Cluster probabilities and outcomes

Authors:

- Daniel E. Coral
- Femke Smit
- Elena Santos
- Ali Farzaneh

In this second part of the analysis, we will examine how the clusters we have validated across cohorts are associated with prevalent diseases at the time of clustering, and also assess whether they add significant information for prediction of MACE events and diabetes progression on top of commonly used risk stratification tools.

Libraries and functions

The libraries needed to run this analysis:

In [1]:
```
library(readr)
library(dplyr, warn.conflicts = FALSE)
library(tidyr, warn.conflicts = FALSE)
library(purrr)
library(survival)
```

And the functions we have prepared to facilitate some steps:

In [2]:
```
source("cross_sectional_FX2.R")
```

Loading data needed

Initial input table of biomarkers and basic covariates

The input table is the same table of 10 traits we had prior to run UMAP. Here is a description of this table:

System targeted	Biomarker	Units	Column name
Individual ID	-	-	eid
Blood pressure	Systolic blood pressure	millimeters of mercury (mmHg)	sbp
	Diastolic blood pressure	millimeters of mercury (mmHg)	dbp
Lipid fractions	High density lipoprotein	mmol/L	hdl
	Low density lipoprotein	mmol/L	ldl
	Tryglicerides	mmol/L	tg
Glycemia	Fasting glucose	mmol/L	fg
Liver metabolism	Alanine transaminase	U/L	alt
Fat distribution	Waist-to-hip ratio	cm/cm	whr
Kidney function	Serum creatinine	umol/L	scr
Inflammation	C reactive protein	mg/L	crp
Basic covariates	Current smoking status	1 if yes, 0 if not	smoking
	Sex	String ("Female" or "Male")	sex
	Age	Years	age

Important note: All columns should be there in the units required, and the names should match, so that the functions we have prepared for the analyses work properly. This is true for this and all the following tables we require for our analysis.

This input table has been preprocessed by:

1. Filtering out values that are possible errors in measurement (>5 SD away from the mean in continuous variables).
2. Only including complete cases.
3. Stratifying by sex.

Here is how the input table should look like - a list of two data frames, one for each sex:

In [3]:
```
load("../data/ukb/strat_dat.RData")
```

In [4]:
```
map(strat_dat, head)
```

$Female

A tibble: 6 × 15

eid	age	sex	bmi	whr	sbp	dbp	alt	scr	crp	hdl	tg	ldl	fg	smoking
<dbl>	<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1000117	47	Female	23.8408	0.7254902	147.5	84.0	14.07	61.0	0.24	1.972	0.591	2.252	4.395	0
1000132	43	Female	35.6559	0.8403361	137.0	100.5	18.89	60.5	4.31	1.236	2.037	3.686	5.214	0
1000176	69	Female	38.1271	0.8897638	137.5	93.5	36.39	68.9	3.69	1.601	1.988	4.551	4.266	0
1000223	63	Female	25.4603	0.7789474	163.0	94.0	6.10	67.1	1.29	1.453	2.829	3.491	5.876	0
1000282	48	Female	25.4297	0.7708333	135.5	89.0	9.63	46.2	0.16	2.185	0.722	3.584	5.212	0
1000367	42	Female	19.3280	0.6777778	107.0	72.5	9.34	57.1	0.69	2.346	0.395	3.072	4.649	0

$Male

A tibble: 6 × 15

eid	age	sex	bmi	whr	sbp	dbp	alt	scr	crp	hdl	tg	ldl	fg	smoking
<dbl>	<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1000039	44	Male	36.6959	0.9911504	124.5	64.5	34.97	93.0	3.60	1.158	2.800	3.956	5.427	0
1000071	67	Male	39.4807	0.8857143	179.5	103.0	46.74	68.7	9.41	1.372	1.127	2.311	7.079	0
1000088	60	Male	24.2786	0.8761905	152.0	89.0	13.14	80.6	1.20	0.983	1.590	4.200	5.401	0
1000096	41	Male	26.5744	0.9587629	143.0	90.0	30.32	80.1	6.13	1.041	2.713	4.029	4.239	0
1000109	62	Male	33.8719	1.0818182	156.5	104.5	16.26	89.3	14.42	0.890	2.437	3.525	6.100	0
1000125	66	Male	36.1100	1.0625000	155.0	102.5	25.59	88.7	1.91	1.061	1.320	2.538	4.531	1

Table of validated clusters

The second thing needed is the clusters we have validated. We have put this in an R file called `validclusmod` :

In [5]:
```
load("../data/validclusmod.RData")
print(validclusmod)
```

```
# A tibble: 2 x 3
  sex    residmod          clusmod
  <chr>  <list>            <list>
1 Female <tibble [10 x 6]> <tibble [6 x 4]>
2 Male   <tibble [10 x 6]> <tibble [5 x 4]>
```

This object contais, for each sex:

- `residmod` : The model to obtain residuals for each variable, i.e., the variability beyond what is explained by BMI, adjusting for age and smoking.
- `clusmod` : The clustering model to apply to the residuals.

Table of pre-existing conditions and medications

The third thing we need is a table of pre-existing conditions and medications participants are currently taking:

In [6]:
```
covar_dat <- read_tsv("../data/covar_dat.tsv", show_col_types = FALSE)
head(covar_dat)
```

A tibble: 6 × 15

eid	HT	CHD	Stroke	PAD	CKD	LiverFailure	RA	T2D	T1D	T2Dage	Insulin	AntiDM	AntiHT	LipidLower
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1000027	0	0	0	0	0	0	0	0	0	0.0	0	0	0	0
1000039	0	0	0	0	0	0	0	0	0	0.0	0	0	0	0
1000040	1	0	0	0	0	0	0	0	0	0.0	0	0	0	0
1000053	0	0	0	0	0	0	0	0	0	0.0	0	0	0	0
1000064	1	0	0	0	0	0	0	1	0	49.5	0	1	1	0
1000071	1	0	0	0	0	0	0	1	0	65.5	0	0	1	1

All the columns in this table are coded 1 or 0 representing current diagnosis of a disease or whether the person is taking the medications specified. The exception is `T2Dage` , which is the age of onset of T2D. This is what each column represent:

Group	Column name	Meaning
Diagnoses	HT	Hypertension
	CHD	Coronary heart disease
	Stroke	Stroke
	PAD	Peripheral artery disease
	CKD	Chronic kidney disease
	LiverFailure	Liver failure
	RA	Rheumatoid arthritis
	T2D	Type 2 diabetes
	T1D	Type 1 diabetes
Age at onset	T2Dage	Age at onset of T2D - It is 0 if `T2D` is 0. Needed in SCORE2.
Medication	Insulin	Taking insulin
	AntiDM	Taking medication for diabetes other than insulin
	AntiHT	Taking medication for hypertension
	LipidLower	Taking lipid-lowering medication

If any of the columns in this table are missing in your data, one option is to assume that none in your population had the disease, i.e., you should have a column with 0 for all individuals.

Survival data

Lastly, we need survival data for MACE and diabetes progression. They should look like this:

In [7]:
```
survmacedat <- read_tsv("../data/survmacedat.tsv", show_col_types = FALSE)
head(survmacedat)
```

A tibble: 6 × 3

eid	outcome_value	outcome_timeyrs
<dbl>	<dbl>	<dbl>
1000071	0	10.001369
1000223	1	6.874743
1000324	1	3.101985
1000583	1	3.761807
1001175	1	4.539357
1001892	1	9.185489

In [8]:
```
survdmdat <- read_tsv("../data/survdmdat.tsv", show_col_types = FALSE)
head(survdmdat)
```

A tibble: 6 × 3

eid	outcome_value	outcome_timeyrs
<dbl>	<dbl>	<dbl>
1000109	1	3.600274
1000132	1	1.248460
1004267	1	4.550308
1006281	1	1.957563
1007454	0	9.423682
1010295	1	6.852841

These two tables include individuals followed ***up to 10 years***. This means that any outcome after 10 years is censored. `outcome_value` is 1 if the person experienced the event during the follow-up time and 0 if not. `outcome_timeyrs` is the time of follow-up in years, up to the first event or up to 10 years.

It is important that these tables ***do not include*** individuals who already experience the events we will study. In any case, we will make sure of this in the next step, when we combine all the data. For example, any individual in the `survmacedat` table with a value of 1 in the columns `CHD` , `Stroke` or `PAD` of the `covar_tab` table, will be excluded from the analysis.

In case your cohort does not have survival data, then follow this guideline until the section below entitled "Prevalent diseases and medication".

Calculation of cluster probabilities

With the data needed in place, we can start by calculating cluster allocation probabilities given the biomarker data. For that we will first add a new column called `data` to the `validclusmod` table where we will put the biomarker data for each sex:

In [9]:
```
clusterdfs <- clusterprobcalc(ClusModDf = validclusmod, StratDat = strat_dat)
```

In [10]:
```
print(clusterdfs)
```

```
# A tibble: 2 x 2
  sex    data
  <chr>  <list>
1 Female <tibble [77,207 x 21]>
2 Male   <tibble [67,904 x 20]>
```

Checking that the probabilities were calculated for each sex:

In [11]:
```
head(clusterdfs$data[[1]])
```

A tibble: 6 × 21

eid	age	sex	bmi	whr	sbp	dbp	alt	scr	crp	⋯	tg	ldl	fg	smoking	probBC	probDHT	probDAL	probDLT	probDIS	probDHG
<dbl>	<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	⋯	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1000117	47	Female	23.8408	0.7254902	147.5	84.0	14.07	61.0	0.24	⋯	0.591	2.252	4.395	0	0.1008059	8.982179e-01	8.145289e-08	0.0006249365	0.0002629902	8.816286e-05
1000132	43	Female	35.6559	0.8403361	137.0	100.5	18.89	60.5	4.31	⋯	2.037	3.686	5.214	0	0.3995536	5.942413e-01	2.669602e-03	0.0014612974	0.0009994929	1.074671e-03
1000176	69	Female	38.1271	0.8897638	137.5	93.5	36.39	68.9	3.69	⋯	1.988	4.551	4.266	0	0.9885268	9.507371e-06	5.394490e-05	0.0110183762	0.0002941394	9.725173e-05
1000223	63	Female	25.4603	0.7789474	163.0	94.0	6.10	67.1	1.29	⋯	2.829	3.491	5.876	0	0.9391540	2.297826e-04	5.226406e-02	0.0004234940	0.0001454109	7.783232e-03
1000282	48	Female	25.4297	0.7708333	135.5	89.0	9.63	46.2	0.16	⋯	0.722	3.584	5.212	0	0.7901985	2.027605e-01	5.046971e-07	0.0019040940	0.0015270843	3.609283e-03
1000367	42	Female	19.3280	0.6777778	107.0	72.5	9.34	57.1	0.69	⋯	0.395	3.072	4.649	0	0.9827497	2.825642e-03	2.925853e-07	0.0033190879	0.0093330496	1.772251e-03

Descriptive statistics

At this point we will recheck some of the characteristics of the clusters as we did in our previous script, weighting calculations by cluster probabilities.

The distribution of biomarkers per cluster:

In [12]:
```
markerdistribdf <- markerdistribfx(clusterdfs)
```

In [13]:
```
head(markerdistribdf)
```

A tibble: 6 × 7

sex	Variable	Cluster	Type	N	Summary1	Summary2
<chr>	<chr>	<chr>	<chr>	<dbl>	<chr>	<chr>
Female	whr	BC	Numeric	58707.879	0.82 (0.07)	0.81 (0.7 - 0.77 - 0.86 - 0.96)
Female	whr	DHT	Numeric	7483.777	0.79 (0.05)	0.78 (0.69 - 0.75 - 0.82 - 0.9)
Female	whr	DAL	Numeric	3950.988	0.87 (0.06)	0.86 (0.76 - 0.83 - 0.91 - 0.99)
Female	whr	DLT	Numeric	2835.984	0.84 (0.07)	0.84 (0.71 - 0.79 - 0.89 - 0.98)
Female	whr	DIS	Numeric	2750.474	0.84 (0.07)	0.83 (0.71 - 0.79 - 0.89 - 0.98)
Female	whr	DHG	Numeric	1477.897	0.85 (0.08)	0.85 (0.71 - 0.79 - 0.91 - 1.02)

The effect of BMI on biomarkers specifically within each cluster, adjusted for age and smoking:

In [14]:
```
bmieffmarkerdf <- bmieffmarkerfx(clusterdfs)
```

In [15]:
```
head(bmieffmarkerdf)
```

A tibble: 6 × 6

sex	Variable	Cluster	term	estimate	se
<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>
Female	whr	BC	(Intercept)	0.576753560	1.917410e-03
Female	whr	BC	age	0.001289826	2.768274e-05
Female	whr	BC	smoking	0.020114650	7.707873e-04
Female	whr	BC	bmi	0.006069039	4.309974e-05
Female	whr	DHT	(Intercept)	0.569642086	1.568987e-03
Female	whr	DHT	age	0.001087705	2.096435e-05

Prevalent diseases and medication

To add covariate data to the `alldat` table we will do the following:

In [16]:
```
clusterdfs <- addcovardat(X = clusterdfs, CovarDat = covar_dat)
```

In [17]:
```
print(clusterdfs)
```

```
# A tibble: 2 x 2
  sex    data
  <chr>  <list>
1 Female <tibble [77,151 x 35]>
2 Male   <tibble [67,848 x 34]>
```

Checking again if the columns were added as expected:

In [18]:
```
head(clusterdfs$data[[1]])
```

A tibble: 6 × 35

eid	age	sex	bmi	whr	sbp	dbp	alt	scr	crp	⋯	CKD	LiverFailure	RA	T2D	T1D	T2Dage	Insulin	AntiDM	AntiHT	LipidLower
<dbl>	<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	⋯	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1000117	47	Female	23.8408	0.7254902	147.5	84.0	14.07	61.0	0.24	⋯	0	0	0	0	0	0	0	0	0	0
1000132	43	Female	35.6559	0.8403361	137.0	100.5	18.89	60.5	4.31	⋯	0	0	0	0	0	0	0	0	1	0
1000176	69	Female	38.1271	0.8897638	137.5	93.5	36.39	68.9	3.69	⋯	0	0	0	0	0	0	0	0	1	0
1000223	63	Female	25.4603	0.7789474	163.0	94.0	6.10	67.1	1.29	⋯	0	0	0	0	0	0	0	0	1	1
1000282	48	Female	25.4297	0.7708333	135.5	89.0	9.63	46.2	0.16	⋯	0	0	0	0	0	0	0	0	0	0
1000367	42	Female	19.3280	0.6777778	107.0	72.5	9.34	57.1	0.69	⋯	0	0	0	0	0	0	0	0	0	0

We will first count the number of individuals with disease in each cluster. Here we will also count the number of individuals taking each class of medications in each cluster.

In [19]:
```
countcovarsdf <- countcovarsfx(clusterdfs)
```

In [20]:
```
head(countcovarsdf)
```

A tibble: 6 × 6

sex	Cluster	Covariate	Nclus	Ncases	Nnoncases
<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>
Female	probBC	HT	58673.33	12946.70991	45726.62
Female	probBC	CHD	58673.33	1659.00050	57014.33
Female	probBC	Stroke	58673.33	848.10719	57825.23
Female	probBC	PAD	58673.33	147.89695	58525.44
Female	probBC	CKD	58673.33	67.25413	58606.08
Female	probBC	LiverFailure	58673.33	50.16266	58623.17

We will use this table to calculate prevalences and compare prevalences across clusters.

We are also interesting in looking at the proportion of individuals receiving medications in each cluster, stratified by each condition. This is obtained with the following function:

In [21]:
```
countdxmed <- countdxmedfx(clusterdfs)
```

In [22]:
```
head(countdxmed)
```

A tibble: 6 × 6

sex	Dx	Cluster	Med	Nnoncases	Ncases
<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>
Female	CHD	probBC	NoMed	42000.949	206.108367
Female	CHD	probBC	Insulin	137.000	17.000000
Female	CHD	probBC	AntiDM	615.000	67.000000
Female	CHD	probBC	AntiHT	10480.000	1320.000000
Female	CHD	probBC	LipidLower	8573.000	1206.000000
Female	CHD	probDAL	NoMed	2708.061	9.496002

We will also formally test the association between cluster allocation and diseases using logistic regressions where the outcome is each disease and the predictors are the cluster allocations. We will have two models for each disease, one with only clusters, and a second one adjusting for medication.

In [23]:
```
assocdxdf <- assocdxfx(clusterdfs)
```

In [24]:
```
print(assocdxdf)
```

```
# A tibble: 36 x 5
   sex    Dx_name model        estimates  varcovmat
   <chr>  <chr>   <chr>        <list>     <list>
 1 Female HT      OnlyClusters <dbl [6]>  <dbl [6 x 6]>
 2 Female HT      FullModel    <dbl [23]> <dbl [23 x 23]>
 3 Female CHD     OnlyClusters <dbl [6]>  <dbl [6 x 6]>
 4 Female CHD     FullModel    <dbl [23]> <dbl [23 x 23]>
 5 Female Stroke  OnlyClusters <dbl [6]>  <dbl [6 x 6]>
 6 Female Stroke  FullModel    <dbl [23]> <dbl [23 x 23]>
 7 Female PAD     OnlyClusters <dbl [6]>  <dbl [6 x 6]>
 8 Female PAD     FullModel    <dbl [23]> <dbl [23 x 23]>
 9 Female CKD     OnlyClusters <dbl [6]>  <dbl [6 x 6]>
10 Female CKD     FullModel    <dbl [23]> <dbl [23 x 23]>
# i 26 more rows
```

Adding survival data

As explained before, since we want to be careful when adding survival data for analysis, we have prepared a function separately for both outcomes, and making sure we exclude individuals who already experience the events under study:

In [25]:
```
clustersurvdfs <- addsurvdat(X = clusterdfs, SurvMACEDf = survmacedat, SurvDMDf = survdmdat)
```

In [26]:
```
print(clustersurvdfs)
```

```
# A tibble: 4 x 3
  sex    outcome data
  <chr>  <chr>   <list>
1 Female MACE    <tibble [73,378 x 34]>
2 Male   MACE    <tibble [60,348 x 33]>
3 Female DM      <tibble [34,581 x 32]>
4 Male   DM      <tibble [29,006 x 31]>
```

`data` now contain the data necessary to run survival analysis.

Rates of outcomes by cluster

Similar to what was done in the cross sectiona setting, we will calculate the number of cases and the total follow-up in each cluster using the weighted approach:

In [27]:
```
ratesbyclus <- ratesclusfx(clustersurvdfs)
```

In [28]:
```
head(ratesbyclus)
```

A tibble: 6 × 5

sex	outcome	Cluster	Ncases	TPT
<chr>	<chr>	<chr>	<dbl>	<dbl>
Female	MACE	probBC	2093.9159	492558.11
Female	MACE	probDAL	200.6431	33353.95
Female	MACE	probDHG	95.5165	11139.54
Female	MACE	probDHT	285.8513	64238.47
Female	MACE	probDIS	143.3605	22584.02
Female	MACE	probDLT	107.7127	23773.73

We will also do this by medication status:

In [29]:
```
ratesbyclusmed <- ratesclusmedfx(clustersurvdfs)
```

In [30]:
```
head(ratesbyclusmed)
```

A tibble: 6 × 6

sex	outcome	Cluster	Med_name	Ncases	TPT
<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>
Female	MACE	probBC	AntiDM	67.829857	4966.2019
Female	MACE	probBC	AntiHT	756.785907	85741.8406
Female	MACE	probBC	Insulin	16.833763	1097.1488
Female	MACE	probBC	LipidLower	524.094312	69458.7921
Female	MACE	probBC	NoMed	1123.837308	368913.8639
Female	MACE	probDAL	AntiDM	6.634919	339.1024

We will also calculate overall and cluster-specific cumulative incidence rates using the Kaplan-Meier method:

In [31]:
```
kmestdf <- kmestfx(clustersurvdfs)
```

In [32]:
```
head(kmestdf)
```

A tibble: 6 × 6

sex	outcome	Cluster	risk	lower	upper
<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>
Female	MACE	Overall	0.04651655	0.04468811	0.04834149
Female	MACE	probBC	0.04399324	0.04194391	0.04603818
Female	MACE	probDHT	0.04512474	0.03942353	0.05079211
Female	MACE	probDAL	0.06075845	0.05173248	0.06969851
Female	MACE	probDLT	0.04782438	0.03785284	0.05769258
Female	MACE	probDIS	0.06237728	0.05175902	0.07287664

Cox models

To quantify the association of clusters to MACE, as well as its potential contribution for prediction, we will compare two models. The reference model will include all predictors that are part of SCORE2, the risk stratification tool for CVD recommended by the European Society of Cardiology *(SCORE2 working group and ESC Cardiovascular risk collaboration, 2021)*. We will use a version of this score that has been validated in diabetic populations, and includes some additional clinically useful predictors *(SCORE2-Diabetes Working Group and the ESC Cardiovascular Risk Collaboration, 2023)*. Additionally, for the sake of completeness, some important pre-existing conditions and pharmacological treatments, such as hypertension/antihypertensives, as well as any predictor that we had in our initial input table that are not part of SCORE2, will also be included. We will compare this reference model to one that includes also the cluster probabilities and then we'll compare the ability of these two models to predict MACE.

Similarly, for diabetes we will also fit two models, one containing all biomarkers and another one containing the biomarkers plus the cluster probabilities.

The way we will introduce the cluster probabilities into the Cox models will be employing the log-ratio transformation *(Coenders & Pawlowsky-Glahn, 2020)*.

Fitting models

In [33]:
```
coxmoddf <- coxmodels(clustersurvdfs)
```

In [34]:
```
print(coxmoddf)
```

```
# A tibble: 4 x 7
  sex    outcome data                   survdf   NullMod    mod_base mod_clus
  <chr>  <chr>   <list>                 <list>   <list>     <list>   <list>
1 Female MACE    <tibble [73,378 x 34]> <tibble> <cxph.nll> <coxph>  <coxph>
2 Male   MACE    <tibble [60,348 x 33]> <tibble> <cxph.nll> <coxph>  <coxph>
3 Female DM      <tibble [34,581 x 32]> <tibble> <cxph.nll> <coxph>  <coxph>
4 Male   DM      <tibble [29,006 x 31]> <tibble> <cxph.nll> <coxph>  <coxph>
```

Here `mod_base` contains the baseline model, while `mod_clus` contains the baseline plus clusters model. `mod_null` contains the null model, which we will use to calculate our metrics.

Coefficient estimates

To properly calculate the effect of clusters, we need the coefficients estimated by the log-ratio models as well as their covariance, contained in the `estimates` and `varcovmat` columns. To properly calculate the expected risk for a given phenotype, we need first the means of all variables in the model, contained in the column `Means` and the parameters of the baseline hazard, contained in the `Afit` column.

In [36]:
```
survcoefs <- survcoefx(coxmoddf)
```

In [37]:
```
print(survcoefs)
```

```
# A tibble: 8 x 7
  sex    outcome model estimates  varcovmat       Means      Afit
  <chr>  <chr>   <chr> <list>     <list>          <list>     <list>
1 Female MACE    base  <dbl [30]> <dbl [30 x 30]> <dbl [30]> <named list [10]>
2 Female MACE    clus  <dbl [35]> <dbl [35 x 35]> <dbl [35]> <named list [10]>
3 Male   MACE    base  <dbl [30]> <dbl [30 x 30]> <dbl [30]> <named list [10]>
4 Male   MACE    clus  <dbl [34]> <dbl [34 x 34]> <dbl [34]> <named list [10]>
5 Female DM      base  <dbl [22]> <dbl [22 x 22]> <dbl [22]> <named list [10]>
6 Female DM      clus  <dbl [27]> <dbl [27 x 27]> <dbl [27]> <named list [10]>
7 Male   DM      base  <dbl [22]> <dbl [22 x 22]> <dbl [22]> <named list [10]>
8 Male   DM      clus  <dbl [26]> <dbl [26 x 26]> <dbl [26]> <named list [10]>
```

Comparison of predictive ability

To assess the predictive ability of the two nested models, we will use the gold-standard method: the likelihood ratio test. Given the wide use of the c-statistic, we will also use this metric. However, comparing two c-statistics is not as powerful as the likelihood ratio test.

In [38]:
```
compmoddf <- comparemods(coxmoddf)
```

In [39]:
```
compmoddf
```

A tibble: 4 × 18

sex	outcome	LL0	LLBase	NVBase	LLBaseCl	NVBaseCl	LRTstat	LRTdf	LRTp	AdeqInd	CBase	CseBase	CBaseCl	CseBaseCl	Cdiff	Cdiffse	Cdiffp
<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Female	MACE	-32431.04	-31447.77	30	-31433.72	35	28.10242	5	3.475851e-05	0.9859111	0.7264359	0.004463723	0.7274166	0.004462602	0.000980733	0.0005080031	5.353687e-02
Male	MACE	-53605.73	-52507.15	30	-52447.47	34	119.35444	4	7.337351e-25	0.9484770	0.6854264	0.003528423	0.6905966	0.003516788	0.005170237	0.0008662478	2.393825e-09
Female	DM	-12041.25	-10901.81	22	-10890.46	27	22.70912	5	3.836388e-04	0.9901333	0.8543404	0.005291641	0.8553759	0.005277785	0.001035494	0.0005482979	5.895081e-02
Male	DM	-16105.26	-14836.24	22	-14830.42	26	11.63925	4	2.024552e-02	0.9954350	0.8210544	0.005080388	0.8223072	0.005056662	0.001252796	0.0004243829	3.156796e-03

Some details of these columns:

- `LRTp` is the p-value of the likelihood ratio test comparing models with or without cluster allocations.
- `AdeqInd` is the adequacy index comparing likelihood ratios of the two models. 1 minus this value represent the fraction of added information by cluster allocation.
- `cdiffp` is the p-value of the difference between c-statistics of the two models.

While p-values of both the likelihood ratio tests and the difference between C-statistics show evidence of added value, the adequacy index quantifies how much additional information is obtained when cluster allocations are added on top of the baseline model.

Adequacy index by cluster

Next we quantify how much is the added value of the new model to each cluster. We do this by recalculating the log likelihood of each model but this time weighting individuals by their cluster probabilities.

In [40]:
```
adeqindbyclus <- AdeqIndClusFx(coxmoddf)
```

In [41]:
```
adeqindbyclus
```

A tibble: 22 × 4

sex	outcome	Cluster	AdeqInd
<chr>	<chr>	<chr>	<dbl>
Female	MACE	probBC	0.9802349
Female	MACE	probDHT	0.9850214
Female	MACE	probDAL	1.0000000
Female	MACE	probDLT	1.0000000
Female	MACE	probDIS	1.0000000
Female	MACE	probDHG	1.0000000
Male	MACE	probBC	0.9530781
Male	MACE	probDAL	0.8981550
Male	MACE	probDLT	1.0000000
Male	MACE	probDIS	0.8448494
Male	MACE	probDHG	0.9544304
Female	DM	probBC	0.9929381
Female	DM	probDHT	0.9880135
Female	DM	probDAL	0.9795036
Female	DM	probDLT	0.9802605
Female	DM	probDIS	0.9712778
Female	DM	probDHG	1.0000000
Male	DM	probBC	0.9929576
Male	DM	probDAL	0.9778296
Male	DM	probDLT	0.9943603
Male	DM	probDIS	1.0000000
Male	DM	probDHG	1.0000000

Adequacy index by MACE probability given by baseline model

We are also interested in how the more complex model behaves along the scale of MACE probabilities given by the baseline model:

In [42]:
```
adeqindbypre <- AdeqIndByPreFx(coxmoddf)
```

In [43]:
```
head(adeqindbypre)
```

A tibble: 6 × 4

sex	outcome	threshold	AdeqInd
<chr>	<chr>	<dbl>	<dbl>
Female	MACE	0.00	0.9859111
Female	MACE	0.01	0.9815227
Female	MACE	0.02	0.9755176
Female	MACE	0.03	0.9667583
Female	MACE	0.04	0.9535612
Female	MACE	0.05	0.9255606

Decision curve analysis

The last step in assessing clinical utility of clustering allocations is to perform a decision curve analysis. First we will assess the overall net benefit of both models:

In [44]:
```
dcares <- DCurvFx(coxmoddf)
```

In [45]:
```
head(dcares)
```

A tibble: 6 × 10

sex	outcome	pred	n	threshold	pos_rate	tp_rate	fp_rate	net_benefit	net_intervention_avoided
<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Female	MACE	base	73378	0.00	0.04662649	0.04662649	0.9533735	0.04662649	NaN
Female	MACE	base	73378	0.01	0.04662649	0.04589500	0.8266008	0.03754549	0.05435516
Female	MACE	base	73378	0.02	0.04662649	0.04369664	0.6704002	0.03001501	0.13941108
Female	MACE	base	73378	0.03	0.04662649	0.04078583	0.5411188	0.02405019	0.22340673
Female	MACE	base	73378	0.04	0.04662649	0.03677727	0.4268085	0.01899358	0.29018374
Female	MACE	base	73378	0.05	0.04662649	0.03215838	0.3282494	0.01488210	0.35023019

We will also calculate this by cluster:

In [46]:
```
dcaclusres <- DCurvbyClFx(coxmoddf)
```

In [47]:
```
head(dcaclusres)
```

A tibble: 6 × 11

sex	outcome	Cluster	pred	n	threshold	pos_rate	tp_rate	fp_rate	net_benefit	net_intervention_avoided
<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Female	MACE	probBC	base	55802.39	0.00	0.04412461	0.04412461	0.9558754	0.04412461	NaN
Female	MACE	probBC	base	55802.39	0.01	0.04412461	0.04335303	0.8127737	0.03514320	0.06671509
Female	MACE	probBC	base	55802.39	0.02	0.04412461	0.04095127	0.6476546	0.02773383	0.15272697
Female	MACE	probBC	base	55802.39	0.03	0.04412461	0.03812193	0.5140035	0.02222492	0.24778520
Female	MACE	probBC	base	55802.39	0.04	0.04412461	0.03393248	0.3977640	0.01735898	0.31350023
Female	MACE	probBC	base	55802.39	0.05	0.04412461	0.02939596	0.3005098	0.01357965	0.37552111

Interaction between clusters and medications

Finally, we will assess the interaction between certain medications and clusters:

In [48]:
```
interactmods <- interactmodfx(coxmoddf)
```

In [49]:
```
print(interactmods)
```

```
# A tibble: 24 x 8
   sex    outcome Med_name   model       estimates varcovmat  Means Afit
   <chr>  <chr>   <chr>      <chr>       <list>    <list>     <lis> <list>
 1 Female MACE    Insulin    OnlyClusMod <dbl>     <dbl[...]> <dbl> <named list>
 2 Female MACE    Insulin    FullMod     <dbl>     <dbl[...]> <dbl> <named list>
 3 Female MACE    AntiDM     OnlyClusMod <dbl>     <dbl[...]> <dbl> <named list>
 4 Female MACE    AntiDM     FullMod     <dbl>     <dbl[...]> <dbl> <named list>
 5 Female MACE    AntiHT     OnlyClusMod <dbl>     <dbl[...]> <dbl> <named list>
 6 Female MACE    AntiHT     FullMod     <dbl>     <dbl[...]> <dbl> <named list>
 7 Female MACE    LipidLower OnlyClusMod <dbl>     <dbl[...]> <dbl> <named list>
 8 Female MACE    LipidLower FullMod     <dbl>     <dbl[...]> <dbl> <named list>
 9 Male   MACE    Insulin    OnlyClusMod <dbl [9]> <dbl[...]> <dbl> <named list>
10 Male   MACE    Insulin    FullMod     <dbl>     <dbl[...]> <dbl> <named list>
# i 14 more rows
```

Saving data

As done before, we will ask you to save an R file that does not contain any individual data, only summary statistics, as follows:

In [50]:
```
result_file2 <- list(
    MarkerDistrib = markerdistribdf,
    BMIeffOnMarker = bmieffmarkerdf,
    CountCovars = countcovarsdf,
    CountDXMeds = countdxmed,
    CrossSectAssoc = assocdxdf,
    RatesByClus = ratesbyclus,
    RatesByClusMeds = ratesbyclusmed,
    KaplanMeierDF = kmestdf,
    SurvCoefs = survcoefs,
    Comparison = compmoddf,
    AdeqIndByClus = adeqindbyclus,
    AdeqIndByPre = adeqindbypre,
    DCARes = dcares,
    DCAClusREs = dcaclusres,
    InteractMods = interactmods
)
```

In [51]:
```
save(
    result_file2,
    file = "../data/ukb/result_file2.RData"
)
```

This file should be uploaded to the respective folder of the cohort in Teams:

> CrossWP > Analyst working groups > WG1 > UMAP_project > *cohort_name* > data

---

References

Coenders, G., & Pawlowsky-Glahn, V. (2020). On interpretations of tests and effect sizes in regression models with a compositional predictor. *SORT-Statistics and Operations Research Transactions*, *44*(1), 201–220. https://doi.org/10.2436/20.8080.02.100

SCORE2 working group and ESC Cardiovascular risk collaboration. (2021). SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. *European Heart Journal*, *42*(25), 2439–2454. https://doi.org/10.1093/eurheartj/ehab309

SCORE2-Diabetes Working Group and the ESC Cardiovascular Risk Collaboration. (2023). SCORE2-Diabetes: 10-year cardiovascular risk estimation in type 2 diabetes in Europe. *European Heart Journal*, ehad260. https://doi.org/10.1093/eurheartj/ehad260