TAVR

Adverse outcomes following Transcatheter Aortic Valve Replacement

Luca Carmisciano

19 May, 2023

Contents

1 Introduction

Trans-catheter aortic valve replacement (TAVR) is the treatment of choice for severe aortic stenosis (AS) at high risk for conventional surgery, and it is increasingly being performed in patients at intermediate and low conventional surgery risk. However, several TAVR-related complications may impact short- and long-term outcomes. Careful evaluation of these risks could play a pivotal role in TAVR scenario, potentially supporting clinical decision making and helping in the early identification of patients who could benefit the most from higher peri- and post-procedural surveillance.

To date, the pre-operative evaluation of risk of adverse outcome after AS treatment recommended by the European Society of Cardiology (ESC) is based on EUROSCORE II and STS score, which have been developed in the last decade on cohorts who undergone conventional surgery replacement of the aortic valve. The last ESC guidelines did not recommend any tool to evaluate the baseline risk of adverse outcome following TAVR. Nevertheless, different predictive tools have been recently developed to estimate TAVR-related risks of 30-day bleeding complications, in-hospital mortality, 30-day and 1-year death, and also the probability to undergone a pacemaker implantation. Most of these score have been derived from population of few patients or showed a low performance (Area Under the Curve often less than 0.65), therefore the overall accuracy of these scores and their generalizability to real-world cohorts remain still modest, representing an unmet need for individualized patient management strategies.

1.1 Aims

The aim of this project is to explore the feasibility of using baseline patients' characteristics to predict all-cause death or major adverse cardiovascular events (MACE) occurring after TAVR procedure.

2 Material

2.1 Data

We used data from the TRITAVI registry, a large, multi-center, collection of real-world observations from routinary clinical practice of TAVR from 2007 to 2022. Data are structured in the "tidy" format, where each row is an independent observation unite and represent a subject, each column is a feature/characteristic of the medical history of that subject. Twenty hospitals from Italy, Spain, Finland, Poland and England enrolled 11264 subjects. Of these we consider as an unreliable source of information and therefore excluded from the study 3083 subjects with a follow-up shorter than 1 year but free at the last follow-up from any relevant adverse event.

```
# load the data
load(project.path)
dim(imp$data)
```

```
## [1] 8181   59
```

These 59 columns can be grouped into three timing/meaning based groups: baseline characteristics, short term outcomes and long term outcomes.

Comment: rows are expected to be independent, however, as it is often the cases in multi-center data collection, there could be an hidden structure represented by the single point of care or the geographical region (es. more severe/acute/complex cases reaching the larger/notrorous/specialized hospital). The removal of short followed-up patient could introduce a selection bias (by excluding more subject likely to be 1-year outcome free than not), however in this work we focused on predictive performance rather than interpreting the effect size of each predictor.

2.1.1 Baseline characteristics

Baseline is defined as the time at the TAVR procedure is planned. Prior to the intervention only this set of features is available and will can be used to predict the occurrence of outcomes. STS and Euroscore have a special meaning becouse are directly estimating the risks of the TAVR alternative approach so we decide to exclude them from the Xs and consider them lately, separatly and independently.

Table 1: Baseline characteristics

	Encoding	i
Center identifier	center	1
Country of enrollment	country	2
Age	Age	3
Male sex	Sex_M	4
Weigth (kg)	Weigth_kg	5
Body Mass Index (kg/m2)	bmi	6
Baseline hystory of diabetes	Diabetes	7
Baseline hystory of dyslipidemia	Dyslipidemia	8
Baseline hystory of hypertension	Hypertension	9
Baseline hystory of smoking	Smoking	10
Baseline hystory of active_Cancer	Active_Cancer	11
Baseline hystory of liver disease	Liver_disease	12
Baseline hystory of dialysis	Base_Dialysis	13
Chronic obstructive pulmonary disease	COPD	14
Peripheral artery disease	PAD	15
New York Heart Association Classification	NYHA_class	16
Previous coronary artery disease	CAD	17
Previous myocardial infarction	Prior_MI	18
Previous percutaneous coronary intervention	Prior_PCI	19
Previous coronary artery bypass graft surgery	Prior_CABG	20
Previous cerebrovascular accident	Prior_CVA	21
Baseline farmacological treatment with Aspirin	Base_Aspirin	22
Baseline farmacological treatment with P2Y12i	Base_P2Y12i	23
Baseline farmacological treatment with DAPT	Base_DAPT	24
Baseline farmacological treatment with VKA	Base_VKA	25
Baseline farmacological treatment with NOAC	Base_NOAC	26
Predicted risk score of in hospital stay following classical surgery (STS)	STS	27
Predicted risk score of 1-y mortality following classical surgery (EuroScore)	EuroScore_2	28
Permanent pacemaker implantation	Permanent_PM	29
Porcelain aorta	Porcelain_Aorta	30
Atrial fibrillation	AF	31
Baseline ejection fraction	Base_EF	32
Baseline ejection fraction group	Base_EF_class	33
Baseline mean valvular pressure gradient	Base_Mean_Gradient	34
Baseline cratinin level (mg/dl)	Base_Creat	35
Platelet count (log scale)	Plts	36
Baseline haemoglobin level (g/dl)	Baseline_Hb	37
Year of TAVI procedure	y_of_procedure	38

2.1.2 Short term outcomes

Short term outcomes are defined as event occurring during TAVR procedure or during the hospital stay as defined by the VARC3 standards. Prior to the intervention this set of features is not available and therefore should not be directly used as input for long term outcome prediction.

Table 2: Short term outcomes

	Encoding	Type	i
Presence of Acute Kidey Injury as defined by VARC3 criteria	AKIyesno	Binary	1
Presence of severe Acute Kidey Injury (grade 2 or 3) as defined by VARC3 criteria	AKI23	Binary	2
Increase of creatinin from baseline VARC3 criteria (mg/dl)	Creat_increase	Continuous (laplacian distribution)	3
Require dialysis after TAVI	PostProc_Dialysis	Binary	4
Decrease of hemoglobin from baseline VARC3 criteria (g/dl)	Hb_drop	Continuous (softly left skewed)	5
Require transfusion after TAVI	Transfusion	Binary	6
Femoral (0) or other (1) approach used during TAVI	Approach	Binary	7
Type of implanted valve	Valve_type	Categorical unordered 3-levels	8
Peri-procedural migration of the valve	Valve_Migrat	Binary	9
Echographical amount of paravalvular leak at discarge	Echo_discarge_PVL	Categorical ordered 5-levels	10
Amount of contrast medium used during TAVI	Contrast_Medium	Continuous (softly right skewed)	11
Post-procedural mean valvular pressure gradient	Final_Mean_Gradient	Continuous (strongly right skewed)	12
Type of access closure percutanous (1) or surgical (2) (during TAVI)	Access_Closure_1perc_2surg	Binary	13
Occurrence of major vascular complication (VARC3 criteria)	Major_vasc_compl	Binary	14
Occurrence of minor vascular complication (VARC3 criteria)	Minor_vasc_compl	Binary	15
Presence and type of bleeding	bleed	Categorical ordered 4-levels	16
Length of hospital stay	Hp_stay_dd	Continuous (strongly right skewed)	17
Type of farmacological treatment set after TAVI	Post_therapy	Categorical unordered 3-levels	18

2.1.3 Long term outcomes

Long term outcomes are time dependent events occurring after TAVR procedure. Prior to the intervention this set of features is not available. 1-Year MACE prediction is the goal of this project.

Table 3: Long term outcomes

	Encoding	Type
Occurrence of any-cause death or any-stroke or myocardial infarction within 1 year from TAVI	outcome	Binary
Occurrence of non cardio vascular related death within 1 year from TAVI	frail	Binary

2.2 Train-test split

We randomly selected 25% of the 8181 rows to be used for testing purpose only.

```
db <- mice::complete(imp, 1)
test  <- db[test.index,]
train <- db[-test.index,]
train.x <- train[, xs]
```

2.3 Missing values handling and preprocess

We used multiple imputation with chained equations based on predictive mean matching and logistic or polytomous regressions to model the distribution of each variable with missing data. Rows included in the test set were not used to train imputation models but were imputed in the same fashion of the train set to avoid leakage of information between test and train sets.

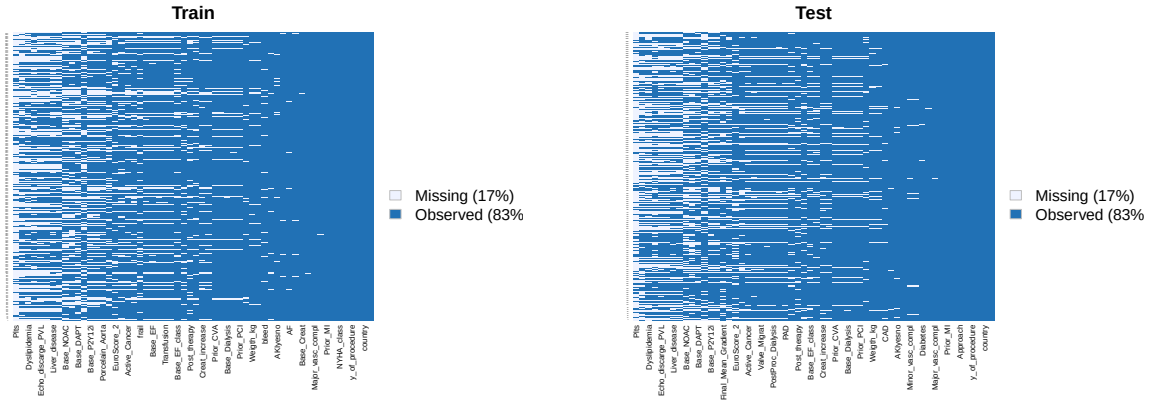


Figure 1: Missing

2.4 Profile

Characteristics of the sampled population are reported in the following tables.

Table 4: Descriptive statistics of baseline characteristics

	Overall
n	8181
Age (mean (SD))	81.10 (6.44)
Male sex = 1 (%)	3887 (47.5)
Weigth (kg) (mean (SD))	71.68 (14.76)
Body Mass Index (kg/m2) (mean (SD))	26.75 (4.90)
Baseline hystory of diabetes = 1 (%)	2360 (28.8)
Baseline hystory of dyslipidemia = 1 (%)	4430 (54.1)
Baseline hystory of hypertension = 1 (%)	6852 (83.8)
Baseline hystory of smoking (%)	
Never	6409 (78.3)
Current	627 (7.7)
Past	1145 (14.0)
Baseline hystory of active_Cancer = 1 (%)	534 (6.5)
Baseline hystory of liver disease = 1 (%)	342 (4.2)
Baseline hystory of dialysis = 1 (%)	165 (2.0)
Chronic obstructive pulmonary disease = 1 (%)	2311 (28.2)
Peripheral artery disease = 1 (%)	1661 (20.3)
New York Heart Association Classification (%)	
1	148 (1.8)
2	1832 (22.4)
3	5500 (67.2)
4	701 (8.6)
Previous coronary artery disease = 1 (%)	2846 (34.8)
Previous myocardial infarction = 1 (%)	1307 (16.0)
Previous percutaneous coronary intervention = 1 (%)	2008 (24.5)
Previous coronary artery bypass graft surgery = 1 (%)	1029 (12.6)
Previous cerebrovascular accident = 1 (%)	1592 (19.5)
Baseline farmacological treatment with Aspirin = 1 (%)	4679 (57.2)
Baseline farmacological treatment with P2Y12i = 1 (%)	1821 (22.3)
Baseline farmacological treatment with DAPT = 1 (%)	1519 (18.6)
Baseline farmacological treatment with VKA = 1 (%)	1654 (20.2)
Baseline farmacological treatment with NOAC = 1 (%)	987 (12.1)
Permanent pacemaker implantation = 1 (%)	879 (10.7)
Porcelain aorta = 1 (%)	434 (5.3)
Atrial fibrillation = 1 (%)	2387 (29.2)
Baseline ejection fraction (mean (SD))	54.20 (11.52)
Baseline ejection fraction group (mean (SD))	1.29 (0.55)
Baseline mean valvular pressure gradient (mean (SD))	47.08 (15.26)
Baseline cratinin level (mg/dl) (mean (SD))	1.24 (0.87)
Platelet count (log scale) (mean (SD))	12.19 (0.37)
Baseline haemoglobin level (g/dl) (mean (SD))	12.12 (1.69)
Year of TAVI procedure (%)	
2007 to 2009	288 (3.5)

Table 4: Descriptive statistics of baseline characteristics *(continued)*

	Overall
2010	287 (3.5)
2011	322 (3.9)
2012	428 (5.2)
2013	484 (5.9)
2014	594 (7.3)
2015	961 (11.7)
2016	1321 (16.1)
2017	1340 (16.4)
2018	657 (8.0)
2019	841 (10.3)
2020 to 2022 (COVID)	658 (8.0)

Table 5: Descriptive statistics of outcomes characteristics

	Overall
n	8181
Presence of Acute Kidey Injury as defined by VARC3 criteria = 1 (%)	907 (11.1)
Presence of severe Acute Kidey Injury (grade 2 or 3) as defined by VARC3 criteria = Yes (%)	679 (8.3)
Increase of creatinin from baseline VARC3 criteria (mg/dl) (mean (SD))	0.17 (0.63)
Require dialysis after TAVI = 1 (%)	215 (2.6)
Decrease of hemoglobin from baseline VARC3 criteria (g/dl) (mean (SD))	-2.30 (1.32)
Require transfusion after TAVI = 1 (%)	1514 (18.5)
Femoral (0) or other (1) approach used during TAVI (%)	
FEM	7441 (91.0)
APIC	491 (6.0)
OTHER	249 (3.0)
Type of implanted valve (%)	
Type 1	3510 (42.9)
Type 2	3589 (43.9)
Type 3	1082 (13.2)
Peri-procedural migration of the valve = 1 (%)	169 (2.1)
Echographical amount of paravalvular leak at discarge (%)	
0	4758 (58.2)
1	2766 (33.8)
2	439 (5.4)
3	114 (1.4)
4	104 (1.3)
Amount of contrast medium used during TAVI (mean (SD))	198.47 (97.52)
Post-procedural mean valvular pressure gradient (mean (SD))	8.89 (5.23)
Type of access closure percutanous (1) or surgical (2) (during TAVI) = 2 (%)	1376 (16.8)
Occurrence of major vascular complication (VARC3 criteria) = 1 (%)	619 (7.6)
Occurrence of minor vascular complication (VARC3 criteria) = 1 (%)	654 (8.0)
Presence and type of bleeding (%)	
none	4408 (53.9)
minor	2071 (25.3)
major	84 (1.0)
lt	1618 (19.8)
Length of hospital stay (mean (SD))	8.21 (8.22)
Type of farmacological treatment set after TAVI (%)	
SAPT	2943 (36.0)
DAPT	2609 (31.9)
OAC	2629 (32.1)
Occurrence of any-cause death or any-stroke or myocardial infarction within 1 year from TAVI = TRUE (%)	2951 (36.1)
Occurrence of non cardio vascular related death within 1 year from TAVI = TRUE (%)	1855 (22.7)

3 Workflow

- **Step 1. Standard approach:** Learn from training set to detect subjects that will have MACE in 1 year using a binary classifier (logistic regression).
- **Step 2. Un-cluster:** Identify pattern between natural grouping structure of the training set to improve the ability to generalize predictions on different grouping structures.
- **Step 3. Re-balance:** Sample the training set to obtain a new set of virtual observations where the classes of interest are equally represented.
- **Step 4. Sources maximization:** Model the probability of secondary outcomes occurrence in the training set to improve the performance to detect MACE.
- **Step 5. Combine predictions:** Learn from a de-clustered and re-balanced training set to detect subjects that will have MACE in 1 year using a classifier trained on baseline covariates and secondary outcomes risk prediction.

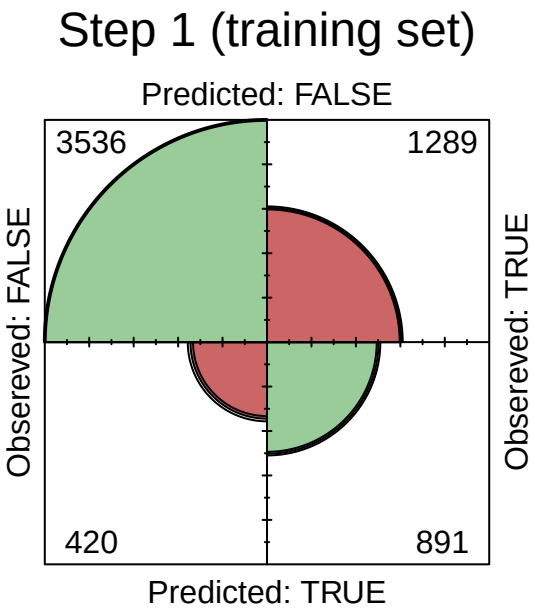
3.1 Step 1. Standard approach

To learn to classify training set subjects that will have MACE in 1 year we fit a multivariable logistic regression model using the occurrence of 1-year MACE as dependent variable and all the other baseline characteristics in the training set as independent variables.

```
step.1.formula <- paste0("outcome~", paste(xs, collapse = "+"))

step.1.model <- glm(
  formula = step.1.formula,
  data    = train,
  family  = binomial
)

step.1.pred <- predict(step.1.model, type = "res") > 0.5
```

comment: although we are focused on predictions rather than effect sizes, we can spot some issue in step 1 model. First, country is not used because center are unique across countries so they borrow and fragment the country effects. Second, I don't want to make different prediction based on centers or country at all, I want the model to be as general as possible. Third, a lot of predictors are significantly associated to the outcome but some pairs of them are associated and, beside betas, their simultaneous inclusion may damage model stability (the more clear pair is body weight and BMI)

We then evaluate the goodness of training set fit.



comment: 72% of accuracy is not bad. However, to put this in context, and because of outcome unbalance in training set, just by knowing the prevalence of MACE, I could have a 64% accuracy (NIR). Also, 1289 subject predicted to be fine within 1 year from TAVR actually had MACE. This is one of the most important goal of the prediction tool and a Specificity of 41% is way to low for this context of clinical practice.

3.2 Step 2. Un-cluster

Country and center was not expected to be associated to MACE outcome. One hypothesis for this finding could be that they are absorbing the effect of some hidden and potentially unmeasured and clinically relevant factor (such as resources invested in TAVR setting or simply different level of center specialization). We can consider the study design as made of clustered samples. Aim of this step is to remove clustering hoping to improve generalizability.

comment: clustering techniques often appear among researchers' aims. However the strong association between the MACE outcome and the hospital (visible in step 1 model coefficients, not shown here) suggest our data are somehow already clustered. We observed a similar behavior for hospital's country when center is removed from step 1 model.

One possible approach could be to stratify the data, which is conceptually similar to train a specific prediction models for each center and then merge the results. Another strategy could be to give the data a structure and account for the within cluster dependency of the rows using something like hierarchical modelling and including a random effect term for center or country or both nested. Although possible, this modelling strategy may be critical in terms of stability because some centers are small. Also, we then want to generalize the model and to make the MACE risk predictions for centers other then the ones used in train. Another possible approach could simply be to balance centers or country or both (but one at a time since they are nested and cannot be balanced together) and then simply ignore those variables. However, in the hypothesis for this clusters to absorb the effect of some hidden relevant factors balancing can be a waste of information.

Lastly, the option we choose was to estimate which cluster a subject is more likely be in based on its baseline characteristics, and then used such likelihood as a potential MACE predictor. Since this choice includes fitting a multiclass classifier, the stability problem due to centers with a small number of subject remain.

Table 6: Number of subjects in each center

Country	Center	Test	Train	i
Country 1	Center 1	65	23	1
	Center 2	86	22	2
	Center 3	364	104	3
	Center 4	297	111	4
	Center 5	35	17	5
	Center 6	69	21	6
	Center 7	373	113	7
	Center 8	297	93	8
	Center 9	1247	439	9
	Center 10	278	75	10
	Center 11	15	12	11
	Center 12	515	157	12
Country 2	Center 1	168	70	13
Country 3	Center 1	276	76	14
	Center 2	260	109	15
	Center 3	423	149	16
	Center 4	203	73	17
	Center 5	424	137	18
Country 4	Center 1	168	59	19
Country 5	Center 1	573	185	20

comment: in this random split a center has as low as 12 subjects but train and test sets have a non null probability of having different centers (so a minimum strata size of zero) just by chance.

As a possible remediation to this problem we might check if centers could further be clustered. For this specific task the observation units are centers so we can deal with categorical predictors just by using class prevalence to obtain a completely numerical dataset of average/percentage by center.

```
settings.db <- train %>%
  select(-ID) %>%
  group_by(center, country) %>%
  summarise_all(summariser) %>%
  ungroup() %>%
  left_join(summarise(group_by(train, center, country), n = n()))
```

Before clustering subjects were made more comparable among features by scaling each column to a z-score with center located of the mean/percentage located at zero and a size indicating the number of standard deviations from the center.

```
settings.db.scaled <- apply(settings.db[, -c(1, 2)], 2, scale)
```

We performed a grid search approach to select an optimal number of cluster. We computed distance matrices using five different methods: complete, average, single, and the two versions of the Ward distance. We then applied hierarchical clustering or K-mean clustering. Agglomeration of hierarchical clustering used either euclidean, maximum or the manhattan between-cluster distance.

We selected a clustering strategy based on an a custom metric, designed to find the 2 or more large and similarly sized clusters, with the following formula:

$$metric = \frac{\sum(N_{max} - N_k)}{K}$$

Where N_{max} represented the size of the larger cluster, and small k is the cluster index ranging from 2 to big K possible clusters. The lower the metric, the better the score.

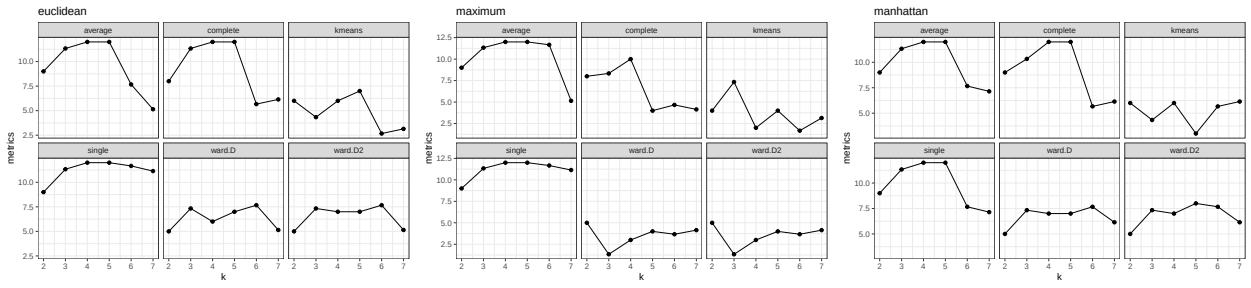


Figure 2: Grid search of optimal cluster number and method

distance	k	method	metrics
maximum	3	ward.D	1.333333

comment: hierarchical clustering based on the maximum between cluster distance and ward between subject distance cutting at 3 clusters produced the largest and similarly sized clusters.

```
clustering <- hclust(dist(settings.db.scaled, "maximum"), method = "ward.D")
settings.db$cluster <- cutree(clustering, k = 3)
```

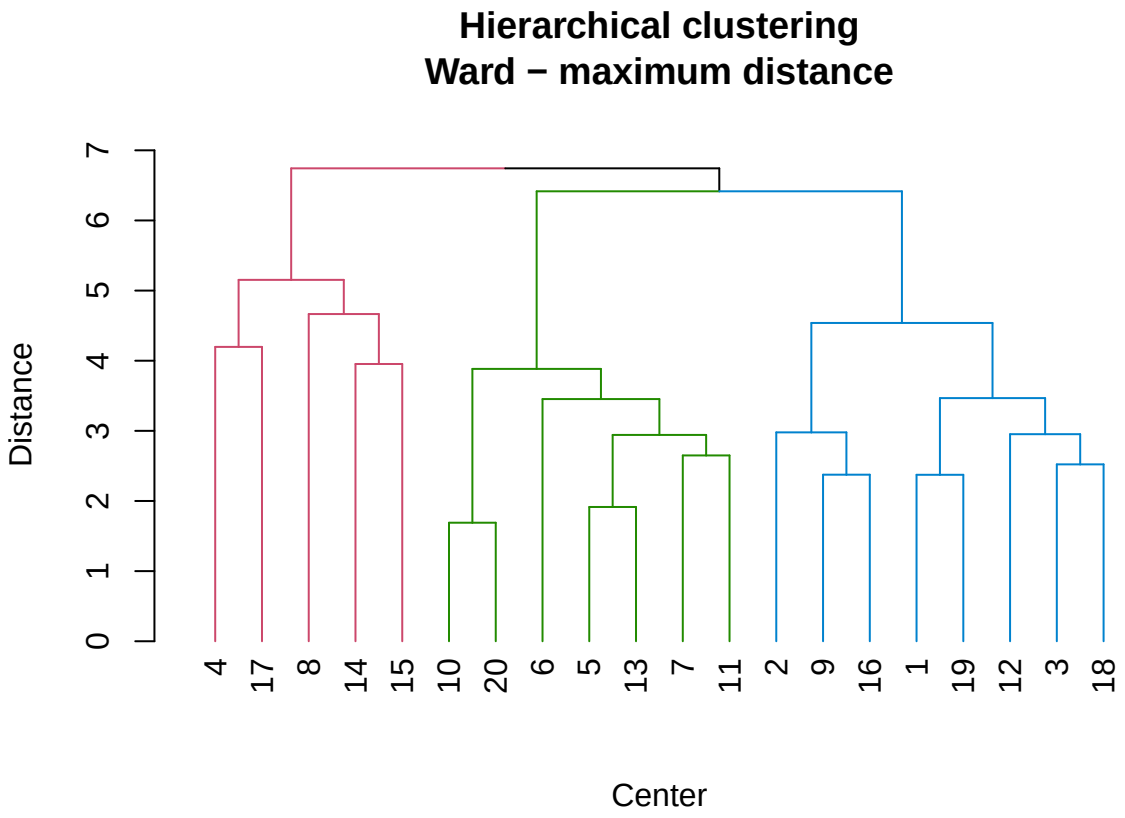


Table 7: Descriptive statistics by center cluster.

	Cluster				
	1	2	3	p	test
n	8	5	7		
outcome (mean (SD))	0.71 (0.13)	0.60 (0.28)	0.65 (0.08)	0.527	
frail (mean (SD))	0.87 (0.08)	0.81 (0.11)	0.74 (0.05)	0.018	
Age (mean (SD))	81.52 (1.28)	80.12 (1.30)	80.88 (0.88)	0.134	
Sex_M (mean (SD))	0.54 (0.08)	0.45 (0.08)	0.52 (0.05)	0.096	
Weigth_kg (mean (SD))	69.99 (1.82)	72.86 (3.61)	72.47 (3.17)	0.150	
bmi (mean (SD))	26.38 (0.81)	27.00 (1.28)	26.70 (0.76)	0.503	
Diabetes (mean (SD))	0.71 (0.05)	0.66 (0.08)	0.74 (0.03)	0.060	
Dyslipidemia (mean (SD))	0.49 (0.12)	0.35 (0.19)	0.49 (0.06)	0.136	
Hypertension (mean (SD))	0.18 (0.08)	0.37 (0.36)	0.16 (0.02)	0.143	
Smoking (mean (SD))	0.81 (0.10)	0.80 (0.17)	0.75 (0.06)	0.512	
Active_Cancer (mean (SD))	0.91 (0.05)	0.94 (0.03)	0.96 (0.03)	0.164	
Liver_disease (mean (SD))	0.97 (0.02)	0.97 (0.02)	0.93 (0.03)	0.022	
Base_Dialysis (mean (SD))	0.98 (0.02)	0.98 (0.01)	0.99 (0.01)	0.316	
COPD (mean (SD))	0.77 (0.04)	0.53 (0.33)	0.78 (0.03)	0.032	
PAD (mean (SD))	0.80 (0.09)	0.67 (0.16)	0.78 (0.08)	0.104	
NYHA_class (mean (SD))	2.78 (0.14)	2.82 (0.25)	2.87 (0.12)	0.555	
CAD (mean (SD))	0.59 (0.13)	0.50 (0.25)	0.66 (0.13)	0.276	
Prior_MI (mean (SD))	0.83 (0.07)	0.78 (0.16)	0.84 (0.04)	0.533	
Prior_PCI (mean (SD))	0.71 (0.06)	0.68 (0.16)	0.74 (0.07)	0.571	

Table 7: Descriptive statistics by center cluster. *(continued)*

	1	2	3	p	test
Prior_CABG (mean (SD))	0.89 (0.05)	0.87 (0.08)	0.83 (0.04)	0.118	
Prior_CVA (mean (SD))	0.91 (0.03)	0.93 (0.04)	0.84 (0.03)	0.001	
Base_Aspirin (mean (SD))	0.41 (0.07)	0.46 (0.05)	0.43 (0.08)	0.439	
Base_P2Y12i (mean (SD))	0.69 (0.09)	0.78 (0.07)	0.82 (0.07)	0.023	
Base_DAPT (mean (SD))	0.76 (0.09)	0.80 (0.07)	0.87 (0.07)	0.055	
Base_VKA (mean (SD))	0.82 (0.07)	0.91 (0.03)	0.70 (0.11)	0.001	
Base_NOAC (mean (SD))	0.88 (0.07)	0.82 (0.07)	0.96 (0.03)	0.003	
Permanent_PM (mean (SD))	0.89 (0.03)	0.87 (0.05)	0.90 (0.03)	0.397	
Porcelain_Aorta (mean (SD))	0.95 (0.03)	0.80 (0.31)	0.94 (0.03)	0.208	
AF (mean (SD))	0.73 (0.05)	0.72 (0.08)	0.63 (0.09)	0.054	
Base_EF (mean (SD))	54.19 (3.26)	52.32 (4.32)	54.55 (2.01)	0.471	
Base_EF_class (mean (SD))	1.29 (0.06)	1.35 (0.20)	1.31 (0.06)	0.638	
Base_Mean_Gradient (mean (SD))	47.04 (3.03)	45.70 (4.11)	48.22 (3.65)	0.487	
Base_Creat (mean (SD))	1.28 (0.13)	1.21 (0.09)	1.17 (0.08)	0.170	
Plts (mean (SD))	12.16 (0.05)	12.19 (0.06)	12.17 (0.03)	0.646	
Baseline_Hb (mean (SD))	11.99 (0.13)	12.03 (0.65)	12.41 (0.38)	0.124	
y_of_procedure (mean (SD))	7.53 (1.67)	10.07 (1.25)	6.70 (0.33)	0.001	

comment: clustering seem to capture a mix of different pharmacological treatments, different clinical history and frailty.

To use this clusters at intervention time we need to estimate the setting (country and cluster) from baseline data. We attempted to predict the setting fitting two neural networks-based multinomial log-linear model (one for country and one for centers cluster).

Table 8: Country prediction based on baseline covariates.

	Country 1	Country 2	Country 3	Country 4	Country 5	Sum
Predicted country 1	3188	128	687	91	12	4106
Predicted country 2	12	18	6	0	0	36
Predicted country 3	388	21	882	10	5	1306
Predicted country 4	22	0	2	65	2	91
Predicted country 5	31	1	9	2	554	597
Sum	3641	168	1586	168	573	6136

```
mean(predict(country.model) == train.x$country)
```

```
## [1] 0.7671121
```

We can estimate from baseline characteristic the country-like hidden factor with 77% accuracy on the training set.

```
mean(predict(center.model) == train.x$cluster)
```

```
## [1] 0.6313559
```

Table 9: Cluster prediction based on baseline covariates.

	Cluster 1	Cluster 2	Cluster 3	Sum
Predicted cluster 1	817	326	367	1510
Predicted cluster 2	382	1306	138	1826
Predicted cluster 3	643	406	1751	2800
Sum	1842	2038	2256	6136

We can estimate from baseline characteristic the center-like hidden factor with 63% accuracy on the training set.

We can now summarise the hidden setting factor into one or a few dimensions. We perfomed a principal component analysis to obtain these dimension of maximal explainable variance of the setting prediction (each multiclass classificater return information on estimated probability of a subject to appartain to each possible class)

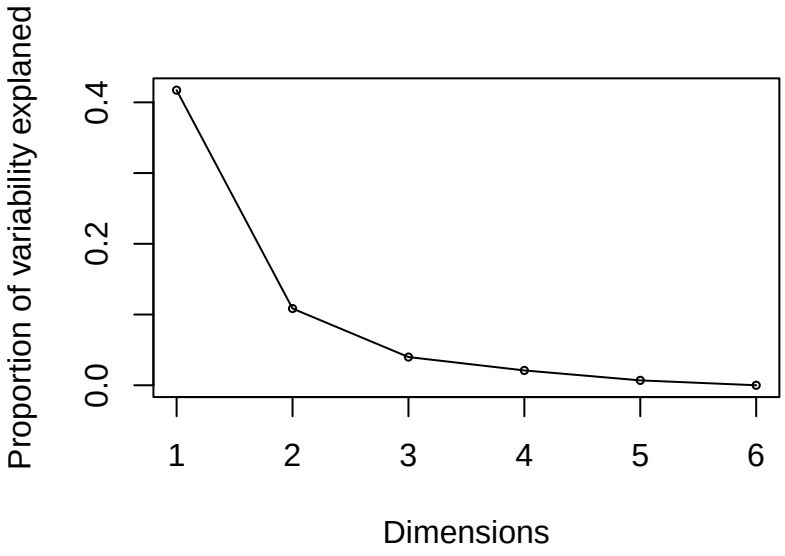


Figure 3: Visual inspection for elbow

```
cntry.pred <- predict(country.model, newdata = train.x, type = "probs")[,-1]
cnter.pred <- predict(center.model, newdata = train.x, type = "probs")[,-1]
pca <- prcomp(cbind(cntry.pred, cnter.pred), scale = F, center = F)
summary(pca)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6
## Standard deviation     0.6085 0.4429 0.20876 0.10970 0.09448 0.06631
## Proportion of Variance 0.5828 0.3087 0.06858 0.01894 0.01405 0.00692
## Cumulative Proportion  0.5828 0.8915 0.96009 0.97903 0.99308 1.00000
```

The first 2 dimensions explain almost 90% of the setting latent variable and can be computed from baseline characteristics. We visually selected 3 as the number of components to keep from PCA using screeplot.

3.3 Step 3. Re-balance

Prevalence of outcome in the training set not optimally balanced (35.5%). Better training could be achieved with balancing.

Table 10: MACE prevalence in training set.

	Overall
n	6136
outcome = TRUE (%)	2180 (35.5)

We had two main options to balance data: (1) to sub-sample the most represented outcome group or (2) to over-sample the less represented outcome group. We chosed to under-sample.

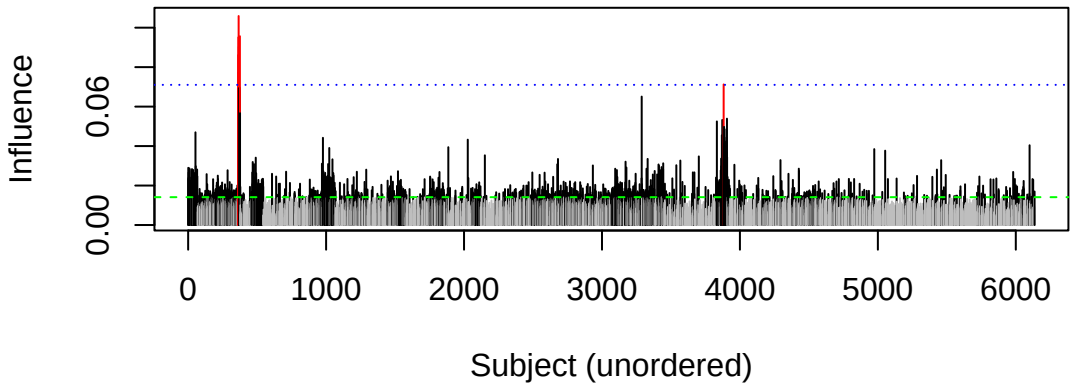
comment: over sample is an harder choice since sampling from a limited option may drop the natural variability of the less represented outcome group. However, we could simply delete a random percentage of our over sampled data and use the already trained imputation model to generate noisier synthetic data.

To construct a weight for sampling we could use the inverse of the relative group size. So for example in a 10 element grouping scenario with a first group of the first 3 and a second group of the last 7, each element of the two groups would have get a weight of 1 over 3/10 and 1 over 7/10 respectively. More then one weights can be used simultaneously by stratifying groups using two or more variables or by combining sampling probabilities.

```
group.size <- prop.table(table(train$outcome))
1 / group.size
```

```
##
##    FALSE     TRUE
## 1.551062 2.814679
```

When sampling we can assign an higher weight (or equivalently an higher sampling probability) based on how important that single data point appeared to be for the classification the training of the model developed in step 1. We can therefore use also importance measures, such as leverage to balance.



Blue and green horizontal lines represented the 99.9th and 80th leverage percentile respectively. We down-weighted extreme outliers (1‰ highest silenced), under-weighted subject below the 80th leverage percentile and over-weighted subject above the 80th leverage percentile as they both were ~50%.

```r
leverage.q <- quantile(train$leverage, probs = c(0, 0.8, 0.999, 1))

train$leverage.group <- cut(
  x = train$leverage, breaks = leverage.q, include.lowest = TRUE)

1 / prop.table(table(train$leverage.group))
```

```
##
## [0.0014,0.0141]  (0.0141,0.071]   (0.071,0.106]
##        1.249949        5.029508      876.571429
```

Combine the weights and silence extreme outliers.

```r
resampling.weights <- train %>%
  group_by(outcome, leverage.group) %>%
  transmute(ip = 1 / (n()/nrow(train))) %>%
  mutate(ip = ifelse(leverage.group == last(levels(leverage.group)), 0, ip)) %>%
  {.[,"ip", drop = TRUE]}

set.seed(1234)
resampling.ids <- sample(
  1:nrow(train),
  size = 2*min(table(train$outcome)),
  replace = T,
  prob = resampling.weights)

resw.train <- train[resampling.ids,]
```

Check if balance worked

```
##                  MACE
## leverage          FALSE TRUE  Sum
##   [0.0014,0.0141]  1076 1122 2198
##   (0.0141,0.071]   1073 1089 2162
##   (0.071,0.106]       0    0    0
##   Sum              2149 2211 4360
```

3.4 Step 4. Source maximization

Aim of this step is to borrow predictive power from other source of outcome. We hypnotized as possible source of information the non cardiovascular related deaths (which we refers to as frailty) and how the procedure ran (short term outcomes such as bleeding during intervention or length of ospital stay).

comment: death cause adjudication by physician and short term outcomes have been collected but are mainly ignored because are not available before the intervention. These are a way too precious resource to be wasted.

3.4.1 Frailty

Frailty is a cross-medical domain concept of death proximity, which is independent from the TAVR risk by design. In this data we have a frailty outcome which exclude TAVR related deaths. Age and previous diseases are expected to be the main drivers of frailty. We can model frailty to then adjust MACE prediction (adding it as interaction term with other predictors) allowing different MACE contribution per risk factor based on (predicted) frailty status.

To estimate frailty we fit a logistic regression using 1-year non TAVR related death as outcome and all baseline variables as predictors. At first, we included a penalization term in the residual computation using an L1 penalty equal to the absolute value of the magnitude of the coefficients.

```
frail.db <- resw.train
frail.db <- cbind(frail.db, estimate.setting(data = frail.db))

x_var <- model.matrix(
  frail ~ .,
  frail.db[, c(xs[-c(1:2)], "frail", "Setting_1", "Setting_2", "Setting_3")])
y_var <- frail.db$frail == "TRUE"

lasso <- glmnet::glmnet(
  x      = x_var,
  y      = y_var,
  family = "binomial",
  alpha  = 1)
```

Train set was divided into 10 folds to cross validate an optimal shrinkage coefficient.

```
cv_lasso <- cv.glmnet(
  x = x_var,
  y = y_var,
  nfolds = 10,
  family = "binomial",
  alpha = 1)

plot(cv_lasso)
```

And we picked the most shrinking one within a 1 standard error from the minimum one, to retain as low predictors as reasonable.

```
se1_lasso <- glmnet::glmnet(
  x      = x_var,
  y      = y_var,
  alpha  = 1,
```

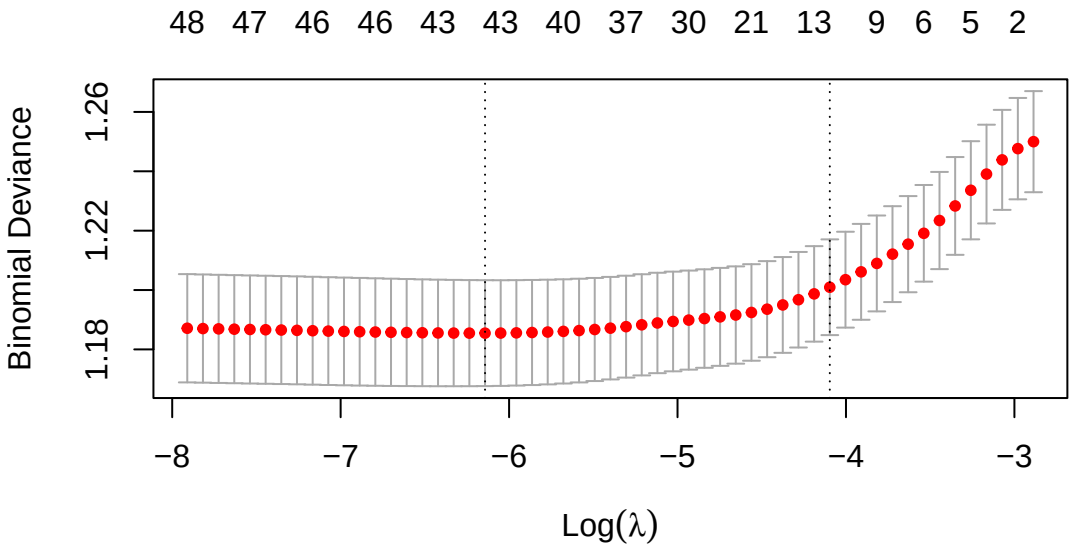


Figure 4: Labmda cross-validation

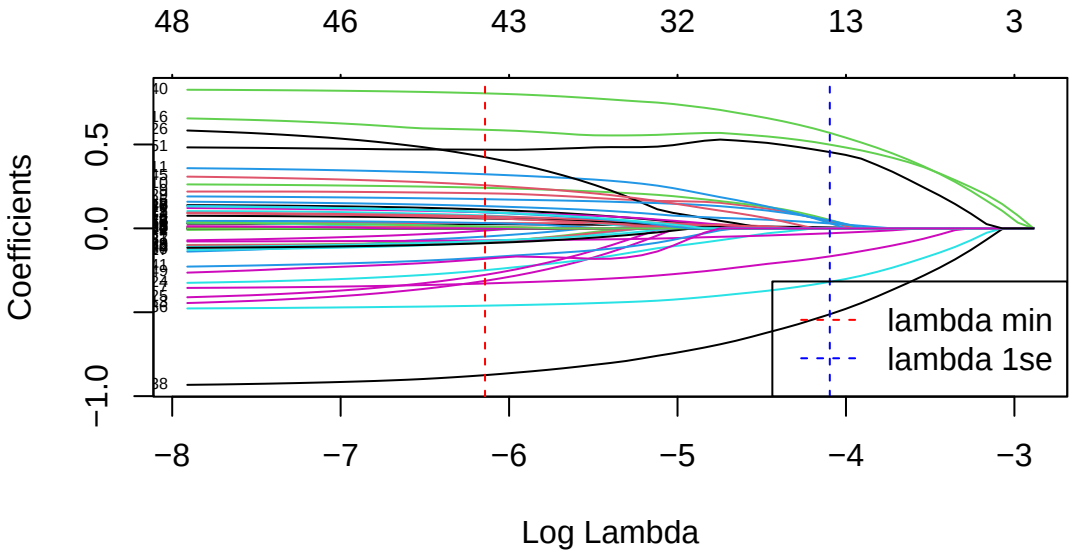


Figure 5: Shrinkage of coefficient at the grow of lambda

```
  family = "binomial",
  lambda = cv_lasso$lambda.1se)

lasso.exc <- as.logical(coef(se1_lasso) == 0)
prop.table(table(excluded = lasso.exc))
```

```
## excluded
##     FALSE      TRUE
## 0.2692308 0.7307692
```

73% of the predictor terms were excluded using LASSO. We then used the selected predictors to fit a classical mutivariable logistic model to mitigate the bias effects related to the shrinkage and access the goodness of frailty fit on the re-sempled training set.

```
frailty.formula <- paste0("frail~", paste(lasso.sel, collapse = "+"))
frailty.model <- glm(frailty.formula, frail.db, family = binomial)

pred.quintiles <- ntile(predict(frailty.model), 5)

round(prop.table(table(
  `Prediction quintiles` = pred.quintiles,
  `Frail (%)` = frail.db$frail), margin = 1)*100, 1)
```

```
##                      Frail (%)
## Prediction quintiles FALSE TRUE
##                    1  85.1 14.9
##                    2  74.2 25.8
##                    3  70.3 29.7
##                    4  63.1 36.9
##                    5  48.9 51.1
```

3.4.2 Short term outcome

To de-noise and summarize all the 18 short term outcomes we used sliced inverse regression. Categorical variables were combined to make ordinal scores (such as grade of kidney failure, intensity of bleeding, and absent-minor-major vascular complication). Two purely categorical variables (implanted valve type and apprach) were excluded.

```
sir.db <- resw.train
sir.db <- cbind(sir.db, estimate.setting(sir.db))
sir.db$frailty_pred <- predict(frailty.model, newdata = sir.db)
sir.db <- preprocess_STO(sir.db)
dr_res <- dr::dr(outcome ~ ., data = sir.db, method = "sir")
dr::plot.dr(dr_res, mark.by.y = T)
```

```
summary(dr_res)$test
```

```
##                  Stat df p.value
## 0D vs >= 1D 401.8086 13       0
```

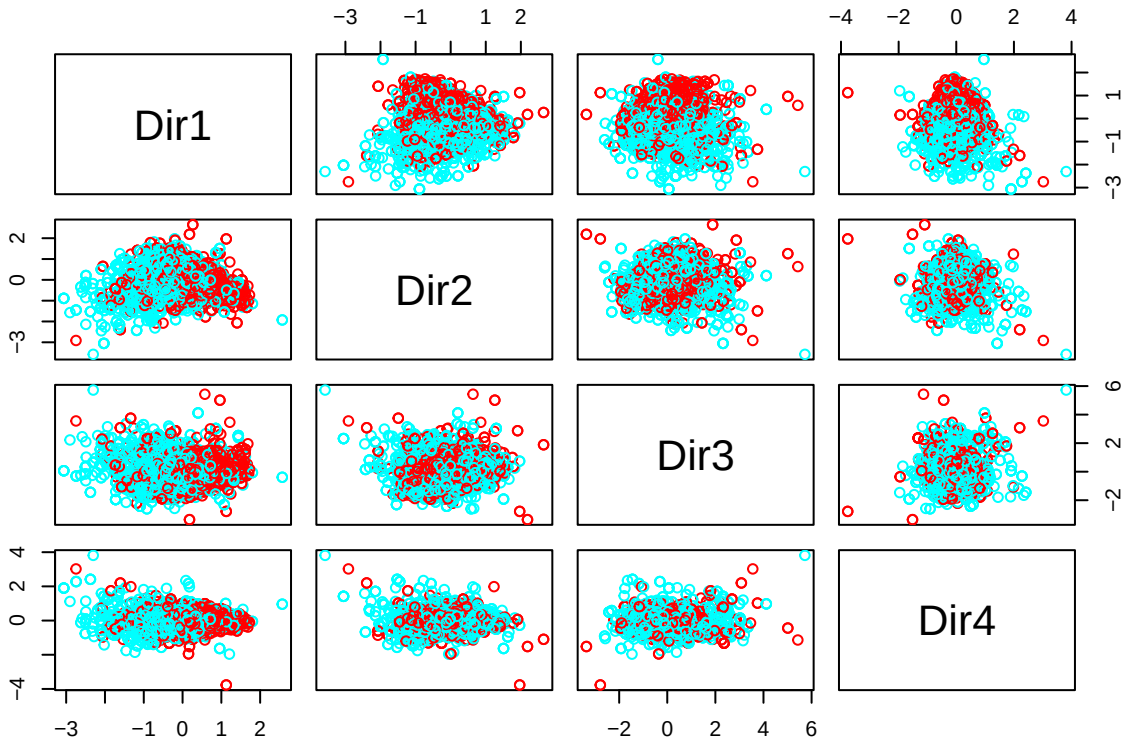


Figure 6: Sliced inverse regression of short term outcomes on MACE

Chi-squared tests indicated that using 2 or more directions might be irrelevant to summarize all the covariance between short term outcomes and MACE. We therefore kept only the first direction. However the direct discrimination capability of the MACE outcome based on this short term outcome approximation was low.

To map baseline covariates to the main sir direction we applied a simple linear model with the short term outcome estimate as dependent variable and baseline covariates as predictors.

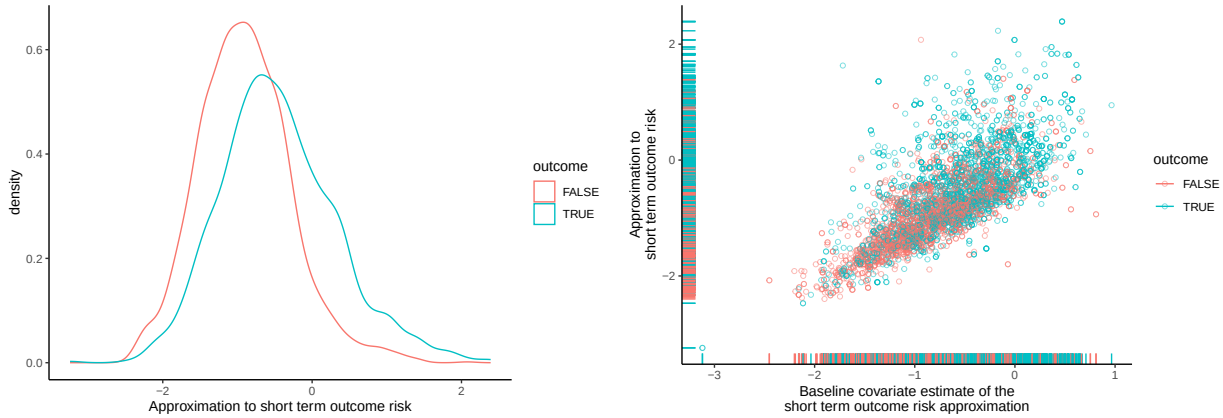


Figure 7: Summarized short term outcome risk by MACE group and aproximation fit to baseline covariates.

```
summary(sto.model)$adj.r.squared
```

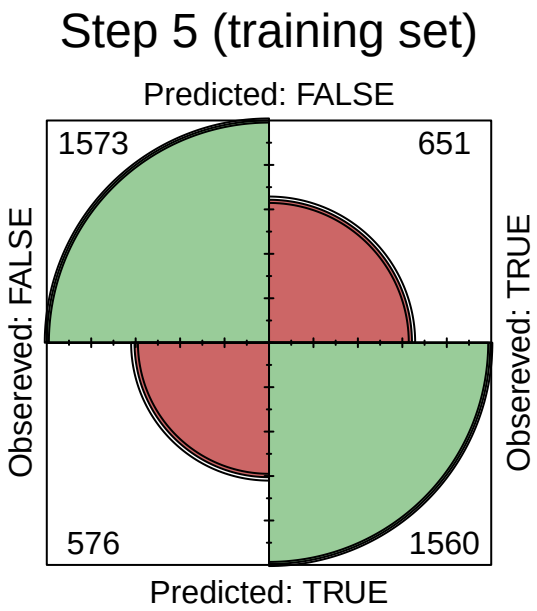
```
## [1] 0.5009899
```

50% of the short term outcome risk approximation in the training set is explanable with the baseline covariats variability. This approximation will then be used as MACE predictor.

3.5 Step 5. Combine predictions

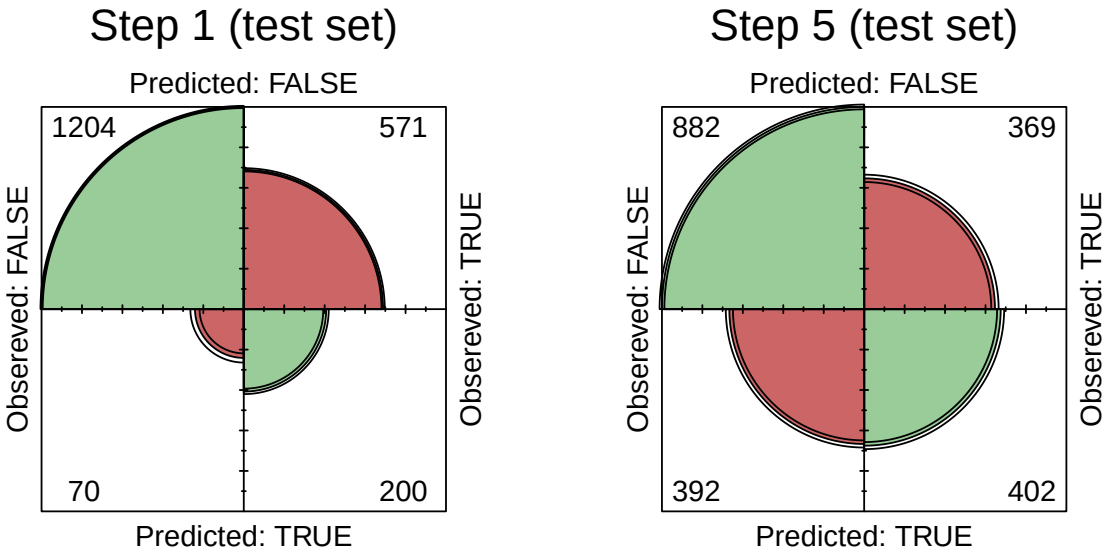
We estimated with baseline covariates: (1) setting, (2) short therm outcome occurrence and (3) frailty. We then combined these new features with baseline covariate as predictor for 1-year MACE occurrence. A generalize additive logistic model was used for the final fit. MACE was used as outcome. All the baseline covariates were used as predictors. The year of intervention was included with its polynomial terms to allow time trends non linearity. All the baseline covariates were also allowed to have a different effect based on frailty, setting and short therm estimates via a two way interaction term.

```
step.5.db <- resw.train
step.5.db_prep <- preprocess.mace(step.5.db)

step.5.model <- glm(
  formula = outcome ~ . * (
    sto + frailty_pred * (Setting_1 + Setting_2 + Setting_3)),
  data    = step.5.db_prep,
  family  = "binomial")

eval.pred(predicted = predict.mace(step.5.model, newdata = step.5.db),
          obsereved = step.5.db$outcome,
          title = "Step 5 (training set)")
```



4 Results

The step 1 model was applied to the test set to have a reference for the improvement due to steps 2 to 5. The model developed at step 5 was applied to the train set. An automatized procedure estimated (1) setting, (2) short therm outcome occurrence and (3) frailty and included them as predictor.



To get a sense results variability (of this specific final model) we used bootstrap to sample with replacement subjects from the test set (in equal sample size) and evaluate predictive performance 100 times.

Table 11: Distribution of the final model performance metrics over 100 times test set replication

	Min.	1st Qu.	Mean	Median	3rd Qu.	Max.
Accuracy	60.2	62.1	62.7	62.7	63.4	65.8
Kappa	15.7	19.9	21.1	21.2	22.3	27.7
AccuracyLower	58.1	60.0	60.6	60.5	61.3	63.7
AccuracyUpper	62.4	64.2	64.8	64.8	65.5	67.9
AccuracyNull	59.8	61.8	62.3	62.3	62.8	65.3
Sensitivity	66.2	68.4	69.2	69.2	69.9	73.1
Specificity	47.7	51.0	52.1	52.1	53.1	56.2
Pos Pred Value	67.9	69.7	70.4	70.4	71.1	73.8
Neg Pred Value	46.3	49.7	50.6	50.6	51.7	55.6
Precision	67.9	69.7	70.4	70.4	71.1	73.8
Recall	66.2	68.4	69.2	69.2	69.9	73.1
F1	67.4	69.0	69.8	69.8	70.5	72.3
Prevalence	59.8	61.8	62.3	62.3	62.8	65.3
Detection Rate	40.7	42.3	43.1	43.1	43.8	45.5
Detection Prevalence	58.7	60.5	61.2	61.1	61.8	64.6
Balanced Accuracy	57.9	60.0	60.6	60.7	61.2	63.9

Over 100 replicas, specificity ranged from 47.5% to 55.6, accuracy from 61.9% to 67.3%. The no information rate (MACE prevalence in the training set) was 35.5%.

5 Limits

The procedure of train-test split could be performed more than once to assess the robustness of the procedure. Multiple imputation was used but a single set of imputation was actually implemented while multiple iteration with other imputation set could also assess the robustness of the procedure. Thresholding could be further optimized by a risk-benefit study.

Steps use very heterogeneous methods with what appeared to be a minimal prediction performance benefit. To improve clarity and presentation of results some analysis could be removed.

6 Conclusion

The final model presented low Accuracy and low Specificity. The biggest performance contribution (based on intermediate results not shown) seemed to be given by re-balancing rather than de-clustering or source maximization. The drop in performance from train to test in step 5 compared to step 1 could be a sign of overfit.

Overall the performance metrics indicates a scoring system not ready for real world clinical application, however, results should be put into context. When stratifying the predicted 1-year MACE risk of TAVR and the predicted 1-year MACE risk of invasive intervention, risk appeared to be well stratified, suggesting that there could be room to such tool for clinical decision support. A specifically designed study with both TAVR and invasive intervention could validate this score and highlight its role for clinical decision support.

Table 12: EUROSCORE and Step 5 model integration

	Euro Score			
TAVR predicted risk	1 quartile	2 quartile	3 quartile	4 quartile
1 quartile	172 (21.5%)	113 (26.5%)	114 (27.2%)	113 (29.2%)
2 quartile	142 (31.7%)	137 (35.0%)	131 (36.6%)	101 (37.6%)
3 quartile	112 (27.7%)	153 (33.3%)	121 (30.6%)	125 (40.0%)
4 quartile	86 (44.2%)	108 (51.9%)	145 (62.1%)	172 (62.8%)

7 Aknowledgement

We thank Vincenzo De Marzo for data access and his subject matter expert contribution to understand the meaning of the data and their clinical relevance.