

# SAS macros for estimation of the cumulative incidence functions based on a Cox regression model for competing risks survival data.

S. Rosthøj<sup>1</sup>, P. K. Andersen<sup>1</sup> and S. Z. Abildstrøm<sup>2</sup>.

<sup>1</sup> Department of Biostatistics, University of Copenhagen, Denmark,  
(sr@biostat.ku.dk, pka@biostat.ku.dk).

<sup>2</sup> National Institute of Public Health, Copenhagen, Denmark,  
(stabil@dadlnet.dk).

The SAS macros `CumInc` and `CumIncV` are used for estimation of the cumulative incidence functions based on a Cox regression model for the cause specific hazards in a competing risks model. In addition `CumIncV` computes the estimated variances of the estimated cumulative incidence functions. The macros are available from the page [www.pubhealth.ku.dk/~pka/](http://www.pubhealth.ku.dk/~pka/). Here, the macro `CumInc` is found in the file `CumInc.sas` while `CumIncV` is found in `CumIncV.sas`.

Some work has to be done to prepare the data for the use of the macros and how to do this is described in this document. First the data has to be rearranged as described in Section 1. Next a Cox regression model has to be fitted using the SAS procedure `PROC PHREG` but, dependent on the macro, different output from this procedure is needed. How to establish the right output and how to use the macros is described in Section 2 (`CumInc`) and 3 (`CumIncV`).

The preparation of the data and the use of the macros are demonstrated by an example on survival with malignant melanoma.

## 1. PREPARATION OF THE THE DATA.

Cox regression in the competing risks model is usually performed by fitting separate models for each cause of failure. The hazard corresponding to a failure of a specific cause is analyzed considering failures of other causes as censored observations. Let  $k$  denote the number of causes of failure. Instead of fitting  $k$  separate Cox regressions it is possible to make a single analysis of the  $k$  cause specific hazards at the same time. This can be done using the methods described in Andersen et al. (1993) p. 495:

Assume the data set has  $n$  observations containing the failure time, the cause of failure and the covariates. A new data set with  $kn$  records has to be created by stacking the data set  $k$  times. Next, a numeric stratum indicator must be created i.e. by letting  $h = 1$  for the first  $n$  records corresponding to the first failure type,  $h = 2$  for the next  $n$  records corresponding to the second failure type etc. It is important, when using the macros, that the  $n$  observations corresponding to the failure type

chosen as the first failure type gets the smallest value of the stratum indicator  $h$ , the  $n$  observations corresponding to the failure type chosen as the second failure type gets the second smallest value of the stratum indicator  $h$ , etc. Furthermore, a failure indicator  $D$  has to be defined attaining the value 1 for each observation of death of cause 1 in stratum 1, for each observation of death of cause 2 in stratum 2 etc., and 0 otherwise. For each covariate effective for a specific cause of failure a cause-specific covariate has to be included in the stacked data set: If the covariate  $Z$  is effective for failure of cause 1 a covariate  $Z_1$  has to be defined letting  $Z_1 = Z$  for the  $n$  observations corresponding to stratum 1 (observations  $1, \dots, n$ ), and letting  $Z_1 = 0$  otherwise. If this covariate is also effective for failure of cause 2, but has a different effect on this type of failure, a covariate  $Z_2$  has to be defined letting  $Z_2 = Z$  for the  $n$  observations corresponding to stratum 2 (observations  $n + 1, \dots, 2n$ ) and 0 otherwise. If the covariate  $Z$  is assumed to have exactly the same effect on failures of causes 1 and 2 a covariate  $Z_{12}$  equal to  $Z$  both in stratum 1 and 2 and 0 otherwise has to be included. If no covariates are assumed to have identical effects on several cause-specific hazards then a stratified Cox regression analysis on the stacked data set will give the same results as the  $k$  Cox regressions mentioned above. For models where some covariates do have exactly the same effect for several causes it is not possible to fit  $k$  separate Cox regressions but instead the stacked data set technique has to be used.

The data are described in example I.3.1 in Andersen et al. (1993) and consists of  $n = 205$  observations for patients with malignant melanoma. Two causes of failure are present: 1) death from malignant melanoma and 2) death from other causes. Two covariates, sex and tumor thickness, are included in the Cox regression analysis. In example VII.2.5 of Andersen et al. (1993) it is demonstrated that it is reasonable to assume that sex has exactly the same effect on both causes of failure whereas tumor thickness only has an effect on failure of cause 1.

The text data file `MalignantMelanoma.dat`, also available from the page [www.pubhealth.ku.dk/~pka/](http://www.pubhealth.ku.dk/~pka/), contains the variables `id`, `time`, `thick`, `sex` and `cause`, where `id` identifies the patient, `time` is the failure time of the patient, `thick` is the tumor thickness in mm minus the mean tumor thickness of 2.92 mm, `sex` is 0 for females, 1 for males and finally `cause` attains the value 1 if the patient died from malignant melanoma, 2 if the patient died of other causes and 0 if the patient was right censored.

The data step to create the stacked data set is:

```
DATA melanoma;
  INFILE 'MalignantMelanoma.dat';
  INPUT id time thick sex cause;

DATA melanom2;
  SET melanoma melanoma;
  h = 1 + (_N_ GT 205);
  d = (cause=1)*(h=1) + (cause=2)*(h=2);
  thick1 = thick*(h=1);
```

The first 10 records and records no. 206-215 (no.  $n + 1$  to  $n + 10$ ) of the data set `melanom2` are printed in Appendix A1 p. 8.

The cumulative incidence functions are computed for a fixed value of the covariates. Output from the SAS procedure `PROC PHREG` based on this value of the covariates is needed. A separate data set containing the value of these covariates has to be made in order to give `PROC PHREG` this value as an input. This data set must contain a variable corresponding to each covariate included in the `MODEL` statement of the `PROC PHREG` procedure and these variables must have the same names as the covariates. Furthermore, there has to be as many records as strata (failure types) and record no. 1 corresponds to the first failure type, record no. 2 corresponds to the second failure type, etc. In each of these records the value of a variable has to be the fixed value of the corresponding covariate if the covariate has an effect on the failure type corresponding to the number of the record, and has to be 0, if the covariate does not have an effect on this failure type.

Assume that it is chosen to estimate the cumulative incidence functions for a male with a tumor thickness of 2 mm. above the average value of 2.92 mm. The data set is made in this way:

```
DATA cov;
  INPUT sex thick1;
CARDS;
  1 2
  1 0
  ;
```

If instead the estimate of the cumulative incidence function for a male with a tumor thickness of 2.92 mm. is needed, the fourth line (1 2) of the above data step has to be replaced by the line 1 0.

The next step is to perform the stratified Cox regression analysis. This is done using the SAS procedure `PROC PHREG`. Different output from this procedure is needed depending on which macro is used.

## 2. USING THE MACRO CumInc.

When a model based on a single covariate `Covariat` is used, the PHREG procedure has to be used in the following way:

```
PROC PHREG data=StackDat;  
  MODEL Time*D(0)= Covariat;  
  STRATA Strata;  
  BASELINE OUT=ciData COVARIATES=cov SURVIVAL=Surv / NOMEAN METHOD=CH;
```

`StackDat` is the name of the stacked data set created above, `Time` is the failure time variable, `D` is the failure indicator  $D$ , `Strata` is the stratum indicator ( $h$  as defined above) and `cov` is the name of the data set containing the fixed value of the covariates.

*Important: No statements must be omitted and the NOMEAN METHOD=CH statement must not be changed. Furthermore, the output data set ciData must not be changed.*

The variable `Surv` in the output data set `ciData` from the BASELINE statement contains the exponential of the negative estimated cumulative cause-specific hazards at different time points for each combination of strata and covariates.

In the malignant melanoma example the Cox regression can be performed using the code:

```
PROC PHREG data=melanom2;  
  MODEL time*d(0)= sex thick1;  
  STRATA h;  
  BASELINE OUT=ciData COVARIATES=cov SURVIVAL=surv / NOMEAN METHOD=CH;
```

The output from the procedure and the first 10 records of the output data set `ciData` are printed in Appendix B1 and B2 p. 9 and 10, respectively.

The macro `CumInc` can now be used. The syntax is:

```
%INCLUDE 'CumInc.sas'; /* The file containing the macro is read */  
%CumInc(ciData,Strata,Time,Surv);
```

Here `ciData` is the name of the data set constructed by PHREG as described above and contains the variables `Strata`, `Time` and `Surv`, where `Strata` is the name of the stratum variable ( $h$ ), `Time` is the name of the failure time variable and `Surv` is the name of the exponential of the negative estimated cumulative cause-specific hazards.

The macro produces a data set named `data` containing the time variable `Time`, the estimated cumulative incidences `P01,P02,...,P0k` where  $k$  is the number of strata

and P01 is the estimated cumulative incidence corresponding to the first failure type with the smallest value of the stratum indicator  $h$ , P02 is the estimated cumulative incidence corresponding to the second failure type with the second smallest value of the stratum indicator  $h$ , etc. Furthermore, the data set contains the estimated overall survival probability  $P00=1-P01-\dots-P0k$  at every time point in Time.

The macro is now used on the malignant melanoma data set:

```
%INCLUDE 'CumInc.sas';
%CumInc(ciData,h,time,surv);

proc print data=data;
```

The output data set `data` will contain the four variables `time`, `P01`, `P02` and `P00`. The first 10 records of this data set are printed in Appendix B3 p. 10.

### 3. USING THE MACRO CumIncV.

The macro `CumIncV` also computes the estimated variances of the estimated cumulative incidence functions as described in Andersen et al. (1993) p. 512-515. This macro requires a lot of input, most of which is based on output from the SAS procedure `PROC PHREG`. If a model based on a single covariate `Covariat` is used, the procedure has to be defined in this way:

```
PROC PHREG DATA=StackDat OUTEST=EstData COVOUT;
MODEL Time*D(0)= Covariat;
STRATA Strata;
OUTPUT OUT=ResData XBETA=Xbeta;
BASELINE OUT=ciData COVARIATES=cov XBETA=Xbeta SURVIVAL=Surv
/ NOMEAN METHOD=CH;
```

`StackDat` is the name of the stacked data set created above, `Time` is the failure time variable, `D` is the failure indicator  $D$ , `Strata` is the stratum indicator ( $h$  as defined above) and `cov` is the name of the data set containing the fixed value of the covariates.

*Important: No statements must be omitted and the `NOMEAN METHOD=CH` statement must not be changed. No changes must be made to the output data set `EstData` whereas the output data sets `ResData` and `ciData` have to be changed as described below.*

The output data set `EstData` contains estimates of the regression coefficients and the estimated covariance matrix of the parameter estimates.

The output data set `ResData` contains estimates of the linear predictor, named `Xbeta`. The variables containing the covariates in this data set have to be renamed. These variables must have the names `COV1`, `COV2`, ..., `COVd`,  $d$  being the number of covariates used in the stratified Cox regression.

Finally the output data set named `ciData` contains the values of the fixed covariate values from the covariate data set `cov`, the linear predictors (named `Xbeta`) based on these covariate values and the exponential of the negative estimated cumulative cause-specific hazards for each combination of strata and covariates. The variables containing the fixed covariate values have to be renamed in agreement with the renaming in the data set `ResData` above.

In the example of survival with malignant melanoma the Cox regression analysis is performed using the code:

```
PROC PHREG data=melanom2 OUTEST=EstData COVOUT;
  MODEL time*d(0)= sex thick1;
  STRATA h;
  OUTPUT OUT=ResData XBETA=xbeta;
  BASELINE OUT=ciData COVARIATES=cov XBETA=xbeta SURVIVAL=surv
    / NOMEAN METHOD=CH;
```

The output of the procedure is the output printed in Appendix B1 p. 9.  
The data set `EstData` in Appendix C1 p. 11.

The names of the covariates in the data set `ResData` are changed

```
DATA ResData;
  SET ResData;
  COV1=sex;
  COV2=thick1;
  DROP sex thick1;
```

and the first 10 records of this data set are printed in Appendix C2 p. 11.  
Similarly, the names of the covariates are changed in the data set `ciData` by

```
DATA ciData;
  SET ciData;
  COV1=sex;
  COV2=thick1;
  DROP sex thick1;
```

and the first 10 records of this data set are printed in Appendix C3 p. 11.

The data are now ready for the use of the macro `CumIncV`. The syntax is:

```
%INCLUDE 'CumIncV.sas'; /* The file containing the macro is read */
%CumIncV(StackDat,ciData,Strata,Time,D,Surv,nCov,ResData,Xbeta,EstData);
```

`EstData`, `ResData` and `ciData` are the data sets constructed by the use of PROC PHREG above. Furthermore, `Strata` is the name of the stratum variable, `Time` is the name of the failure time variable, `D` is the name of the failure indicator (the variable `D` in the stacked data set), `Surv` is the name of the variable containing the estimated survival probabilities in `ciData`, `nCov` is the number of covariates included in the stratified Cox regression and `Xbeta` is the name of the linear predictor in the data sets `ResData` and `ciData`.

The macro produces a data set named `data` containing the time variable `Time`, the estimated cumulative incidences `P01`, `P02`, ..., `P0k` where  $k$  is the number of strata and `P01` is the estimated cumulative incidence corresponding to the first failure type with the smallest value of the stratum indicator  $h$ , `P02` is the estimated cumulative incidence corresponding to the second failure type with the second smallest value of the stratum indicator  $h$ , etc. Furthermore, the data set contains the estimated overall survival probability  $P00=1-P01-\dots-P0k$  at every time point in `Time`. The data set also contains the estimated variances `VAR01`, `VAR02`, ..., `VAR0k` of `P01`, `P02`, ..., `P0k` and the estimated variance `VAR00` of `P00` at every time point in `Time`.

The macro is now used on the malignant melanoma data set:

```
%INCLUDE 'CumIncV.sas';
%CumIncV(melanom2,ciData,h,time,d,surv,2,ResData,xbeta,EstData);
```

```
proc print data=data;
```

The output data set will contain the variables `time`, `P00`, `P01`, `P02`, `VAR00`, `VAR01` and `VAR02`. The first 10 records of this data set are printed in Appendix C4 p. 12.

The macro `CumIncV` may be very time consuming depending on the number of observations  $n$ .

It is recommended to include the statement `OPTIONS NONOTES;` before using the macro `CumIncV` to suppress the printing of notes in the `.log` file produced by SAS, since this file may become extremely large. Using the command `OPTIONS NOTES;` after the use of the macro will ensure that the notes are printed again.

## REFERENCES

Andersen, P. K., Borgan, Ø., Gill, R. D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer.

APPENDIX A: DATA

A1. Records no. 1-10 and 206-215 of the data set melanom2

OBS	ID	TIME	THICK	SEX	CAUSE	H	D	THICK1
1	1	10	3.84	1	2	1	0	3.84
2	2	30	-2.27	1	2	1	0	-2.27
3	3	35	-1.58	1	0	1	0	-1.58
4	4	99	-0.02	0	2	1	0	-0.02
5	5	185	9.16	1	1	1	1	9.16
6	6	204	1.92	1	1	1	1	1.92
7	7	210	2.24	1	1	1	1	2.24
8	8	232	9.96	1	1	1	1	9.96
9	9	232	0.30	0	2	1	0	0.30
10	10	279	4.49	0	1	1	1	4.49
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
206	1	10	3.84	1	2	2	1	0.00
207	2	30	-2.27	1	2	2	1	0.00
208	3	35	-1.58	1	0	2	0	0.00
209	4	99	-0.02	0	2	2	1	0.00
210	5	185	9.16	1	1	2	0	0.00
211	6	204	1.92	1	1	2	0	0.00
212	7	210	2.24	1	1	2	0	0.00
213	8	232	9.96	1	1	2	0	0.00
214	9	232	0.30	0	2	2	1	0.00
215	10	279	4.49	0	1	2	0	0.00

APPENDIX B: CumInc

B1. The output from the PHREG procedure

The PHREG Procedure

Data Set: WORK.MELANOM2  
 Dependent Variable: TIME  
 Censoring Variable: D  
 Censoring Value(s): 0  
 Ties Handling: BRESLOW

Summary of the Number of Event and Censored Values

Stratum	H	Total	Event	Censored	Percent Censored
1	1	205	57	148	72.20
2	2	205	14	191	93.17
-----					
Total		410	71	339	82.68

Testing Global Null Hypothesis: BETA=0

Criterion	Without Covariates	With Covariates	Model Chi-Square
-2 LOG L Score	700.985	675.808	25.177 with 2 DF (p=0.0001)
Wald	.	.	33.572 with 2 DF (p=0.0001)
	.	.	30.100 with 2 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
SEX	1	0.585164	0.23774	6.05837	0.0138
THICK1	1	0.159075	0.03271	23.65440	0.0001

Analysis of Maximum Likelihood Estimates

Variable	Risk Ratio
SEX	1.795
THICK1	1.172

**B2. The first 10 records of the output data set ciData**

OBS	SEX	THICK1	H	TIME	SURV
1	1	2	1	0	1.00000
2	1	2	1	185	0.99221
3	1	2	1	204	0.98430
4	1	2	1	210	0.97638
5	1	2	1	232	0.96846
6	1	2	1	279	0.96035
7	1	2	1	295	0.95224
8	1	2	1	386	0.94416
9	1	2	1	426	0.93610
10	1	2	1	469	0.92805

**B3. The first 10 records of the data set data constructed by CumInc**

OBS	TIME	P01	P02	P00
1	0	0.00000	0.000000	1.00000
2	10	0.00000	0.006703	0.99330
3	30	0.00000	0.013406	0.98659
4	99	0.00000	0.020155	0.97984
5	185	0.00766	0.020155	0.97219
6	204	0.01545	0.020155	0.96440
7	210	0.02323	0.020155	0.95661
8	232	0.03102	0.026862	0.94211
9	279	0.03895	0.026862	0.93419
10	295	0.04687	0.026862	0.92627

APPENDIX C: CumIncV

**C1. The output data set EstData**

OBS	_TIES_	_TYPE_	_NAME_	SEX	THICK1	_LNLIKE_
1	BRESLOW	PARMS	ESTIMATE	0.58516	0.15908	-337.904
2	BRESLOW	COV	SEX	0.05652	-0.00012	-337.904
3	BRESLOW	COV	THICK1	-0.00012	0.00107	-337.904

**C2. The first 10 records of the output data set ResData containing the renamed covariates**

OBS	H	TIME	D	COV1	COV2	XBETA
1	1	10	0	1	3.84	1.19601
2	1	30	0	1	-2.27	0.22406
3	1	35	0	1	-1.58	0.33382
4	1	99	0	0	-0.02	-0.00318
5	1	185	1	1	9.16	2.04229
6	1	204	1	1	1.92	0.89059
7	1	210	1	1	2.24	0.94149
8	1	232	1	1	9.96	2.16955
9	1	232	0	0	0.30	0.04772
10	1	279	1	0	4.49	0.71425

Note: The output data set ResData from the PHREG procedure is sorted in decreasing order w.r.t. the time variable Time within each strata. Thus, the data set was sorted w.r.t. increasing order of the time variable Time within each strata before these 10 observations were picked out.

**C3. The first 10 records of the output data set ciData containing the renamed covariates**

OBS	H	TIME	XBETA	SURV	COV1	COV2
1	1	0	0.90331	1.00000	1	2
2	1	185	0.90331	0.99221	1	2
3	1	204	0.90331	0.98430	1	2
4	1	210	0.90331	0.97638	1	2
5	1	232	0.90331	0.96846	1	2
6	1	279	0.90331	0.96035	1	2
7	1	295	0.90331	0.95224	1	2
8	1	386	0.90331	0.94416	1	2
9	1	426	0.90331	0.93610	1	2
10	1	469	0.90331	0.92805	1	2

**C4. The first 10 records of the of the data set data constructed by CumIncV**

OBS	TIME	P01	P02	P00	VAR01	VAR02	VAR00
1	0	0.00000	0.000000	1.00000	.0000000	.00000000	.0000000
2	10	0.00000	0.006703	0.99330	.0000000	.00004489	.0000449
3	30	0.00000	0.013406	0.98659	.0000000	.00009032	.0000903
4	99	0.00000	0.020155	0.97984	.0000000	.00013691	.0001369
5	185	0.00766	0.020155	0.97219	.0000584	.00013691	.0001967
6	204	0.01545	0.020155	0.96440	.0001191	.00013691	.0002589
7	210	0.02323	0.020155	0.95661	.0001801	.00013691	.0003215
8	232	0.03102	0.026862	0.94211	.0002416	.00018342	.0004318
9	279	0.03895	0.026862	0.93419	.0003055	.00018342	.0004979
10	295	0.04687	0.026862	0.92627	.0003695	.00018342	.0005639