Appendix 3

Examples of R-code with specified DSA and cvDSA algorithms used for the illustrative analysis

A.DSA-selected Treatment Model III

```
library(DSA)

a<-as.matrix(read.csv("ns1053_glm_FINAL.csv",header = TRUE, sep = ",", quote="\"",
dec=”.”,fill= TRUE,comment.char=” “))

b<-as.data.frame(a)

#median-center age

agecen<-b[,1]-70

#replace age variable with age-centered variable

ns<-cbind(agecen,b[,2:9])

# “ns” defines a data frame object with single line data (i.e., line for each subject) that
# includes baseline covariates used in the selection of the treatment model.
# Data should include only variables under consideration for the model. Variable
#names and coding are provided:

#Agecen, Sex (1=F, 0=M), Cardiovascular Disease (Cardio 1=Yes, 0=No), Secondhand
#smoke exposure (Total), Body Mass Index (Bmass), Serum cholesterol (Hdl, Ldl),
#Diabetes (Diabnm 1=Yes, 0=No). Agecen, Second-hand smoke, Body Mass Index, Hdl,
#and Ldl are continuous variables.

result <- DSA(fev1 ~ 1, data = ns, maxsize = 8, maxorderint = 2, maxsumofpow = 2)
 summary(result)
```

Arguments:

```
        DSA(formula,data,family=gaussian,maxsize,maxoderint,maxsumofpow)
```

1. “formula” indicates a description of base model which specifies the independent/response variables and all terms forced in the final model. These forced terms might represent part of the model that may be known to the user and incorporated as a “fixed” part of the model. Typically, “formula” is set to ‘Y ~ 1’ when no terms are forced in the final model.
2. *“data”* is a data frame that contains both response variables and candidate covariates for consideration in the procedure.

3. *“family”* is currently ‘binomial’ or ‘gaussian’(default)
4. *“maxsize”* specifies the maximum number of terms in the final model
5. *“maxorderint”* specifies the maximum order of interactions in final model
6. *“maxsumofpow”* specifies the maximum sum of power for each term of the model (e.g., if set to ‘2’, a term could have a polynomial order of 2 or the term could be an interaction of two variables)

Results with the DSA algorithm can vary because v-fold splits of the data are assigned at random. Options are available in the package for setting the v-fold splits for exact, reproducible results.

B. DSA-Selected MSM 8

```
library(cvDSA)

a<-as.matrix(read.csv("ns1053_msm_FINAL.csv",header = TRUE, sep = ",", quote="\"",
dec=".",fill = TRUE, comment.char=""))

b<-as.data.frame(a)

#median-center age

agecen<-b[,2]-70

#replace age variable with age-centered variable

ns0<-cbind(b[,1],agecen)
ns1<-cbind(ns0,b[,3:13])

# “ns1” defines a data frame object with multiple lines of data per subject (i.e., single # line
for each subject for each 6 month interval of follow-up).
#Data can include other variables not considered for inclusion in the model.

result1<-cvMSM(y=deathcv, a=fev1, v=cbind(agecen, sex), w=cbind(agecen,
sex,BMASS,HDL,LDL,cardio,diabnm,total), data=ns1,yfamily='binomial',
afamily='gaussian',model.msm=list(Model=NULL,Size=6,Int=2),
model.aw=list(Model="sex+agecen+sex:agecen+BMASS^2+BMASS:sex"),
model.av=list(Model="sex+agecen+BMASS"),
wt.censor=C_N_WTS, ncv=5, mapping=’IPTW’, fitting=’IPTW’, stable.wt=T, rep.ID=T,
ID=ID)
```

Arguments:

1. “*y=deathcv*” indicates the outcome variable, “*a=fev1*” is the treatment variable, *“v= cbind(agecen, sex)”* and “*w=cbind(agecen, sex,…)*” are the covariates in MSM and the baseline covariates, respectively. “*data=ns1*” is a

dataframe with all the variables for analysis as described above.

2. "*yfamily*" and "*afamily*" indicate the distribution of the outcome variable and the treatment variable, respectively.
3. The following parameters *model.msm, model.aw, model.av* indicate models for *MSM, g(A|W), g(A|V)*, respectively. If the user wishes to fit a particular causal model, he/she can specify the model---e.g., *model. msm=list(Model="agecen+fev1+sex");* otherwise, the cvDSA will select a model given the model search criteria that are provided by the user.
4. "*wt.censor*" is an option to multiply extra weights to the loss function; C_N_WTS is a pre-calculated censoring weight. If "*wt.censor=NULL*", no extra weight will be used.
5. "*ncv=5*" specifies 5-fold cross-validation.
6. "*stabilized.wt=T*" indicates this IPTW estimator uses a stabilized weight -- *g(A|V)/g(A|W)*.
   "*cross-validation='IPTW'"* and *"empirical='IPTW'"* tell the function how the cross-validation risk and the empirical risk are calculated (they could be different). Other options are 'G-comp' and 'DR'.
7. *"rep.ID=T"* indicates that the data include observations with repeating IDs ---i.e., multiple lines of data per subject.
8. *"ID=ID"* used to identify subject IDs

Documentation ("help" files) for the DSA and cvDSA algorithms are included in the packages and include additional details and options