times \log(\text{mean}(Y/N)) + (1 - Y) \times \log(1 - \text{mean}(Y/N))$$

Where N is the number of measurements and $\text{mean}(Y/N)$ represents the average probability of the outcome. The performance is given by the Likelihood Ratio statistic (LR) which follows a χ^2 distribution:

$$\text{LR} = -2(\log L_0 - \log L)$$

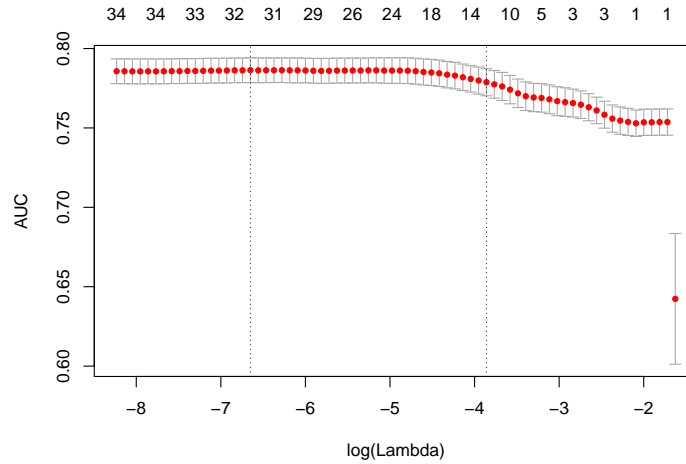


Figure C.6: 10-fold cross validation of LASSO penalty factor λ versus AUC. On the x-axis the penalty $\log(\lambda)$ increases from -8 to -2, we see that an increase in penalty leads to a reduction in AUC. On the top of the graph the number of predictors that are included in the model are shown. So, while the higher penalty leads to a reduction in AUC it also leads to a smaller model containing only relevant predictors. The dotted line between -6 and -7 gives the minimum mean cross-validated error. The other dotted line between -3 and -4 gives the most regularized model such that the error is within one standard error of the minimum. Figure made with `glmnet` package.

The LR can be used for testing the importance of predictors in the model. If we fit our diabetes logistic regression model (see appendix C) with HbA1c as our only predictor we obtain a LR χ^2 of 1215, now we add age to our model and we obtain an LR of LR χ^2 1240. The difference between the LR is $1240 - 1215 = 25$. We can test whether age significantly improves the fit of the model by applying a χ^2 test with 1 degree of freedom, this yields a P-value of <0.001 so we conclude that age significantly improves our model fit. To demonstrate this principle we will fit a logistic regression model of iron to the outcome of diabetes. We do not expect iron levels to be related to the outcome, our model confirms this; the LR drops to 0.47. The associated P values equals 0.50, hence we expect that diabetes is not significantly associated with iron levels.

C.7.2 Wald test

The Wald test is a rough approximation of the LR test. The LR test requires at least two models to test, like demonstrated above. The Wald test can be performed on a single model. The MLE estimates of our regression coefficients $\hat{\beta}$ are tested whether they are significantly different from β_0 , which is usually 0 because we want to test whether a coefficient is significantly different from 0. In our example with HbA1c we want to test whether the odds ratio of having diabetes when increasing HbA1c by one unit is greater than 1. Remember that in logistic regression the coefficients are given in log odds and that an odds ratio of 1 is equal to a regression coefficient

of 0. We know from the theory of MLE, that the difference between $\hat{\beta}$ and β_0 will be approximately normally distributed with mean 0. Dividing a normal distribution with mean 0 and standard deviation σ by its standard deviation will yield the standard normal distribution with mean 0 and standard deviation 1. The Wald statistic for a single parameter is defined as:

$$W = \frac{\hat{\beta} - \beta_0}{\text{se}(\hat{\beta})} \sim \mathcal{N}(0, 1)$$

Where $\text{se}(\hat{\beta})$ is the standard error of the MLE. Because the parameter of interest is usually $\beta = 0$ this simplifies to:

$$W = \frac{\hat{\beta}}{\text{se}(\hat{\beta})} \sim \mathcal{N}(0, 1)$$

Which is simply the estimated regression coefficient divided by its standard error. This results in a z-score and because the Wald statistic is asymptotically distributed as a standard normal distribution, we can use the z-score to calculate the p-value. However, the Wald statistic is usually squared and compared to a χ^2 distribution:

$$W^2 = \frac{\hat{\beta}^2}{\text{var}(\hat{\beta})} \sim \chi^2$$

This allows the Wald test to be performed on multiple parameters and is therefore used by default. The Wald test assumes a large sample size (i.e., 80–100 or more subjects).

C.7.3 Example

In this example we will fit a logistic regression model to the outcome of hypertension in bariatric patients before surgery. First we will consider a model with four variables: age, potassium, calcium and bilirubin.

Logistic Regression Model

```
lrm(formula = hyp.bin ~ age.at.surgery + potassium + calcium +
    bilirubin, data = dat)
```

		Model Likelihood Ratio Test	Discrimination Indexes	Rank Discrim. Indexes
Obs	2367	LR χ^2 466.71	R^2 0.247	C 0.759
FALSE	1541	d.f. 4	g 1.229	D_{xy} 0.518
TRUE	826	Pr(> χ^2) <0.0001	g_r 3.418	γ 0.519
max $ \frac{\partial \log L}{\partial \beta} 1 \times 10^{-11}$			g_p 0.234	τ_a 0.235
			Brier 0.186	

	$\hat{\beta}$	S.E.	Wald Z	Pr(> Z)
Intercept	-8.0402	1.4319	-5.62	<0.0001
age.at.surgery	0.0918	0.0052	17.76	<0.0001
potassium	-0.5577	0.1789	-3.12	0.0018
calcium	2.3829	0.5933	4.02	<0.0001
bilirubin	0.0261	0.0135	1.94	0.0527

In the upper table on the left we see the observations (Obs) that were used to fit the model: 2367 in total of which 1541 did not have hypertension and 826 did. In the Model Likelihood Ratio Test we can see that the full model has a LR χ^2 of 466.71, we have four continuous variables and thus four degrees of freedom, a χ^2 test with four degrees of freedom gives a P-value of <0.0001. So we can expect that these four variables are associated with the outcome of hypertension. The Discrimination and Rank Discrimination indexes we take for granted for now, these apply later when we have to assess predictive performance. In the lower table we see the estimated regression coefficients $\hat{\beta}$, their standard error, the Wald Z-score (which is simply $\hat{\beta}/\text{S.E.}$) and the associated P-value with this Z-score. We can see that age, potassium and calcium all have a P-value < 0.05, only bilirubin is perhaps not significant (P= 0.0527) in predicting hypertension status. We see from the regression coefficients $\hat{\beta}$ that an increase in age and calcium are positively associated with hypertension, and an increase in potassium is negatively associated with hypertension. If we drop bilirubin from our model we obtain a LR χ^2 of 462.99, the difference in LR between the full model and the model without bilirubin is 3.72. Bilirubin takes up one d.f. so a one d.f. χ^2 test results in a P-value of 0.0537 which is very close to the Wald statistic.

D | Ordinal Logistic Regression

Categorical variables can either be nominal or ordinal. Nominal variables are for example gender (male or female) or race (white, black, hispanic, etc.). These variables have no natural ordering. Categorical variables that do have a natural ordering are called ordinal variables. Examples are social class (upper, middle, lower) or patient condition (good, fair, serious, critical). The ordinal variables have different categories but distances between them are unknown. Although a patient's condition can be described as fair, there is no way to quantify *how much* better the condition of a good patient is. From our data we know that bariatric patients that have multiple co-morbidities have a higher measure of co-morbidity than patients with only one or no co-morbidities, but we cannot quantify by *how much*. Therefore we chose an ordinal scale. A basic assumption of ordinal regression models is that the ordinal response Y behaves in an ordinal fashion with respect to each predictor X . A simple way to assess this assumption is to plot, for each predictor X , the mean stratified over the levels of Y [51]. These means should be in consistent order. See figure D.1 for an example.

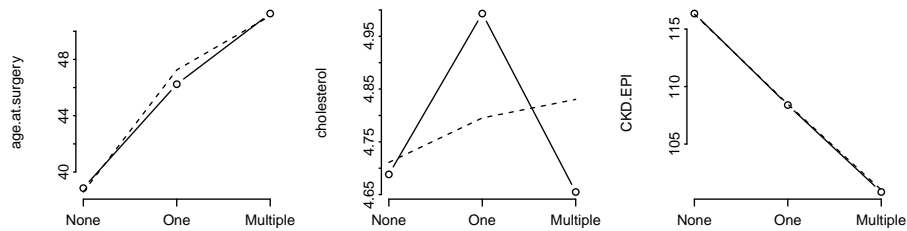


Figure D.1: Means of age at surgery, cholesterol and CKD-EPI stratified over levels of Y . The points connected with solid lines are the observed means, the dotted lines are the lines under the proportional odds assumption. If the proportional odds model fits, these lines overlap. We see that age at surgery and CKD-EPI behave in an ordinal fashion; the average age at surgery increase as the level of Y increases and the average CKD-EPI decreases as the level of Y increases. Cholesterol does not behave in an ordinal fashion; the mean increases when going from the none to one category, but then decreases when going to the multiple category.

D.1 Proportional Odds Model

The most common used ordinal logistic model was described in Walker and Duncan [55] and is called the proportional odds (PO) model by McCullagh [56]. For our outcome Y with levels None, One or Multiple it is stated as follows:

$$\Pr[Y \geq j|X] = \frac{1}{1 + \exp[-(\alpha_j + X\beta)]}$$

Where $j = \text{One or Multiple}$ and $\Pr[Y \geq j|X]$ gives the probability of being *at or above* a certain level versus being below this level. The regression coefficients β are the same for all levels but there are k intercepts (α s). Fitting this model results in a common set of regression coefficients β and two intercept terms: α_{one} and α_{multiple} . E.g. if we wanted to know the probability that a patient has one or more co-morbidities we use the α_{one} intercept to obtain $\Pr[Y \geq \text{One}|X]$. An advantage of the proportional odds model is that it can deal with an ordered outcome but the price we pay is that we assume that the regression coefficients β are independent of the outcome level j . This assumption is called the of proportionality of the odds.

D.2 Penalized Extended Continuation Ratio Model

The PO model is based on *cumulative* probabilities, i.e. the probability of having more than one or multiple co-morbidities, the continuation ratio (CR) model is based on *conditional* probabilities. The forward CR model is stated as follows:

$$\Pr[Y = j|Y \geq j, X] = \frac{1}{1 + \exp[-(\alpha_j + X\beta)]}$$

Where $j = \text{None or One}$. This might seem similar to the PO model, but note the conditional probability: $\Pr[Y = j|Y \geq j, X]$. Here the question is, given the patient has reached level j , what is the chance of going further? At the first category of the CR model we calculate the probability that the patient has no co-morbidities versus one or multiple co-morbidities. At the second category of the CR model we calculate the probability that, *given that the patient has reached this level*, the patient will move to the next level. The patients without co-morbidities are dropped. To illustrate the difference between the PO and CR model we have visualized the data splits in figure D.2.

In its ordinary form the CR model offers no advantage over the PO model. However the CR model can be fitted with a binary logistic regression. This enables the possibility of relaxing the equal slopes assumption without the requirement of specialized software as is the case with the PO model. The β s are allowed change with the Y -cutoff, i.e. we obtain two logistic regression models:

$$\begin{aligned} \Pr[Y = \text{None}|Y \geq \text{None}, X] &= \frac{1}{1 + \exp[-(\alpha + X\beta_{\text{None}})]} \\ \Pr[Y = \text{One}|Y \geq \text{One}, X] &= \frac{1}{1 + \exp[-(\alpha + \theta_{\text{One}} + X\beta_{\text{One}})]} \end{aligned}$$

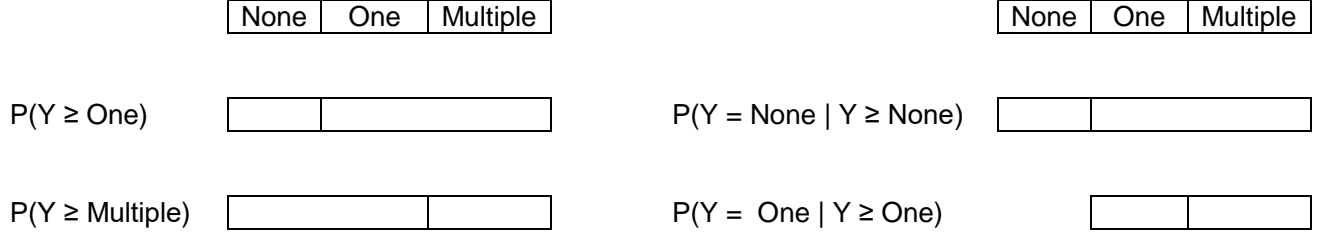


Figure D.2: (Left) The PO model compares the One and Multiple categories to the None category, or the Multiple category to the One and None categories. (Right) The CR model compares the None category to the One and Multiple categories or the One to the Multiple category.

Although these are technically two separate models, a single logistic regression model can be "tricked" in fitting both categories with the technique described by Armstrong and Sloan [57] and Berridge and Whitehead [58] and as implemented in the `rms` package by Harrell. The process does not stop here however. Because we have 38 predictors, who all use 1 degree of freedom, except for "period" which uses an additional 2 degrees of freedom, we have 40 degrees of freedom in the regular PO model¹. Fitting the CR model with a binary logistic trick has 81 degrees of freedom, each β has to be fitted twice and an additional intercept term θ_{One} .

D.3 Overfitting and limits on number of predictors

When a model is fitted that is too complex, that is, it has too many degrees of freedom for the amount of samples available, overfitting can occur. Some findings will arise from fitting noise instead of signal. In Harrell [51] a rule of thumb is that the number of candidate predictors p should be less than $m/10$ or $m/20$. Where m is the "limiting sample size". The limiting sample size of continuous, binary and ordinal response variables is given in table D.1. We have an ordinal response variable with three

Response variable	Limiting sample size m
Continuous	n (total sample size)
Binary	$\min(n_1, n_2)$ (the sample size of the minority class)
Ordinal	$n - \frac{1}{n^2} \sum_{k=1}^{i-1} n_k^3$ (k categories)

Table D.1: Limiting sample sizes for different response variables.

categories and our sample size is equal to 2367. Although we have 5565 records we

¹the number of degrees of freedom is the number of values in the final calculation of a statistic that are free to vary.

only have 2367 *unique* patients. The sample size n_i for each category is given in table 3.3. Calculating m from the table D.1 we obtain $m = 1896$, so we can include between $1896/20 = 95$ and $1986/10 = 190$ predictors in our model. The PO model includes 40 degrees of freedom and the extended CR model 81, so both stay below the $m/20$ -rule of thumb. Therefore we do not expect overfitting to occur. However, as this only a rule of thumb and the minimum sample size needed to just calculate the intercept or residual variance is not taken into account, we will still include penalization in our extended CR model. Moreover we will test model calibration to assess the actual degree of overfitting.

D.4 Penalization of extended CR model

Penalized maximum likelihood is applied to the extended CR model (see appendix C for an explanation) by varying both the penalty for main effects and interaction effects. Selection for the optimal penalty is done based on Hurvich and Tsai's corrected AIC (AIC_c). The set of penalties that combined yield the highest AIC_c is chosen for the final penalized extended CR model. This combination of penalty terms is obtained using the `pentrace` function in the `rms` package. In figure D.3 a diagnostic plot from this function is plotted to illustrate the selection of penalty terms.

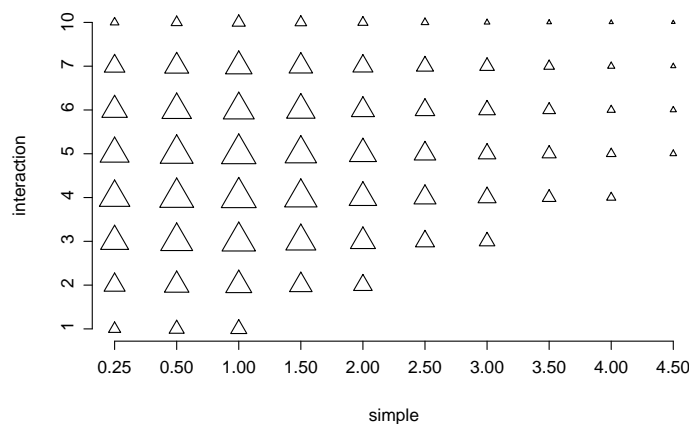


Figure D.3: Diagnostic plot obtained from `pentrace` function. The size of the triangles represents the AIC_c , the x-axis shows the penalty for simple (main effect) terms and the y-axis the penalty for interaction terms. The combination of a penalty 1 for simple terms and a penalty of 4 for interaction terms yields the highest AIC_c .

E | Ordinality Assumption Plots

Examination of the ordinality of the outcome Y for each predictor by assessing how the variations of Y relate to the mean of each predictor X , and whether the trend is monotonic. Solid lines connect the observed means, dotted lines represent the expected means under the PO assumption, and Cs under the CR assumption.

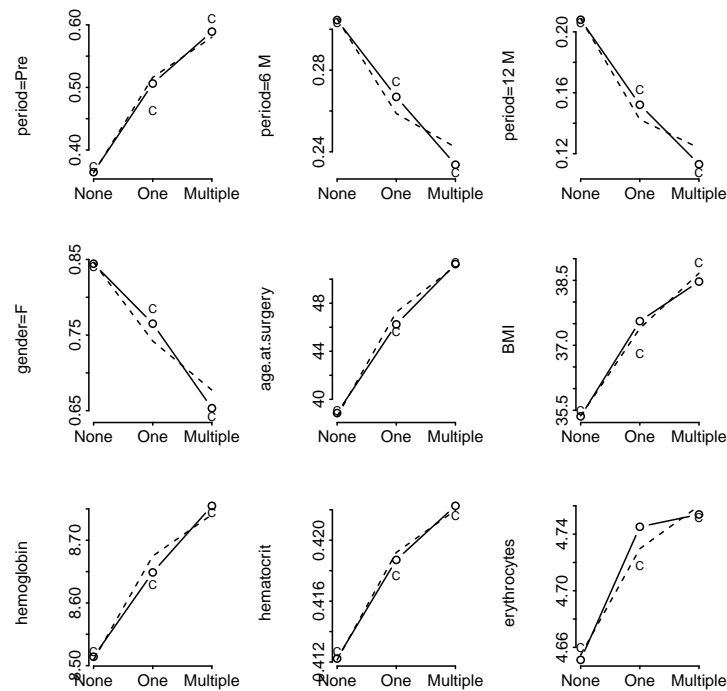


Figure E.1

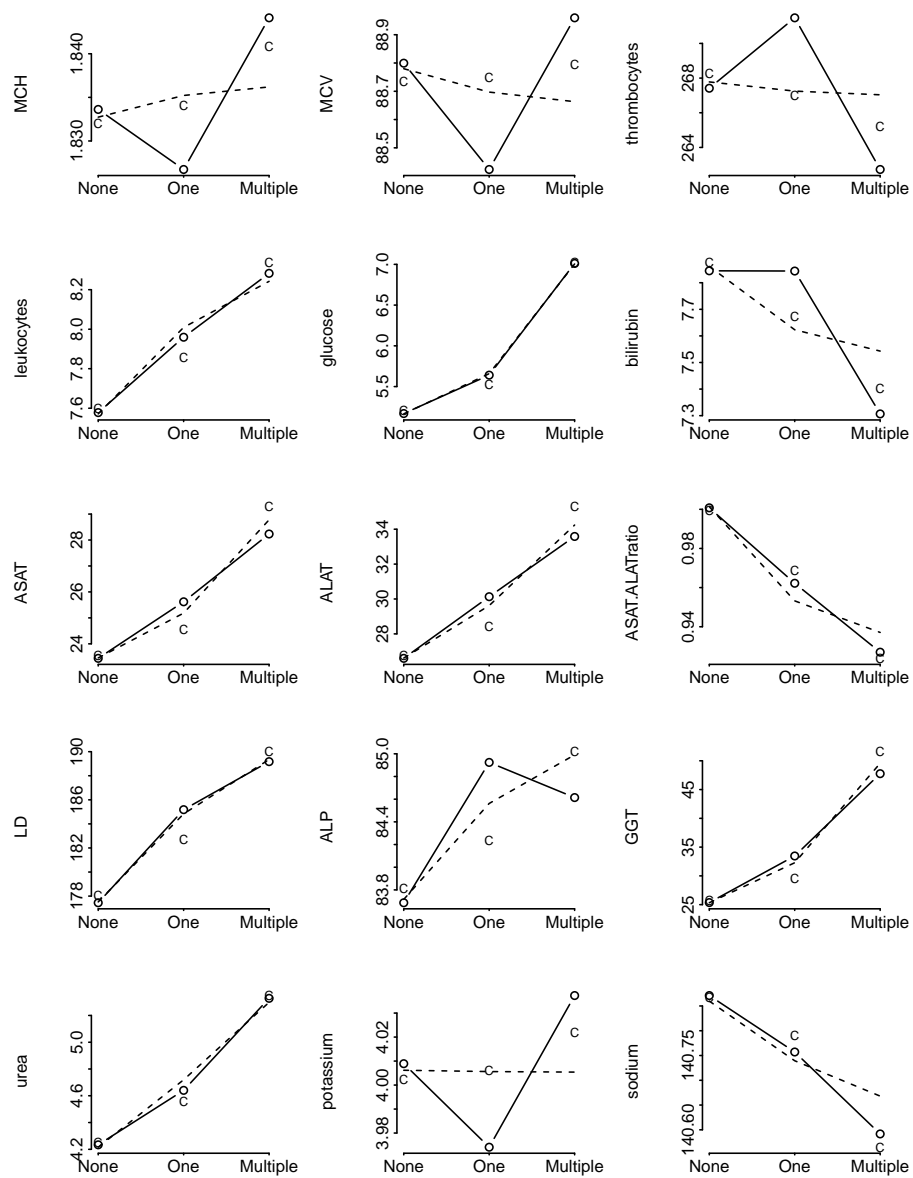


Figure E.2

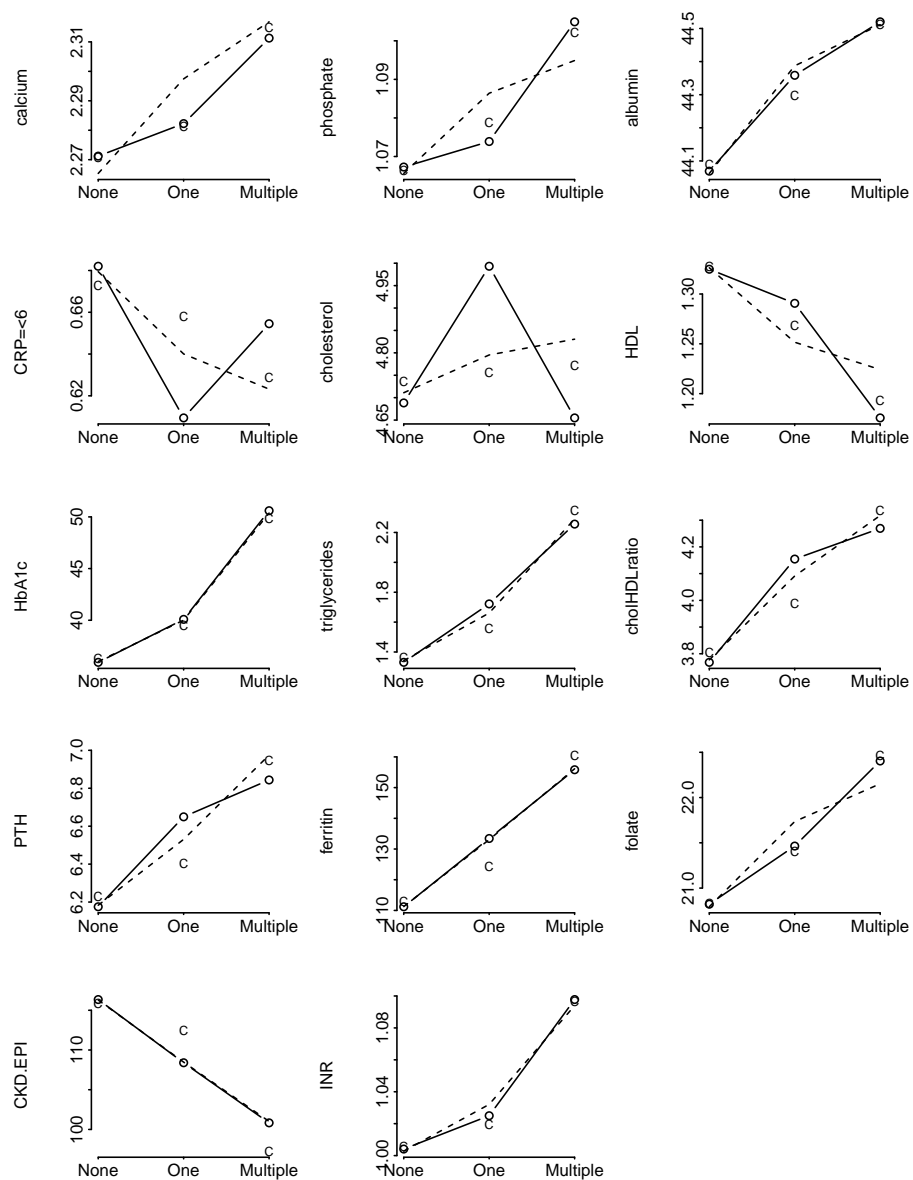


Figure E.3

F | K-Fold Cross Validation

When studying the performance of a model there is an inherent variability in the model generation itself [59]. It is not sufficient to generate a single model from the data and test it on replicates. A models that would memorize the labels would obtain a perfect prediction score but would fail to predict anything on new unseen data. The *generalization* performance of a model relates to its predictive capabilities on independent test data.

F.1 Cross-validation

The simplest way to asses the performance on unseen data is to split the data into a training and test set. The model is trained on the training set and the unseen test data is used to assess the performance. The performance can be assessed with a statistic of choice such as R^2 or AUC. A commonly used split is 70/30 where 70% of the data is used as training and 30% used as test. This split depends on the sample size. While data splitting provides an unbiased estimate of model performance a sacrifice is made in terms of sample size. If using 30% of the sample size as test set, the sample size for training is also reduced by 30%. To obtain nearly unbiased performance estimates without sacrificing sample size, cross validation can be used [54]. In K-fold cross validation the original dataset is randomly split in K roughly equal-sized parts. For the k th part the model is fitted to the $K - 1$ other parts of the data, this model is then used to make a prediction on the unseen k th part of the model. We do this for all $k = 1, 2, \dots, K$ and combine the K estimates of the desired performance statistic. An example of 5-fold cross validation is given in figure F.1. Typical choices of K are 5 or 10, although somewhat arbitrary, 5- and 10-fold cross-validation are recommended as a good compromise [54]. Larger K means less bias towards overestimating the true expected error because the training folds will be closer to the total dataset. However, larger K means higher variance and higher running time. The higher variance can be understood by the limiting case: $K = N - 1$ (also called leave-one-out cross-validation). Since there is only one data point being used for testing, the variance in the estimates of the model's error would be higher. This is especially true in the presence of outliers.

	i = 1	i = 2	i = 3	i = 4	i = 5
1	Training	Training	Training	Training	Test
2	Training	Training	Training	Test	Training
3	Training	Training	Test	Training	Training
4	Training	Test	Training	Training	Training
5	Test	Training	Training	Training	Training

Figure F.1: 5-Fold cross validation. The data is randomly split into 5 roughly equal parts. In the first iteration ($i = 1$) the data is trained on the first four parts of the data (1-4) and the last part (5) is used as unseen test data. The second iteration trains the data on parts 1,2,3,5 and uses part 4 as unseen test data. This process is repeated until $i = 5$.

F.2 Group k-fold

One important aspect of cross-validation is the splitting of the data in 10 roughly equals parts. Since we have about 5500 patient records we could create 10 folds, each containing about 550 records. However with this approach we violate the assumption that the test set has to be *unseen* data. It is possible that the screening record of a patient is included in the training set, and the follow-up record in the test set. This violates the independence assumption because we expect a certain correlation between repeated measurements on a single patient. To overcome this we sample from *patients* instead of *patient records*, also called group K-fold cross validation. Since we have about 2370 patients in our dataset, we create 10 splits each containing roughly 237 patients. For each patient we include all the records. This principle is illustrated in figure F.2

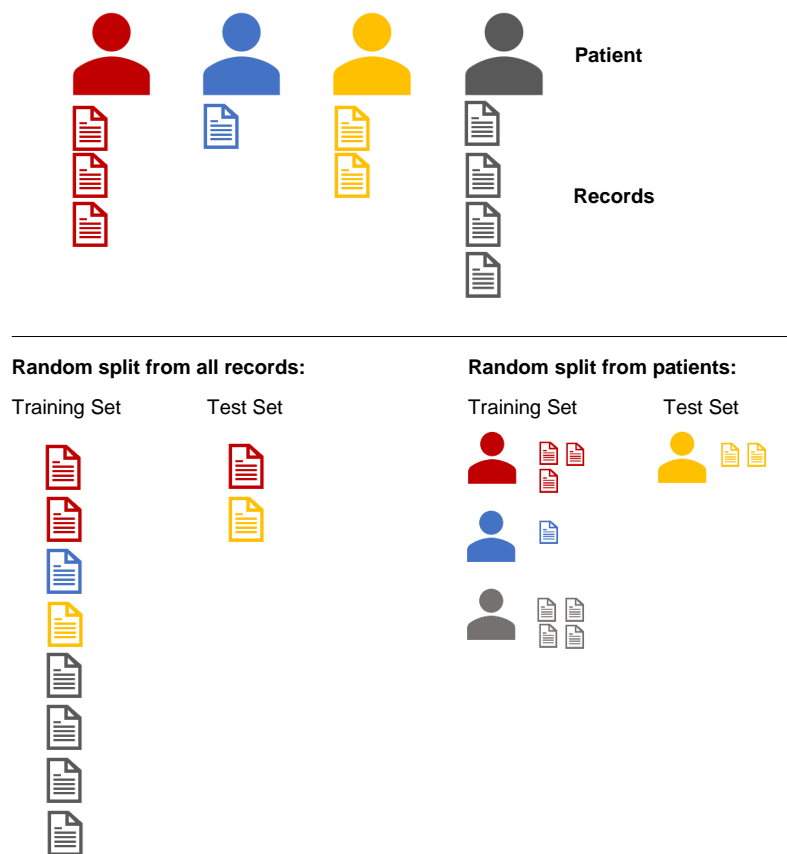


Figure F.2: Difference between sampling from all the records and sampling from patients. When sampling from all records it is possible that one measurement of a patient is included in the training and the other in the test set. Since we assume correlation between repeated measurements on the same patient, the training and test set are not independent. Therefore we sample from patients because we assume that measurements between patients are independent.

G | Performance Measures

G.1 Overall performance

G.1.1 R-squared

The R^2 (or R-squared) is the most common performance measure for continuous outcomes, it is an overall measure to quantify the amount of information in a given dataset. However, with logistic regression we have a binary outcome. To evaluate the fit of logistic models, several pseudo R^2 have been developed. These are called "pseudo" because they look like R^2 in the sense that they are on a scale of 0 to 1 with higher values indicating better model fit, but they cannot be interpreted as one would interpret an ordinary least square R^2 . Here we use Nagelkerke's R^2 :

$$R_N^2 = \frac{1 - (L_0/L)^{2/N}}{1 - L_0^{2/N}}$$

G.1.2 Brier score

The Brier score is simply defined as the sum of the squared differences between the actual outcomes Y and the predicted probabilities P :

$$B = \frac{1}{n} \sum_{i=1}^n (Y_i - P_i)^2$$

A score of 0 equals a perfectly accurate prediction, a score of 1 is totally inaccurate and a score of 0.25 is a model that always predicts $P = 0.5$. An example for a patient with or without diabetes:

- If the model predicts the patient has diabetes with a probability of 100% and the patient has diabetes, the Brier score is 0.
- If the model predicts the patient has diabetes with a probability of 100% and the patient does not have diabetes, the Brier score is 1.
- If the model predicts the patient has diabetes with a probability of 70% and the patient has diabetes, the Brier score is $(1 - 0.7)^2 = 0.09$

- If the model predicts the patient has diabetes with a probability of 50% the Brier score is always $0.5^2 = 0.25$.

G.2 Discriminative ability

G.2.1 Sensitivity

The sensitivity is also called the true positive rate or recall. It measures the fraction of patients that are positive and labeled as positive (true positive) divided by the total number of positives (true positives + false negatives):

$$\text{Sensitivity} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

G.2.2 Specificity

The specificity is also called the true negative rate. It measures the fraction of patients that are negative and labeled as negative (true negatives) divided by the total number of negatives (true negatives + false positives):

$$\text{Specificity} = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}}$$

G.2.3 Area under ROC curve

To determine whether a patient has to be classified as positive or negative, a cutoff has to be chosen. The default cutoff for probabilities is to classify patients with a $P > 0.5$ as positive. This cutoff can be applied to any value, not necessarily a probability. A ROC curve which plots the false positive rate against the true positive rate is often used in diagnostic research to quantify the diagnostic value of a test. An example can be seen in figure G.1 for HbA1c and diabetes. As the cutoff for HbA1c is varied, starting at the minimum value of 22 mmol/mol and increasing until the maximum of 124 mmol/mol, the specificity and sensitivity change. We start at a specificity of 1 and sensitivity of 0 (all patients are classified as negative) and end at sensitivity of 1 and specificity of 0 (all patients are classified as positive). The choice of the "best" cutoff can be made based on the Youden Index or the point that is the closest to the top-left. However, the cutoff can also be chosen to maximize either the sensitivity or specificity depending on the preference. This trade-off is visualized in a ROC curve. The area under the ROC curve (AUC) can be interpreted as the probability that a patient with the outcome is given a higher probability than a randomly chosen patient. A perfect model will have an AUC of 1 and a useless model, which is the same as flipping a coin, will have an AUC of 0.5. The AUC offers an advantage over the Brier score in the sense that it is independent of the incidence of the outcome.

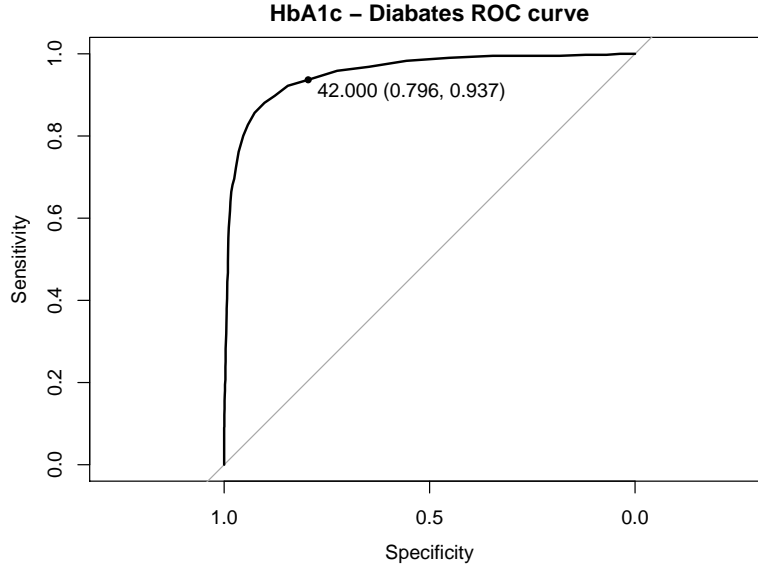


Figure G.1: ROC curve for HbA1c with outcome of diabetes. Patients with a HbA1c > 45 are classified as diabetic. In the graph we see this corresponds to a sensitivity of 0.796 and a specificity of 0.937.

G.2.4 Somers' D_{xy} rank correlation

The Somers' D_{xy} rank correlation is a measure of how many more concordant than discordant pairs exist, divided by the total number of pairs. If $D_{xy} = -1$ all pairs are discordant (i.e. disagree) if $D_{xy} = 1$ all pairs are concordant (i.e. agree). It is used as a quality measure of logistic or ordinal regression. Somers' D_{xy} gives us a measure of how much the prediction of the dependent (or outcome) variable improves when knowing the independent variable (or predictor). The Somers' D_{xy} is calculated from the concordant and discordant pairs as follows:

$$D_{xy} = \frac{N_s - N_r}{N_s + N_r + T_l}$$

N_s = the number of concordant pairs

N_r = the number of discordant pairs

T_l = the number of tied pairs on the independent variable

If we have two pairs (X_i, Y_i) and (X_j, Y_j) :

- the pairs are concordant if $X_i > X_j$ and $Y_i > Y_j$ or if $X_i < X_j$ and $Y_i < Y_j$
- the pairs are discordant if $X_i > X_j$ and $Y_i < Y_j$ or if $X_i < X_j$ and $Y_i > Y_j$

A Somers' D_{xy} of -1 means that all the pairs are discordant, +1 that all pairs are concordant. A D_{xy} of -1 or +1 implies a perfect correlation with the outcome, when $D_{xy} = 0$ there is no correlation or predictive value. One property of the Somers' D_{xy} is that it is related tot the AUC (or C-statistic): $AUC = \frac{D_{xy}+1}{2}$.

Example

For 6 patients we have a score X calculated by the model and an ordinal outcome label Y as shown in table G.1.

	Patient A	Patient B	Patient C	Patient D	Patient E	Patient F
Label	None	None	Multiple	One	Multiple	One
Score	1.06	-2.55	3.37	0.37	3.08	1.06

Table G.1: Example data.

We count the number of concordant pairs N_s , the number of discordant pairs N_r and the number of ties on the independent variable (score) T_l :

Concordant Pairs N_s	Discordant Pairs N_r	Ties on score T_l
Patient A and C	Patient A and D	Patient A and F
Patient A and E		
Patient B and C		
Patient B and D		
Patient B and E		
Patient B and F		
Patient C and D		
Patient C and F		
Patient D and E		
Patient E and F		

From this the Somers' D_{xy} and AUC are calculated:

$$D_{xy} = \frac{(10 - 1)}{10 + 1 + 1} = 0.75$$

$$AUC = \frac{0.75 + 1}{2} = 0.875$$

A large absolute D_{xy} suggests that the model has good predictive ability.

G.3 Calibration

Instead of simply classifying a patient as positive or negative we want to obtain a probability for this classification. An important but sometimes overlooked property of prediction models is calibration. Calibration refers to the agreement between observed

outcomes and predictions. I.e. if we predict that a patient has an 80% chance of having diabetes how can we be sure that this probability is actually 80% and not 70% or 90%? If the model is well calibrated the probability can be directly interpreted as a confidence interval.

G.3.1 Calibration plot

An example of a calibration plot for the hypertension logistic regression model containing age, calcium, potassium can be seen in figure G.2. Horizontally the predicted

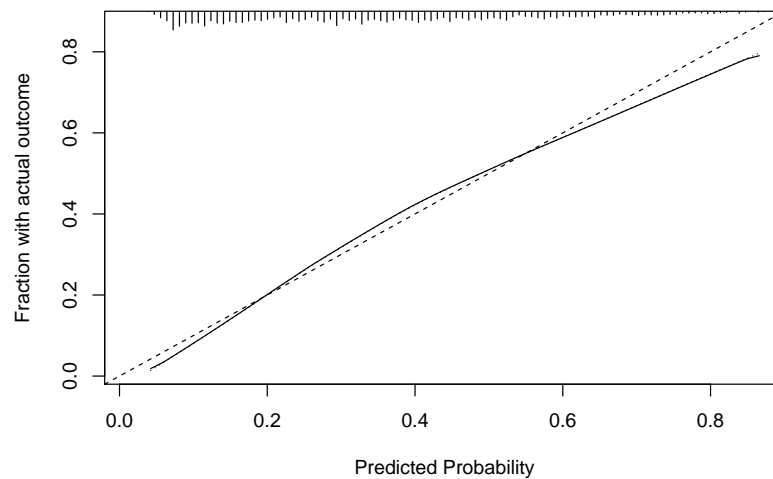


Figure G.2: Calibration plot.

probabilities of the model can be seen and vertically the actual probability or the fraction with the actual outcome. E.g. if the model predicts a probability of 0.6, there should be a fraction of 0.6 of all patients with hypertension. The straight dotted line from (0,0) to (1,1) represents a perfect/ideal fit. The solid line is a non-parametric fit between the predicted probabilities and the actual probabilities. The rug plot at the top of the chart reflects the relative occurrence of predictions of the corresponding probability. From this plot we can conclude that there is some mis-calibration, especially at the higher probabilities the predicted probability is higher than the actual probability. I.e. if we predict that patient has a 80% probability of having hypertension, the actual probability will be lower. Harrell notes that in a calibration plot overfitting can be observed when "Typically, low predictions will be too low and high predictions too high." [51].

G.3.2 Calibration slope and intercept

If the model is perfectly calibrated it coincides with the ideal 45° line as shown in figure G.2. The intercept of this line is 0 and the slope is 1. The slope and intercept of the predicted probability can be estimated with a logistic regression and using the linear predictor as the only coefficient:

$$\log \text{ odds (hypertension} = 1) = \alpha + \beta \times \text{linear predictor}$$

Where α is the intercept and β the slope. As stated above, ideally $\alpha = 0$ and $\beta = 1$. If the slope β is < 1 this is an indication of optimism: too low estimates for low predictions and too high estimates for high probabilities. We can see that our logistic regression model of hypertension is also suffers from optimism. A β larger than 1, indicates that the predicted probabilities are pessimistic, or not extreme enough. The intercept tells us something about a the systematic error. If the intercept is > 0 the probabilities are systematically too low and v.v. if the intercept is < 0 the probabilities are systematically too high.

G.4 Model parsimony

To asses which model is more parsimonious when it is approximated, the number of variables that are included in the final reduced model is a good measure. When the performance of two models is not significantly different, yet one model requires fewer predictors than the other, we prefer the model with fewer predictors because it is more parsimonious.

H | Difference in Patient Population between Periods

The number of patients that are included before surgery is 2367, at the 6, 12 and 24 month follow-up (FUP) after surgery only 1596, 1019 and 583 patients are included, respectively. Here we show whether the population of patients that is not included in each follow-up period differs at baseline from the population that is included. The p-values for comparing equal means of continuous variables are obtained using a two sided Student's t-Test, p-values for binary variables are obtained using a χ^2 test of equal proportions.

	In 6-month FUP (N = 1596)	Not in 6-month FUP (N = 771)	P-value
Mean age at surgery (years)	42.04	42.67	0.197
Surgery type = bypass (%)	44.9	49.3	0.051
Pre-op mean BMI (kg/m2)	43.47	43.90	0.092
Pre-op Diabetes (%)	3.3	3.8	0.668
Pre-op Diab. + Hypert. (%)	4.1	4.9	0.396
Pre-op Diab. + Dyslip. (%)	2.6	2.9	0.860
Pre-op Hypertension (%)	18.0	17.4	0.763
Pre-op Hypert. + Dyslip. (%)	5.8	6.2	0.724
Pre-op Dyslipidemia (%)	2.4	3.5	0.183
Pre-op Diab. + Hypert. + Dyslip. (%)	6.3	7.9	0.179
Pre-op No co-morbidities (%)	57.5	53.4	0.071
Pre-op One co-morbidity (%)	23.7	24.6	0.669
Pre-op Multiple co-morbidities (%)	18.8	21.9	0.083

Table H.1: Difference in patient population before surgery of patients that did attend the 6-month FUP and patients that did not. Significant P-values are marked with *.

	In 12-month FUP (N = 1019)	Not in 12-month FUP (N = 1348)	P-value
Gender female (%)	82.1	76.4	*0.001
Mean age at surgery (years)	42.03	42.40	0.431
Surgery type = bypass (%)	48.0	45.1	0.176
Pre-op mean BMI (kg/m2)	43.58	43.63	0.820
Pre-op Diabetes (%)	3.0	3.8	0.388
Pre-op Diab. + Hypert. (%)	3.8	4.7	0.325
Pre-op Diab. + Dyslip. (%)	2.7	2.7	1.000
Pre-op Hypertension (%)	18.9	16.9	0.222
Pre-op Hypert. + Dyslip. (%)	5.3	6.4	0.310
Pre-op Dyslipidemia (%)	2.2	3.3	0.136
Pre-op Diab. + Hypert. + Dyslip. (%)	6.6	7.0	0.713
Pre-op No co-morbidities (%)	57.4	55.2	0.301
Pre-op One co-morbidity (%)	24.1	24.0	0.958
Pre-op Multiple co-morbidities (%)	18.4	20.8	0.163

Table H.2: Difference in patient population before surgery of patients that did attend the 12-month FUP and patients that did not. Significant P-values are marked with *.

	In 24-month FUP (N = 583)	Not in 24-month FUP (N = 1784)	P-value
Gender female (%)	81.8	77.9	0.052
Mean age at surgery (years)	42.03	42.31	0.594
Surgery type = bypass (%)	46.7	46.2	0.901
Pre-op mean BMI (kg/m2)	43.89	43.52	0.165
Pre-op Diabetes (%)	2.4	3.8	0.137
Pre-op Diab. + Hypert. (%)	4.5	4.3	0.976
Pre-op Diab. + Dyslip. (%)	2.7	2.7	1.000
Pre-op Hypertension (%)	14.2	18.9	*0.012
Pre-op Hypert. + Dyslip. (%)	5.8	5.9	1.000
Pre-op Dyslipidemia (%)	2.6	2.9	0.827
Pre-op Diab. + Hypert. + Dyslip. (%)	6.9	6.8	1.000
Pre-op No co-morbidities (%)	60.9	54.6	*0.009
Pre-op One co-morbidity (%)	19.2	25.6	*0.002
Pre-op Multiple co-morbidities (%)	19.9	19.8	1.000

Table H.3: Difference in patient population before surgery of patients that did attend the 24-month FUP and patients that did not. Significant P-values are marked with *.

I | Influential Observations

A measure that quantifies how much impact each observation has on a particular predictor is DFBETAS. The DFBETA for a particular observation is the difference between the regression coefficient calculated from the entire data, and the regression coefficient calculated from the data with the observation deleted, scaled by the standard error calculated with the observation deleted. As stated in the equation:

$$\text{DFBETAS}((-i), \text{fit}) = \frac{\hat{b}_k - \hat{b}_{k(-i)}}{\text{SE} \hat{b}_{k(-i)}}$$

Where \hat{b}_k is the k -th regression coefficient and $\hat{b}_{k(-i)}$ stands for the k -th regression coefficient after deleting the i -th observation from the data. In table [I.1](#) all the influential observations are shown with $|\text{DFBETAS}| > 2$ by using the `which.influence` function in the `rms` package by Harrell.

	Count	hemoglobin	erythrocytes	MCH	MCV	leukocytes	glucose	bilirubin	ASAT	ALAT	ASAT:ALATratio	LD	GGT	urea	calcium	albumin	cholesterol	HbA1c	cholHDLratio	PTH	iron	ferritin	folate	INR	comorb
150	1	8.5	4.8	1.78	85	10.6	4.4	6.7	* 240	110	2.18	190	180	4.4	2.34	45	6.94	38	6.7	2.5	15.0	293	14.0	0.97	Multiple
241	1	9.4	5.3	1.76	85	8.6	5.2	13.0	360	510	0.71	330	* 1700	2.7	2.57	40	4.90	37	4.9	4.4	17.0	565	35.0	0.95	None
343	1	9.6	5.4	1.78	83	10.4	5.2	4.2	61	* 160	0.38	189	77	3.6	2.28	46	4.48	40	4.9	4.9	10.0	72	17.0	1.01	One
442	1	10.5	5.2	2.02	92	10.6	4.9	18.0	33	27	1.22	312	44	5.5	2.36	43	3.19	44	3.2	6.8	13.0	268	17.0	* 3.28	None
538	1	8.8	4.3	2.05	99	8.8	5.2	5.0	35	38	0.92	185	158	3.6	2.30	49	7.16	44	* 12.0	5.0	14.0	82	9.8	0.98	None
734	1	10.0	5.4	1.84	86	7.7	5.2	7.3	78	* 190	0.41	232	63	5.1	2.32	48	6.16	43	5.7	8.6	14.0	419	17.0	0.99	One
922	1	10.5	5.5	1.91	92	8.8	5.5	7.6	36	50	0.72	200	78	5.5	2.31	48	4.78	45	5.4	7.3	18.0	257	21.0	* 3.31	None
979	1	9.3	5.0	1.85	88	8.0	* 17.2	9.3	50	69	0.72	213	54	5.8	2.31	45	5.08	77	4.9	6.3	19.0	571	14.0	1.02	None
1018	1	9.3	4.4	2.12	101	* 27.1	13.7	6.0	24	73	0.33	226	101	3.3	2.35	45	5.79	67	4.9	8.8	17.0	62	25.0	0.95	None
1025	1	9.5	5.4	1.78	85	11.5	5.4	3.0	25	34	0.74	203	30	3.0	2.12	46	5.94	41	* 12.0	4.4	12.0	38	18.0	0.91	None
1413	1	11.0	5.6	1.95	91	9.5	6.0	17.0	30	60	0.50	181	117	2.3	2.15	42	4.26	66	4.7	7.6	14.0	333	11.0	* 3.13	None
1439	1	10.9	5.6	1.94	90	11.4	5.1	8.3	34	48	0.71	199	39	4.3	2.27	46	3.52	35	3.5	6.1	34.0	* 1580	14.0	0.97	One
2033	4	* 5.0	* 5.0	* 1.10	* 56	10.4	6.0	2.6	18	21	0.86	201	24	4.7	2.13	45	5.72	49	7.3	13.0	3.7	5	8.6	1.03	None
2049	1	9.3	5.0	1.84	85	7.6	4.8	8.2	39	50	0.78	* 439	42	4.1	2.22	48	6.16	37	6.1	8.4	16.0	398	9.4	0.99	One
2192	2	9.6	4.8	2.01	99	10.2	6.7	3.0	21	30	0.70	193	32	4.9	2.22	44	* 5.96	38	* 12.0	8.7	9.0	65	8.2	0.95	One
2202	1	8.4	5.3	1.59	81	7.2	4.9	5.9	20	18	1.11	181	24	2.7	2.30	43	5.02	40	4.6	5.0	* 51.0	24	12.0	1.06	One
2530	1	7.9	4.1	1.94	101	6.9	10.3	7.3	15	23	0.65	255	80	* 27.0	2.32	41	4.33	67	3.7	3.9	17.0	195	22.0	2.48	None
2680	1	9.3	5.3	1.74	83	7.1	* 19.4	8.1	17	27	0.63	150	133	6.1	2.34	49	4.36	82	4.1	3.3	18.0	125	24.0	0.90	None
2809	1	9.3	4.8	1.96	91	9.5	* 18.3	5.4	17	29	0.59	198	32	4.2	2.26	43	4.51	92	4.1	7.2	18.0	51	34.0	2.72	None
3353	2	9.3	5.0	1.84	88	8.1	* 21.8	9.9	26	38	0.68	179	32	3.6	2.32	43	3.51	* 112	1.5	6.7	18.0	198	45.0	1.01	None
3384	1	11.4	5.8	1.97	91	5.6	4.4	15.0	21	27	0.78	180	21	4.7	2.29	52	4.13	32	4.7	6.3	* 49.0	262	11.0	1.01	One
3581	1	7.3	4.0	1.82	84	9.0	10.2	5.3	52	56	0.93	213	87	5.8	2.19	* 29	3.20	72	4.2	5.1	18.0	293	36.0	0.97	None
3910	1	11.3	6.2	1.81	86	6.3	5.4	* 46.0	20	15	1.33	187	17	3.6	2.35	44	4.77	29	3.2	6.2	25.0	268	18.0	1.13	One
3967	1	9.1	4.7	1.94	94	8.7	* 18.1	6.6	21	26	0.81	196	34	4.6	2.33	42	4.23	86	3.7	8.8	19.0	39	27.0	2.19	None
4147	1	7.9	4.3	1.82	89	4.1	5.1	* 54.0	33	45	0.73	157	41	7.9	2.29	42	3.90	33	2.8	2.8	18.0	181	35.0	1.07	None
4815	1	8.5	5.2	1.62	82	5.1	4.8	6.3	240	* 300	0.80	443	180	5.8	2.19	47	4.66	37	2.8	6.5	7.9	57	12.0	0.98	One
4857	1	8.5	4.7	1.81	85	7.2	13.5	5.5	20	29	0.69	183	18	3.5	2.30	39	3.42	* 94	1.6	7.3	8.8	75	40.0	0.95	None
4860	6	8.2	4.6	1.78	88	8.6	5.2	22.0	* 870	* 670	* 1.30	* 578	* 121	5.8	2.21	46	4.85	37	4.4	5.4	20.0	429	* 13.0	1.02	None
4879	1	10.0	5.0	2.00	95	9.1	7.7	11.0	30	16	1.88	172	38	4.0	2.40	42	2.76	41	2.8	4.9	16.0	175	41.0	* 4.56	None
4908	2	9.5	5.2	1.84	86	5.4	10.0	6.9	29	27	1.07	222	22	9.8	* 2.74	46	5.95	54	3.8	* 28.0	14.0	353	28.0	0.98	None
5252	1	6.5	3.5	1.85	93	9.8	6.0	5.1	29	32	0.91	185	133	5.8	2.14	* 28	3.75	42	2.7	11.0	14.0	166	24.0	1.07	Multiple
5322	1	8.2	4.5	1.83	89	8.6	9.9	5.5	23	25	0.92	188	37	5.4	2.29	44	4.04	58	3.5	8.4	15.0	21	33.0	* 2.97	None
5433	1	8.6	4.9	1.75	94	5.1	4.6	8.7	31	23	1.35	358	48	4.1	2.27	47	4.94	38	4.0	11.0	14.0	668	19.0	* 3.54	One
5453	1	7.6	3.8	1.98	92	8.0	14.7	3.0	19	24	0.79	167	* 553	6.0	2.28	45	6.81	68	6.7	3.4	22.0	120	40.0	1.05	None

Table I.1: Influential observations as determined by DFBETAS. The first column gives the row number of the record and the "Count" column gives the number of influential observations in that record. Influential observations are marked with *.

J | Full Models

J.1 Diabetes Logistic Regression Model

		Model Likelihood Ratio Test	Discrimination Indexes	Rank Discrim. Indexes
Obs	5565	LR χ^2 2136.64	R^2 0.606	C 0.938
FALSE	4880	d.f. 41	g 2.340	D_{xy} 0.875
TRUE	685	Pr($> \chi^2$) <0.0001	g_r 10.385	γ 0.878
Cluster on df.all.correct\$patient			g_p 0.183	τ_a 0.189
Clusters	2367		Brier 0.047	
$\max \frac{\partial \log L}{\partial \beta} $	1×10^{-6}			

	$\hat{\beta}$	S.E.	Wald Z	Pr(> Z)
Intercept	0.7987	10.4219	0.08	0.9389
period=6 M	-1.9736	0.8517	-2.32	0.0205
period=12 M	-3.5476	1.0621	-3.34	0.0008
period=24 M	-3.7006	2.1092	-1.75	0.0794
BMI	-0.0324	0.0172	-1.89	0.0592
gender=F	-0.2702	0.2707	-1.00	0.3181
age.at.surgery	0.0396	0.0106	3.71	0.0002
hemoglobin	-0.2088	1.2634	-0.17	0.8687
hematocrit	34.9821	18.0380	1.94	0.0525
erythrocytes	-2.7183	1.7524	-1.55	0.1209
MCH	-0.5708	6.1168	-0.09	0.9257
MCV	-0.1377	0.0900	-1.53	0.1262
thrombocytes	-0.0003	0.0013	-0.23	0.8144
leukocytes	-0.0283	0.0440	-0.64	0.5200
glucose	0.0598	0.0834	0.72	0.4735
bilirubin	0.0014	0.0182	0.08	0.9376
ASAT	0.0005	0.0122	0.04	0.9644
ALAT	0.0001	0.0072	0.02	0.9842
ASAT.ALATratio	-0.0214	0.2778	-0.08	0.9387
LD	-0.0056	0.0027	-2.09	0.0366

	$\hat{\beta}$	S.E.	Wald Z	Pr(> Z)
ALP	-0.0042	0.0036	-1.16	0.2454
GGT	0.0011	0.0031	0.35	0.7226
urea	0.1158	0.0580	2.00	0.0460
potassium	0.1696	0.2366	0.72	0.4735
sodium	0.0157	0.0412	0.38	0.7038
calcium	0.8965	0.9007	1.00	0.3196
phosphate	0.4795	0.4168	1.15	0.2500
albumin	0.0021	0.0330	0.06	0.9490
CRP= ≥ 6	-0.0962	0.1744	-0.55	0.5812
cholesterol	-0.1012	0.2091	-0.48	0.6283
HDL	-1.2761	0.7720	-1.65	0.0983
HbA1c	0.2077	0.0179	11.59	<0.0001
triglycerides	0.4176	0.0954	4.38	<0.0001
cholHDLratio	-0.4965	0.2276	-2.18	0.0292
PTH	-0.0147	0.0144	-1.02	0.3070
ferritin	-0.0008	0.0007	-1.16	0.2479
folate	0.0042	0.0082	0.51	0.6075
CKD.EPI	0.0018	0.0076	0.24	0.8099
INR	-0.0530	0.2577	-0.21	0.8369
period=6 M \times BMI	0.0612	0.0224	2.73	0.0063
period=12 M \times BMI	0.1016	0.0306	3.32	0.0009
period=24 M \times BMI	0.0927	0.0647	1.43	0.1519

J.2 Hypertension Logistic Regression Model

		Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrim. Indexes	
Obs	5565	LR χ^2	1456.95	R^2	0.335	C	0.812
FALSE	4065	d.f.	41	g	1.572	D_{xy}	0.624
TRUE	1500	Pr(> χ^2)	<0.0001	g_r	4.819	γ	0.625
Cluster on df.all.correct\$patient				g_p	0.243	τ_a	0.246
Clusters	2367			Brier	0.148		
max $ \frac{\partial \log L}{\partial \beta} $	9×10^{-7}						

	$\hat{\beta}$	S.E.	Wald Z	Pr(> Z)
Intercept	-3.4717	6.2138	-0.56	0.5764
period=6 M	-1.5833	0.4683	-3.38	0.0007
period=12 M	-3.0646	0.6842	-4.48	<0.0001
period=24 M	-2.7078	0.7751	-3.49	0.0005
BMI	-0.0073	0.0100	-0.73	0.4673
gender=F	-0.2540	0.1646	-1.54	0.1228
age.at.surgery	0.0731	0.0072	10.15	<0.0001

	$\hat{\beta}$	S.E.	Wald Z	Pr(> Z)
hemoglobin	-1.1384	0.7326	-1.55	0.1202
hematocrit	13.8911	10.6550	1.30	0.1923
erythrocytes	0.9331	1.0156	0.92	0.3582
MCH	5.9162	3.5118	1.68	0.0921
MCV	-0.1049	0.0523	-2.00	0.0451
thrombocytes	0.0016	0.0009	1.75	0.0805
leukocytes	0.0833	0.0258	3.22	0.0013
glucose	-0.0047	0.0343	-0.14	0.8907
bilirubin	0.0043	0.0121	0.36	0.7196
ASAT	0.0101	0.0059	1.71	0.0869
ALAT	-0.0056	0.0047	-1.19	0.2344
ASAT.ALATratio	-0.0063	0.1926	-0.03	0.9737
LD	-0.0019	0.0019	-1.02	0.3084
ALP	-0.0048	0.0022	-2.19	0.0284
GGT	0.0018	0.0013	1.38	0.1688
urea	0.1306	0.0403	3.24	0.0012
potassium	-1.0230	0.1646	-6.22	<0.0001
sodium	-0.0289	0.0221	-1.31	0.1908
calcium	1.8847	0.5655	3.33	0.0009
phosphate	0.0490	0.2760	0.18	0.8590
albumin	0.0576	0.0206	2.80	0.0051
CRP= \geq 6	0.0433	0.1042	0.42	0.6777
cholesterol	-0.1230	0.1057	-1.16	0.2446
HDL	-0.1889	0.3659	-0.52	0.6057
HbA1c	0.0166	0.0066	2.52	0.0119
triglycerides	0.2247	0.0644	3.49	0.0005
cholHDLratio	-0.1135	0.1106	-1.03	0.3048
PTH	0.0416	0.0177	2.36	0.0183
ferritin	-0.0001	0.0004	-0.33	0.7421
folate	0.0027	0.0052	0.52	0.6065
CKD.EPI	-0.0180	0.0050	-3.59	0.0003
INR	0.0646	0.2134	0.30	0.7620
period=6 M \times BMI	0.0406	0.0127	3.20	0.0014
period=12 M \times BMI	0.0853	0.0206	4.15	<0.0001
period=24 M \times BMI	0.0722	0.0230	3.15	0.0017

J.3 Dyslipidemia Logistic Regression Model

		Model Likelihood Ratio Test	Discrimination Indexes	Rank Discrim. Indexes
Obs	5565	LR χ^2 1051.90	R^2 0.308	C 0.823
FALSE	4773	d.f. 41	g 1.467	D_{xy} 0.647
TRUE	792	Pr($> \chi^2$) <0.0001	g_r 4.338	γ 0.649
Cluster on df.all.correct\$patient			g_p 0.155	τ_a 0.158
Clusters	2367		Brier 0.093	
$\max \frac{\partial \log L}{\partial \beta} $	2×10^{-5}			

	$\hat{\beta}$	S.E.	Wald Z	Pr(> Z)
Intercept	-13.2646	7.7012	-1.72	0.0850
period=6 M	-1.4202	0.6753	-2.10	0.0355
period=12 M	-2.1908	0.7862	-2.79	0.0053
period=24 M	-2.7175	0.9788	-2.78	0.0055
BMI	-0.0227	0.0130	-1.75	0.0802
gender=F	-0.0342	0.2071	-0.17	0.8688
age.at.surgery	0.0610	0.0088	6.90	<0.0001
hemoglobin	0.0741	0.8581	0.09	0.9312
hematocrit	11.8530	13.2913	0.89	0.3725
erythrocytes	-1.2182	1.1985	-1.02	0.3094
MCH	1.0882	4.1144	0.26	0.7914
MCV	-0.0448	0.0659	-0.68	0.4966
thrombocytes	0.0002	0.0011	0.21	0.8368
leukocytes	0.0431	0.0304	1.42	0.1564
glucose	-0.0664	0.0383	-1.73	0.0829
bilirubin	-0.0084	0.0146	-0.58	0.5652
ASAT	0.0106	0.0053	2.00	0.0457
ALAT	-0.0088	0.0060	-1.48	0.1382
ASAT.ALATratio	-0.0024	0.2222	-0.01	0.9915
LD	-0.0022	0.0022	-0.99	0.3199
ALP	-0.0109	0.0027	-3.98	<0.0001
GGT	0.0060	0.0014	4.28	<0.0001
urea	-0.0218	0.0431	-0.51	0.6130
potassium	-0.1139	0.1808	-0.63	0.5286
sodium	0.0459	0.0296	1.55	0.1206
calcium	1.7103	0.7241	2.36	0.0182
phosphate	0.5637	0.3378	1.67	0.0952
albumin	0.0768	0.0247	3.11	0.0019
CRP= ≥ 6	-0.3907	0.1332	-2.93	0.0034
cholesterol	-0.0739	0.1379	-0.54	0.5920
HDL	-1.0892	0.4752	-2.29	0.0219
HbA1c	0.0524	0.0072	7.26	<0.0001

	$\hat{\beta}$	S.E.	Wald Z	Pr(> Z)
triglycerides	0.2731	0.0796	3.43	0.0006
cholHDLratio	-0.2341	0.1346	-1.74	0.0820
PTH	0.0264	0.0159	1.66	0.0965
ferritin	0.0001	0.0005	0.19	0.8474
folate	0.0080	0.0064	1.24	0.2137
CKD.EPI	-0.0084	0.0061	-1.39	0.1638
INR	-0.0155	0.1931	-0.08	0.9362
period=6 M \times BMI	0.0289	0.0187	1.54	0.1227
period=12 M \times BMI	0.0525	0.0233	2.25	0.0242
period=24 M \times BMI	0.0719	0.0284	2.53	0.0114

J.4 Proportional Odds Model

		Model Likelihood Ratio Test	Discrimination Indexes	Rank Discrim. Indexes
Obs	5565	LR χ^2 2391.75	R^2 0.422	C 0.823
None	3645	d.f. 41	g 1.833	D_{xy} 0.645
One	1124	Pr(> χ^2) <0.0001	g_r 6.250	γ 0.646
Multiple	796		g_p 0.296	τ_a 0.329
Cluster on df.all.correct\$patient			Brier 0.153	
Clusters	2367			
max $ \frac{\partial \log L}{\partial \beta} $	1×10^{-7}			

	$\hat{\beta}$	S.E.	Wald Z	Pr(> Z)
y \geq One	-10.2419	5.5229	-1.85	0.0637
y \geq Multiple	-11.9792	5.5272	-2.17	0.0302
period=6 M	-1.9131	0.4570	-4.19	<0.0001
period=12 M	-3.3327	0.6020	-5.54	<0.0001
period=24 M	-3.0700	0.7347	-4.18	<0.0001
BMI	-0.0235	0.0094	-2.51	0.0121
gender=F	-0.1331	0.1496	-0.89	0.3737
age.at.surgery	0.0671	0.0067	10.06	<0.0001
hemoglobin	-0.9279	0.6578	-1.41	0.1584
hematocrit	15.0381	9.9800	1.51	0.1319
erythrocytes	0.4708	0.8578	0.55	0.5831
MCH	5.0897	3.1519	1.61	0.1064
MCV	-0.0896	0.0497	-1.80	0.0714
thrombocytes	0.0008	0.0008	0.95	0.3404
leukocytes	0.0533	0.0238	2.24	0.0252
glucose	0.0220	0.0376	0.58	0.5595
bilirubin	0.0086	0.0101	0.85	0.3943
ASAT	0.0095	0.0059	1.61	0.1084

	$\hat{\beta}$	S.E.	Wald Z	Pr(> Z)
ALAT	-0.0065	0.0047	-1.38	0.1675
ASAT.ALATratio	-0.0155	0.1717	-0.09	0.9279
LD	-0.0022	0.0017	-1.31	0.1887
ALP	-0.0062	0.0020	-3.13	0.0017
GGT	0.0031	0.0013	2.43	0.0153
urea	0.0937	0.0384	2.44	0.0146
potassium	-0.5908	0.1414	-4.18	<0.0001
sodium	0.0113	0.0210	0.54	0.5908
calcium	1.3689	0.5413	2.53	0.0114
phosphate	0.2202	0.2485	0.89	0.3756
albumin	0.0522	0.0191	2.73	0.0064
CRP= ≥ 6	-0.0941	0.0953	-0.99	0.3235
cholesterol	-0.1270	0.1024	-1.24	0.2149
HDL	-0.4950	0.3552	-1.39	0.1635
HbA1c	0.0851	0.0074	11.54	<0.0001
triglycerides	0.3026	0.0645	4.69	<0.0001
cholHDLratio	-0.1600	0.1072	-1.49	0.1355
PTH	0.0294	0.0158	1.86	0.0630
ferritin	-0.0002	0.0004	-0.49	0.6258
folate	0.0033	0.0047	0.69	0.4889
CKD.EPI	-0.0133	0.0046	-2.87	0.0041
INR	0.1065	0.1680	0.63	0.5264
period=6 M \times BMI	0.0480	0.0123	3.89	<0.0001
period=12 M \times BMI	0.0914	0.0180	5.07	<0.0001
period=24 M \times BMI	0.0789	0.0218	3.62	0.0003

J.5 Penalized extended Continuation Ratio Model

Penalty factors
simple nonlinear interaction nonlinear.interaction
0.5 0.5 4 4

		Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrim. Indexes	
Obs	7485	LR χ^2	2549.04	R^2	0.393	C	0.824
0	2716	d.f.	72.568	g	1.778	D_{xy}	0.649
1	4769	Pr(> χ^2)	<0.0001	g_r	5.918	γ	0.650
Cluster on df.all.correct[u\$subs,]\$patient		Penalty	17.03	g_p	0.295	τ_a	0.300
Clusters	2367			Brier	0.159		
max $ \frac{\partial \log L}{\partial \beta} $	1×10^{-10}						

	$\hat{\beta}$	S.E.	Wald Z	Pr(> Z)	Penalty Scale
Intercept	8.6436	4.7833	1.81	0.0708	0.0000
cohort=Y \geq One	0.0560	0.8288	0.07	0.9462	0.5000
period=6 M	1.3135	0.3423	3.84	0.0001	0.6124
period=12 M	2.3094	0.4502	5.13	<0.0001	0.6124
period=24 M	2.1365	0.5194	4.11	<0.0001	0.6124
BMI	0.0175	0.0093	1.88	0.0604	5.8265
gender=F	0.1425	0.1539	0.93	0.3544	0.5000
age.at.surgery	-0.0650	0.0068	-9.60	<0.0001	7.7647
hemoglobin	0.5060	0.5095	0.99	0.3207	0.5316
hematocrit	-11.6349	8.2292	-1.41	0.1574	0.0225
erythrocytes	0.0062	0.6907	0.01	0.9928	0.2820
MCH	-3.0671	2.4403	-1.26	0.2088	0.0809
MCV	0.0780	0.0411	1.90	0.0576	3.3061
thrombocytes	-0.0010	0.0009	-1.11	0.2664	45.1632
leukocytes	-0.0436	0.0255	-1.71	0.0880	1.5928
glucose	-0.0627	0.0413	-1.52	0.1288	1.3243
bilirubin	-0.0146	0.0110	-1.33	0.1823	3.3329
ASAT	-0.0087	0.0053	-1.65	0.0996	11.7712
ALAT	0.0055	0.0042	1.30	0.1929	16.0204
ASAT.ALATratio	0.0065	0.1729	0.04	0.9700	0.2210
LD	0.0022	0.0018	1.26	0.2092	24.2574
ALP	0.0053	0.0021	2.51	0.0120	17.5909
GGT	-0.0024	0.0013	-1.84	0.0665	28.8170
urea	-0.0948	0.0400	-2.37	0.0179	1.1612
potassium	0.5771	0.1370	4.21	<0.0001	0.2125
sodium	-0.0126	0.0189	-0.67	0.5039	1.5369
calcium	-1.2441	0.4896	-2.54	0.0110	0.0591
phosphate	-0.0462	0.2506	-0.18	0.8536	0.1212
albumin	-0.0463	0.0177	-2.61	0.0091	1.8634
CRP= \geq 6	0.0027	0.1021	0.03	0.9791	0.5000
cholesterol	0.0969	0.0962	1.01	0.3136	0.7160
HDL	0.2905	0.3254	0.89	0.3720	0.2499
HbA1c	-0.0872	0.0085	-10.30	<0.0001	7.8818
triglycerides	-0.3410	0.0673	-5.06	<0.0001	0.8955
cholHDLratio	0.0941	0.0990	0.95	0.3417	1.0129
PTH	-0.0320	0.0173	-1.85	0.0641	2.5735
ferritin	0.0002	0.0004	0.60	0.5478	92.4228
folate	-0.0032	0.0049	-0.66	0.5103	7.1381
CKD.EPI	0.0117	0.0045	2.62	0.0087	11.7200
INR	-0.1435	0.2174	-0.66	0.5093	0.2133

	$\hat{\beta}$	S.E.	Wald Z	Pr(> Z)	Penalty Scale
period=6 M \times BMI	-0.0331	0.0093	-3.55	0.0004	29.2336
period=12 M \times BMI	-0.0627	0.0138	-4.55	<0.0001	22.5633
period=24 M \times BMI	-0.0529	0.0156	-3.39	0.0007	18.0721
cohort=Y \geq One \times period=6 M	-0.3872	0.5173	-0.75	0.4541	0.4928
cohort=Y \geq One \times period=12 M	-0.0312	0.7344	-0.04	0.9661	0.3669
cohort=Y \geq One \times period=24 M	-0.6805	1.1433	-0.60	0.5517	0.2662
cohort=Y \geq One \times BMI	-0.0056	0.0120	-0.47	0.6404	34.1315
cohort=Y \geq One \times gender=F	-0.1429	0.2481	-0.58	0.5646	0.7756
cohort=Y \geq One \times age.at.surgery	0.0249	0.0105	2.38	0.0175	43.2359
cohort=Y \geq One \times hemoglobin	-0.0632	0.0504	-1.25	0.2102	7.6338
cohort=Y \geq One \times hematocrit	-0.3440	1.0432	-0.33	0.7416	0.3685
cohort=Y \geq One \times erythrocytes	0.2056	0.0988	2.08	0.0375	4.1689
cohort=Y \geq One \times MCH	-0.5184	0.2555	-2.03	0.0425	1.6063
cohort=Y \geq One \times MCV	-0.0081	0.0052	-1.56	0.1181	77.5817
cohort=Y \geq One \times thrombocytes	0.0006	0.0014	0.41	0.6826	243.0871
cohort=Y \geq One \times leukocytes	0.0152	0.0418	0.36	0.7168	7.4363
cohort=Y \geq One \times glucose	0.0664	0.0541	1.23	0.2194	5.9105
cohort=Y \geq One \times bilirubin	0.0350	0.0206	1.70	0.0898	8.0678
cohort=Y \geq One \times ASAT	-0.0090	0.0088	-1.02	0.3058	27.8842
cohort=Y \geq One \times ALAT	0.0081	0.0073	1.11	0.2678	35.4874
cohort=Y \geq One \times ASAT.ALATratio	0.2533	0.3156	0.80	0.4223	0.8812
cohort=Y \geq One \times LD	-0.0002	0.0026	-0.08	0.9338	167.1642
cohort=Y \geq One \times ALP	0.0006	0.0036	0.16	0.8734	77.8414
cohort=Y \geq One \times GGT	-0.0024	0.0027	-0.88	0.3796	54.2215
cohort=Y \geq One \times urea	0.0330	0.0559	0.59	0.5543	4.7245
cohort=Y \geq One \times potassium	-0.4153	0.1574	-2.64	0.0083	3.5096
cohort=Y \geq One \times sodium	-0.0016	0.0024	-0.67	0.5038	122.9099
cohort=Y \geq One \times calcium	-0.5597	0.2384	-2.35	0.0189	2.0058
cohort=Y \geq One \times phosphate	-0.5931	0.3910	-1.52	0.1293	0.9660
cohort=Y \geq One \times albumin	0.0123	0.0145	0.85	0.3960	38.9015
cohort=Y \geq One \times CRP= \geq 6	0.4575	0.1753	2.61	0.0091	0.5875
cohort=Y \geq One \times cholesterol	0.1878	0.1057	1.78	0.0756	4.3743
cohort=Y \geq One \times HDL	0.4683	0.3838	1.22	0.2224	1.1430
cohort=Y \geq One \times HbA1c	0.0239	0.0111	2.14	0.0322	41.1985
cohort=Y \geq One \times triglycerides	0.1741	0.0935	1.86	0.0627	2.3201
cohort=Y \geq One \times cholHDLratio	0.0876	0.1045	0.84	0.4022	3.9942
cohort=Y \geq One \times PTH	0.0201	0.0246	0.81	0.4153	7.3930
cohort=Y \geq One \times ferritin	-0.0004	0.0007	-0.57	0.5695	195.0627
cohort=Y \geq One \times folate	0.0046	0.0088	0.53	0.5974	21.5783
cohort=Y \geq One \times CKD.EPI	-0.0027	0.0052	-0.52	0.6003	93.6284

	$\hat{\beta}$	S.E.	Wald Z	Pr(> Z)	Penalty Scale
cohort=Y \geq One \times INR	0.0589	0.2921	0.20	0.8403	1.0009
cohort=Y \geq One \times period=6 M \times BMI	0.0092	0.0149	0.61	0.5386	16.2292
cohort=Y \geq One \times period=12 M \times BMI	-0.0011	0.0231	-0.05	0.9635	11.3714
cohort=Y \geq One \times period=24 M \times BMI	0.0120	0.0352	0.34	0.7335	8.4695

Appendix References

1. Thomas, M. C., Tsalamandris, C., MacIsaac, R. J. & Jerums, G. The epidemiology of hemoglobin levels in patients with type 2 diabetes. *American journal of kidney diseases* **48**, 537–545 (2006).
2. Hanley, A. J. *et al.* Association of hematological parameters with insulin resistance and β -cell dysfunction in nondiabetic subjects. *The Journal of Clinical Endocrinology & Metabolism* **94**, 3824–3832 (2009).
3. Barbieri, M. *et al.* New aspects of the insulin resistance syndrome: impact on haematological parameters. *Diabetologia* **44**, 1232–1237 (2001).
4. Emamian, M. *et al.* Association of hematocrit with blood pressure and hypertension. *Journal of clinical laboratory analysis* (2017).
5. Cirillo, M., Laurenzi, M., Trevisan, M. & Stamler, J. Hematocrit, blood pressure, and hypertension. The Gubbio Population Study. *Hypertension* **20**, 319–326 (1992).
6. Vozarova, B. *et al.* High white blood cell count is associated with a worsening of insulin sensitivity and predicts the development of type 2 diabetes. *Diabetes* **51**, 455–461 (2002).
7. Ohshita, K. *et al.* Elevated white blood cell count in subjects with impaired glucose tolerance. *Diabetes care* **27**, 491–496 (2004).
8. Tong, P. C. *et al.* White blood cell count is associated with macro-and microvascular complications in Chinese patients with type 2 diabetes. *Diabetes care* **27**, 216–222 (2004).
9. Shankar, A., Klein, B. E. & Klein, R. Relationship between white blood cell count and incident hypertension. *American journal of hypertension* **17**, 233–239 (2004).
10. Gillum, R. F. & Mussolino, M. E. White blood cell count and hypertension incidence. The NHANES I Epidemiologic Follow-up Study. *Journal of clinical epidemiology* **47**, 911–919 (1994).
11. Kim, D.-J. *et al.* The associations of total and differential white blood cell counts with obesity, hypertension, dyslipidemia and glucose intolerance in a Korean population. *Journal of Korean medical science* **23**, 193–198 (2008).

12. Ohnaka, K. *et al.* Inverse associations of serum bilirubin with high sensitivity C-reactive protein, glycated hemoglobin, and prevalence of type 2 diabetes in middle-aged and elderly Japanese men and women. *Diabetes research and clinical practice* **88**, 103–110 (2010).
13. Cheriya, P. *et al.* High total bilirubin as a protective factor for diabetes mellitus: an analysis of NHANES data from 1999–2006. *Journal of clinical medicine research* **2**, 201–206 (2010).
14. Fukui, M. *et al.* Relationship between serum bilirubin and albuminuria in patients with type 2 diabetes. *Kidney international* **74**, 1197–1201 (2008).
15. Papadakis, J. A., Ganotakis, E. S., Jagroop, I. A., Mikhailidis, D. P. & Winder, A. F. Effect of hypertension and its treatment on lipid, lipoprotein (a), fibrinogen, and bilirubin levels in patients referred for dyslipidemia. *American journal of hypertension* **12**, 673–681 (1999).
16. Giral, P. *et al.* Plasma bilirubin and gamma-glutamyltransferase activity are inversely related in dyslipidemic patients with metabolic syndrome: relevance to oxidative stress. *Atherosclerosis* **210**, 607–613 (2010).
17. Nakanishi, N., Suzuki, K. & Tatara, K. Serum γ -glutamyltransferase and risk of metabolic syndrome and type 2 diabetes in middle-aged Japanese men. *Diabetes care* **27**, 1427–1432 (2004).
18. Clark, J. M., Brancati, F. L., Diehl, A. M., *et al.* The prevalence and etiology of elevated aminotransferase levels in the United States. *The American journal of gastroenterology* **98**, 960–967 (2003).
19. Vozarova, B. *et al.* High alanine aminotransferase is associated with decreased hepatic insulin sensitivity and predicts the development of type 2 diabetes. *Diabetes* **51**, 1889–1895 (2002).
20. Sattar, N. *et al.* Elevated alanine aminotransferase predicts new-onset type 2 diabetes independently of classical risk factors, metabolic syndrome, and C-reactive protein in the west of Scotland coronary prevention study. *Diabetes* **53**, 2855–2860 (2004).
21. Fraser, A. *et al.* Alanine aminotransferase, γ -glutamyltransferase, and incident diabetes. *Diabetes care* **32**, 741–750 (2009).
22. Crawford, S. O. *et al.* Association of blood lactate with type 2 diabetes: the Atherosclerosis Risk in Communities Carotid MRI Study. *International journal of epidemiology* **39**, 1647–1655 (2010).
23. Maxwell, D. B., Fisher, E. A., Ross-Clunis 3rd, H. & Estep, H. L. Serum alkaline phosphatase in diabetes mellitus. *Journal of the American College of Nutrition* **5**, 55–59 (1986).
24. Lee, D.-H. *et al.* Gamma-glutamyltransferase and diabetes—a 4 year follow-up study. *Diabetologia* **46**, 359–364 (2003).
25. Lee, D. S. *et al.* Gamma glutamyl transferase and metabolic syndrome, cardiovascular disease, and mortality risk. *Arteriosclerosis, thrombosis, and vascular biology* **27**, 127–133 (2007).

26. Ikai, E., Honda, R. & Yamada, Y. Serum gamma-glutamyl transpeptidase level and blood pressure in nondrinkers: a possible pathogenetic role of fatty liver in obesity-related hypertension. *Journal of human hypertension* **8**, 95–100 (1994).
27. Shrestha, S., Gyawali, P., Shrestha, R., Poudel, B. & Sigdel, M. Serum urea and creatinine in diabetic and non-diabetic subjects. *Journal of Nepal Association for Medical Laboratory Sciences P* **11**, 12 (2008).
28. Taniguchi, Y. *et al.* Serum uric acid and the risk for hypertension and Type 2 diabetes in Japanese men: The Osaka Health Survey. *Journal of hypertension* **19**, 1209–1215 (2001).
29. Jossa, F. *et al.* Serum uric acid and hypertension: the Olivetti heart study. *Journal of human hypertension* **8**, 677–681 (1994).
30. Shulman, N. B. *et al.* Prognostic value of serum creatinine and effect of treatment of hypertension on renal function. Results from the hypertension detection and follow-up program. The Hypertension Detection and Follow-up Program Cooperative Group. *Hypertension* **13**, 180 (1989).
31. Coresh, J. *et al.* Prevalence of high blood pressure and elevated serum creatinine level in the United States: findings from the third National Health and Nutrition Examination Survey (1988–1994). *Archives of internal medicine* **161**, 1207–1216 (2001).
32. Mänttari, M., Tiula, E., Alikoski, T. & Manninen, V. Effects of hypertension and dyslipidemia on the decline in renal function. *Hypertension* **26**, 670–675 (1995).
33. Adrogué, H. J. & Madias, N. E. Sodium and potassium in the pathogenesis of hypertension. *New England Journal of Medicine* **356**, 1966–1978 (2007).
34. Whelton, P. K. *et al.* Effects of oral potassium on blood pressure: meta-analysis of randomized controlled clinical trials. *Jama* **277**, 1624–1632 (1997).
35. Pittas, A. G., Lau, J., Hu, F. B. & Dawson-Hughes, B. The role of vitamin D and calcium in type 2 diabetes. A systematic review and meta-analysis. *The Journal of Clinical Endocrinology & Metabolism* **92**, 2017–2029 (2007).
36. Saltevo, J. *et al.* Serum calcium level is associated with metabolic syndrome in the general population: FIN-D2D study. *European journal of endocrinology* **165**, 429–434 (2011).
37. Pradhan, A. D., Manson, J. E., Rifai, N., Buring, J. E. & Ridker, P. M. C-reactive protein, interleukin 6, and risk of developing type 2 diabetes mellitus. *Jama* **286**, 327–334 (2001).
38. Yudkin, J. S., Stehouwer, C., Emeis, J. & Coppack, S. C-reactive protein in healthy subjects: associations with obesity, insulin resistance, and endothelial dysfunction. *Arteriosclerosis, thrombosis, and vascular biology* **19**, 972–978 (1999).
39. Freeman, D. J. *et al.* C-reactive protein is an independent predictor of risk for the development of diabetes in the West of Scotland Coronary Prevention Study. *Diabetes* **51**, 1596–1600 (2002).
40. Sesso, H. D. *et al.* C-reactive protein and the risk of developing hypertension. *Jama* **290**, 2945–2951 (2003).

41. Bautista, L. E. *et al.* Is C-reactive protein an independent risk factor for essential hypertension? *Journal of hypertension* **19**, 857–861 (2001).
42. Bonaa, K. & Thelle, D. Association between blood pressure and serum lipids in a population: the Tromsø Study. *Circulation* **83**, 1305–1314 (1991).
43. Modan, M. *et al.* Hyperinsulinemia. A link between hypertension obesity and glucose intolerance. *Journal of clinical investigation* **75**, 809 (1985).
44. Chiu, K. C. *et al.* Insulin sensitivity is inversely correlated with plasma intact parathyroid hormone level. *Metabolism* **49**, 1501–1505 (2000).
45. Snijder, M. *et al.* Vitamin D status and parathyroid hormone levels in relation to blood pressure: a population-based study in older men and women. *Journal of internal medicine* **261**, 558–565 (2007).
46. Jorde, R., Svartberg, J. & Sundsfjord, J. Serum parathyroid hormone as a predictor of increase in systolic blood pressure in men. *Journal of hypertension* **23**, 1639–1644 (2005).
47. Ford, E. S. & Cogswell, M. E. Diabetes and serum ferritin concentration among US adults. *Diabetes care* **22**, 1978–1983 (1999).
48. Forouhi, N. *et al.* Elevated serum ferritin levels predict new-onset type 2 diabetes: results from the EPIC-Norfolk prospective study. *Diabetologia* **50**, 949–956 (2007).
49. Hämäläinen, P., Saltevo, J., Kautiainen, H., Mäntyselkä, P. & Vanhala, M. Erythropoietin, ferritin, haptoglobin, hemoglobin and transferrin receptor in metabolic syndrome: a case control study. *Cardiovascular diabetology* **11**, 116 (2012).
50. Piperno, A. *et al.* Increased serum ferritin is common in men with essential hypertension. *Journal of hypertension* **20**, 1513–1518 (2002).
51. Harrell, F. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis* (Springer, 2015).
52. Steyerberg, E. *Clinical prediction models: a practical approach to development, validation, and updating* (Springer Science & Business Media, 2008).
53. Nelder, J. A. & Wedderburn, R. W. M. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)* **135**, 370–384 (1972).
54. Hastie, T., Tibshirani, R. & Friedman, J. *The elements of statistical learning: data mining, inference and prediction* 2nd ed. <http://www-stat.stanford.edu/~tibs/ElemStatLearn/> (Springer, 2009).
55. Walker, S. H. & Duncan, D. B. Estimation of the probability of an event as a function of several independent variables. *Biometrika* **54**, 167–179 (1967).
56. McCullagh, P. Regression models for ordinal data. *Journal of the royal statistical society. Series B (Methodological)*, 109–142 (1980).
57. Armstrong, B. & Sloan, M. Ordinal regression models for epidemiologic data. *American Journal of Epidemiology* **129**, 191–204 (1989).
58. Berridge, D. M. & Whitehead, J. Analysis of failure time data with ordinal categories of response. *Statistics in medicine* **10**, 1703–1710 (1991).

59. Malley, J. D., Malley, K. G. & Pajevic, S. *Statistical learning for biomedical data* (Cambridge University Press, 2011).