Paper AS05

Eat, Sleep, R, Repeat

Miriam Amor, Veramed, Madrid, Spain

ABSTRACT

Not only are statistics in clinical research evolving in an exciting way, but also the open software R is gaining popularity for study submissions in clinical trials. This combination opens a new area of professional growth for statisticians and programmers.

Over the last decades, there has been a growing interest in analysing recurrent and repeated occurrences of an event. With time-to-event analysis being one of the most common analyses in clinical trials, it should come as no surprise that survival analysis techniques are evolving to provide robust estimators for time-to-event in the presence of recurrent events.

This paper aims to show how to use R and SAS® to program different models for time-to-event with recurrent events. These models will include the conditional approaches proposed by Anderson-Gill (AG) [1] and Prentice, Williams and Peterson [2], in addition to the marginal model described by Wei, Lin and Weissfeld [3].

INTRODUCTION

Although SAS® has long been the preferred software for statistical analysis and reporting in the pharmaceutical industry, the open-source R software is gaining in popularity and being used in study submissions. As initiatives such as “ADaM in R Asset Library” (ADMIRAL) package [4] develop open-source code and gain traction with pharmaceutical companies and institutions comfortable in using open-source code, we, as data analysts, should add programming in R to our personal toolkit, and be able to replicate our results in either software.

At the same time, there is a growing interest in analysing recurrent and repeated occurrences of an event, as well as the time to the occurrence of the events. Typically, Mixed Models for Repeated Measures (MMRM) and Generalized Linear Mixed Models (GLMMs) are common models used in longitudinal studies, and the Cox Regression Model is used to analyse the risk of having an event over a unit of time. However, when the outcome of interest is time to event of a recurrent event, these models are not suitable as the Cox Regression Model is limited to a unique occurrence of an event, and MMRM and GLMMs would not provide information about the time to the occurrence.

Different approaches are published to resolve this problem by Anderson-Gill (AG) [1] and Prentice, Williams and Peterson [2] and Wei, Lin and Weissfeld [3]. This paper aims to illustrate, with different examples, how to program these analyses in both R and SAS®.

MODELS FOR TIME TO EVENT WITH RECURRENT EVENTS: AN OVERVIEW

The first method presented is by **Anderson-Gill (AG) [1],** which is also called the “*proportional intensity model*”. The idea behind this model is that a subject can experience an event several times during their observation period. An individual is therefore at risk of an event regardless of whether a previous event occurred or not, and the order of the occurrence of events is not important. An example would be to analyse time to hospitalization due to all causes in the elderly [5]. However, in practice the assumption of total independence within subjects is very rare, and therefore it is recommendable to apply the ‘sandwich correction’ to adjust for the within-subject correlation.

The second methodology explained corresponds to the conditional models described by **Prentice, Williams and Peterson (PWP) [2].** In contrast to AG, these models analyse events by stratification assuming the subject level of risk is conditioned by the previous event, meaning a subject cannot be at risk until the end of the previous event. For example, if the event of interest is the occurrence of heart attacks: as the first heart attack occurs to a subjects, chances of a second heart attack happening increase because during the first heart attack part of the heart becomes damaged. Depending on the outcome of interest, PWP suggested two different models:

- PWP-CP (Counting Process, also called Total Time): This model assesses the effect of treatment from the beginning of the observation period.
- PWP-GT (Gap Time): Assess effect of treatment from previous events. Gaps between events are often used with infrequent events when the interest lies in predicting the next event [5].

PWP models assume that the subjects can only be considered at risk for the next event after they experienced the previous event. A limitation for the use of PWP models is that the risk sets for the later events get quite small, making the estimates unstable. To overcome this issue, it is common to restrict the data to the k^{th} occurrence of the event.

Depending on the outcome of interest, the interpretation of results will be different: while the Hazard Ratio (HR) obtained with PWP-CP model provides the rate risk of having a new event, while PWP-GT would indicate the rate risk of recurrence after a previous event.

The last method presented in this paper is the marginal method proposed by **Wei, Lin and Weissfeld WLW [3]**. This approach is different to the ones above as it considers each event is a separate process and there is no ordering among events within subject. For example, during a neonatal intensive care unit (NICU) stay, a baby is at risk of several events simultaneously such as infection, meningitis, diarrhea etc. This model assumes each subject is at risk of each event for the whole observation period.

In summary, the selection of the model to use would depend on the type of events, the importance of the order of the occurrence and the time to be analysed. Table 1 displays a summary of the differences and similarities between the models.

	AG	PWP-CP	PWP-GT	WLW
Type of Events	Same type	Same type	Same type	Different type
Order of Events is important	No	Yes	Yes	No
Outcome	Total Time	Total Time	Gap Time	Total Time
Hazard Ratio (HR) interpretation	Risk of any new recurrence	Risk of any new recurrence	Risk of new recurrence after a previous event	Risk of event of any type

Table 1. Comparison of different models for time to event with repeated measures

ANALYSIS OF RECURRENT EVENTS WITH SAS® AND R

Analyses of survival data based on the Cox regression model typically use the PHREG procedure in SAS, or the COXPH function in R. When it comes to analysing time to event data in presence of recurrent events, the same commands can be used to perform the analysis providing an adequate data structure is used and the corresponding options for each methodology are specified.

In the following sections, the examples show the data structure for the four methodologies described above, as well as the specific code with the different software SAS® and R. The data selected to illustrate these examples is the "bladder cancer data" listed in Wei, Lin, and Weissfeld (1989) [3], which has been widely used in literature and publications related to this topic.

The dataset is available in R in the *survival* package [6], where there are two different datasets containing the same data with different structures: bladder and bladder2. Both files contain data from 85 subjects (47 randomized into the placebo, and 38 randomized into the group receiving thiotepa) with bladder tumours, which were removed at the start of the study. The dataset contains the information on the first four recurrences of the tumour for each subject. A screenshot of some records taken from each dataset is shown in Figure 1.

Bladder

ID	RX	NUMBER	SIZE	START	STOP	EVENT	ENUM
1	1	1	3	0	1	0	1
1	1	1	3	0	1	0	2
1	1	1	3	0	1	0	3
1	1	1	3	0	1	0	4
5	1	4	1	0	6	1	1
5	1	4	1	0	10	0	2
5	1	4	1	0	10	0	3
5	1	4	1	0	10	0	4

Bladder2

ID	RX	NUMBER	SIZE	START	STOP	EVENT	ENUM
1	1	1	3	0	1	0	1
5	1	4	1	0	6	1	1
5	1	4	1	6	10	0	2
6	1	1	1	0	14	0	1
7	1	1	1	0	18	0	1
8	1	1	3	0	5	1	1
8	1	1	3	5	18	0	2

Figure 1. Screenshot of different data structures for bladder data.

The variables included in both files are listed below:

- ID: subject identification
- RX: treatment group (1: placebo and 2: thiotepa)
- Number: number of initial tumours
- Size: initial tumour size
- Start: Initial time, set to zero as it refers to start of the observation period.
- Stop: time of the occurrence of the event, or follow-up time if the event did not occur.
- Event: occurrence of the event (1: yes, 0: no).
- Enum: order of the recurrence (1,2,3,4)

These files collect the data as a counting process structure, that is: one record for each subject and time interval, where a time interval is defined as the time to the k^{th} recurrence, or time to follow-up if the event did not occur. The main difference between these two files is that the first one (bladder), contains one entry for each of the 4 possible tumour recurrences regardless of whether the subject had the event or not, while the second one (bladder2), contains one record for each occurrence of the event or time to follow-up. For example:

- Subject ID #1: was in the study for one month without an event. In the first dataset, this subject has 4 entries (one for each possible occurrence), the interval is stated by variables START=0 and STOP=1, EVENT is set to zero as the subject did not have any recurrence of the tumour and baseline/initial variables remain constant. In the second dataset there is one unique record for this subject.

- Subject ID #5: was in the study for ten months, experiencing a tumour recurrence at month 6. In the first dataset there are 4 records for this subject, with the first interval is set to STOP=6 and EVENT=1, because they had an event at month six, while remaining records contain the follow-up time. However, in dataset bladder2, there are two entries instead of 4, as the time in the study for this subject can be split in two separate intervals:
    - 1^{st} interval: from beginning of the study until month 6, where there was an event (EVENT=1).
    - 2^{nd} interval: from month 6 until month 10, without any other recurrence of the event (EVENT=0).

The correct data structure is key when working with recurrent events and depends on the methodology being used. The next section shows how to analyse the time to the recurrence of bladder tumours using the treatment (placebo vs thiotepa), number of initial tumours and initial tumour size as covariables, for each of the four methodologies described earlier.

ANDERSON-GILL (AG): WITH SAS® AND R

Using the AG model to analyse recurrences of an event requires the data to use a counting process structure. Table 2 shows the data structure used for the data, similar to the dataset bladder2 described above.

Note variables ID, START and STOP are now renamed to SUBJID, TSTART and TSTOP to avoid using SAS key words as variable names.

SUBJID	INTERVAL	TSTART	TSTOP	EVENT (*1: Yes, 0: No*)
1	(0,1]	0	1	0
*….]	*….]	*….]	*….]	*….]
5	(0,6]	0	6	1
5	(6,10]	6	10	0
*….]	*….]	*….]	*….]	*….]

Table 2. Data structure example for AG model

Using this data structure, the data can be analysed using the AG methodology by using the counting process syntax: adding variables TSTART, TSTOP instead of one variable for time, and adding the following options in order to obtain the robust sandwich estimators:

- SAS: COVS(aggregate) and ID <subject identification>
- R: CLUSTER <subject identification>

The codes used for this analysis are shown in Table 3. Since default options may differ between SAS and R (in SAS, Breslow is used for ties by default, whilst R uses the Efron approximation). In these examples, Breslow is used in both programs in order to obtain same results. In SAS the option rl displays the 95% confidence interval for the HR.

SAS

```
proc phreg data=Bladder2 covs(aggregate);
 class rx (ref='1');
 model (tstart, tstop) * event(0) = rx Size Number /rl;
 id subjid;
run;
```

R

```
AG <- coxph(Surv(tstart, tstop, event) ~  as.factor(rx)+size+number , ties="breslow",  cluster=subjid,
data=bladder2)
 summary(AG)
```

Table 3. Comparison of code in SAS® and R for AG model

Results obtained with these programs are printed in Figure 2. As the option COVS(aggregate) was included, the SAS estimation with robust Sandwich variance estimate is printed. The results in R include both standard error estimators, but the 95%CI obtained for HR is calculated using the Sandwich correction.

SAS

Analysis of Maximum Likelihood Estimates
with Sandwich Variance Estimate

Parameter		DF	Parameter Estimate	Standard Error	StdErr Ratio	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits		Label
rx	2	1	-0.45979	0.25801	1.290	3.1757	0.0747	0.631	0.381	1.047	rx 2
Size		1	-0.04256	0.07555	1.094	0.3174	0.5732	0.958	0.826	1.111	
Number		1	0.17165	0.06131	1.296	7.8373	0.0051	1.187	1.053	1.339	

R

```
                    coef exp(coef) se(coef) robust se      z Pr(>|z|)
as.factor(rx)2 -0.45979   0.63142  0.19996   0.25801 -1.782  0.07474 .
size           -0.04256   0.95833  0.06903   0.07555 -0.563  0.57317
number          0.17164   1.18726  0.04733   0.06131  2.799  0.00512 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

               exp(coef) exp(-coef) lower .95 upper .95
as.factor(rx)2    0.6314     1.5837    0.3808     1.047
size              0.9583     1.0435    0.8264     1.111
number            1.1873     0.8423    1.0528     1.339
```

Figure 2. AG model results obtained with SAS® and R

In this analysis, the HR estimates the risk of having any new tumour recurrence in the group of subjects treated with thiotepa versus the placebo group. As expected, results look the same regardless of the software used and result in a HR 95%CI 0.631 [0.381, 1.047] with a non-significant p-value: 0.0747.

PRENTICE, WILLIAMS AND PETERSON (PWP)

The first step is to choose the model: PWP-CP should be used if the outcome of interest is time since beginning of study, or PWP-GT if it is the time between events. In both cases dataset 'bladder2' can be used as the source data, but a modification would be needed if using the PWP-GT methodology.

Table 4 displays the data structure to be followed for the two PWP methodologies. While the data structure for PWP-CP is the same as in previous example for AG, the definition of TSTART and TSTOP is different in order to analyse the GAP time. The scenario presented is for SUBJID #5: now TSTART is set to zero, and TSTOP is the time since the previous event: 4 months. Additionally, a new variable to collect the gap time can be derived as GAP_TIME= TSTOP-TSTART, to be used in the PWP-GT model.

	SUBJID	TSTART	TSTOP	GAP_TIME	EVENT *1: Yes 0: No*	ENUM
PWP-CP	1	0	1	NA	0	1
	….]	*….]*	[….]	[….]	*….]*	*….]*
	5	0	6	NA	1	1
	5	6	10	NA	0	2
PWP-GT	1	0	1	1	0	1
	….]	*….]*	*….]*	*….]*	*….]*	*….]*
	5	0	6	6	1	1
	5	0	**4**	**4**	0	2

TSTOP = Time since beginning the study

TSTOP = Gap Time

Table 4. Data structure example for PWP models

With the correct data structure, the results can be easily obtained with SAS or R by using same code as for the AG model, but adding the STRATA and specifying the variable that identifies each occurrence of the event, or censorship if the event did not occur, in our dataset this is the variable ENUM. Table 5 shows example code. Note for PWP-GT analysis, the counting process structure using variables TSTART and TSTOP is no longer used, as these variables are replaced by the variable GAP_TIME in both software:

SAS

```
proc phreg data=Bladder2 covs(aggregate);
 class rx (ref='1');
 model (tstart, tstop) * event(0) = rx Size Number/rl; /* [1] PWP-CP*/
 model gap_time * event(0) = rx Size Number/rl; /* [2] PWP-GT*/
 id subjid;
 strata enum;
run;
```

R

```
PWP_CP <- coxph(Surv(tstart, tstop, event) ~ as.factor(rx)+size+number + strata(enum) , ties="breslow",
cluster=subjid, data=bladder2) # [1] PWP-CP
summary(PWP_CP)

PWP_GT <- coxph(Surv(gap_time, event) ~ as.factor(rx)+size+number + strata(enum) , ties="breslow",
cluster=subjid, data=bladder2) # [2] PWP-GT
summary(PWP_GT)
```

[1] Code used in PWP-CP analysis.

[2] Code used in PWP-GT analysis.

Table 5. Comparison of code in SAS® and R for PWP models

Results obtained for PWP-CP and PWP-GT models are displayed in Figure 3 and Figure 4 respectively. The HR obtained in the PWP-CP estimates the risk of any recurrence during the follow-up period, so the HR: 0.72 [95%CI: 0.49, 1.05] obtained for the treatment variable would indicate that the risk of recurrence is reduced by 28% in the group of subjects treated vs the placebo group. The HR obtained with PWP-GT estimates the risk of having a new tumour after previous recurrence. In PWP-GT, the HR: 0.76 [95%CI: 0.51, 1.49] obtained for treatment variable would indicate that treatment reduces the risk of having a new recurrence after the previous one. However, in neither of the models do we obtain a significant p-value for variable treatment (p-value: 0.089 in PWP-CP and p-value: 0.1952 in PWP-GT).

SAS

Analysis of Maximum Likelihood Estimates

with Sandwich Variance Estimate

Parameter		DF	Parameter Estimate	Standard Error	StdErr Ratio	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits		Label
rx	2	1	-0.33430	0.19706	0.912	2.8777	0.0898	0.716	0.486	1.053	rx 2
Size		1	-0.00805	0.06012	0.827	0.0179	0.8935	0.992	0.882	1.116	
Number		1	0.11565	0.04991	0.930	5.3690	0.0205	1.123	1.018	1.238	

R

```
                    coef exp(coef)  se(coef) robust se      z Pr(>|z|)
as.factor(rx)2 -0.334295  0.715842  0.216087  0.197064 -1.696   0.0898 .
size           -0.008051  0.991982  0.072725  0.060124 -0.134   0.8935
number          0.115653  1.122606  0.053681  0.049913  2.317   0.0205 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

               exp(coef) exp(-coef) lower .95 upper .95
as.factor(rx)2    0.7158     1.3970    0.4865     1.053
size              0.9920     1.0081    0.8817     1.116
number            1.1226     0.8908    1.0180     1.238
```

Figure 3. PWP-CP model results obtained with SAS® and R

SAS

Analysis of Maximum Likelihood Estimates

with Sandwich Variance Estimate

Parameter		DF	Parameter Estimate	Standard Error	StdErr Ratio	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits		Label
rx	2	1	-0.26952	0.20808	1.002	1.6778	0.1952	0.764	0.508	1.148	rx 2
Size		1	0.00684	0.06222	0.889	0.0121	0.9125	1.007	0.891	1.137	
Number		1	0.15353	0.04889	0.938	9.8620	0.0017	1.166	1.059	1.283	

R

```
                    coef exp(coef) se(coef) robust se      z Pr(>|z|)
as.factor(rx)2 -0.26952   0.76375  0.20766   0.20808 -1.295  0.19522
size            0.00684   1.00686  0.07001   0.06222  0.110  0.91246
number          0.15353   1.16595  0.05211   0.04889  3.140  0.00169 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

               exp(coef) exp(-coef) lower .95 upper .95
as.factor(rx)2    0.7637     1.3093    0.5080     1.148
size              1.0069     0.9932    0.8913     1.137
number            1.1659     0.8577    1.0594     1.283
```

Figure 4. PWP-GT model results obtained with SAS® and R

WEI, LIN AND WEISSFELD (WLW)

The last methodology presented in this paper is the marginal model described by WLW. In this case it is assumed there is not a specific order among these events or they are different type of events. Thus, for this specific analysis it is assumed that variable ENUM is no longer an ordinal variable collecting the order of each occurrence, but that it records the information related to the location of the tumour occurrence. Additionally, under this model it is implied that the occurrence of an event in one location does not increase/reduce the odds of having a tumour in a different location. To

make this example more visual, a format is assigned to the variable ENUM, so that each number (1-4) now refers to a different tumour site.

An example for the data structure including the new variable is shown in Table 6. The dataset now contains 4 entries for each subject, including the time from beginning of the study to the occurrence of each tumour location. If a subject does not experience an event in any location, the variable TSTOP is set to the last month in the study.
Demonstrating this scenario, Subject ID #5 had a new tumour in bone at month 6 but had no events in any other location.

SUBJID	INTERVAL	TSTART	TSTOP	EVENT	ENUM
1	(0,1]	0	1	0	1 (Bone)
1	(0,1]	0	1	0	2 (Lung)
1	(0,1]	0	1	0	3 (Colon)
1	(0,1]	0	1	0	4 (Liver)
[….]	*[….]*	*[….]*	*[….]*	*[….]*	*[….]*
5	(0,6]	0	6	1	1 (Bone)
5	(0,10]	0	10	0	2 (Lung)
5	(0,10]	0	10	0	3 (Colon)
5	(0,10]	0	10	0	4 (Liver)

TSTART = Set to zero

TSTOP = Time since beginning of the study.

ENUM = Different types of events (instead of order of the events)

Table 6. Data structure example for WLW models

Table 7 shows the code used for this analysis, which is the same as the one used for PWP-CP (see Table 5). However, the obtained results (Figure 5) are different to the ones obtained with the conditional model. Though the p-value is not statistically significant in any of the models, the result obtained in WLW is p-value:0.0560, while in PWP-CP it was p-value: 0.0898. This is because the data structure is key to obtaining the results from the selected methodology. In this analysis, the HR estimates the risk of having a recurrence in any location, so a HR: 0.56 [95%CI: 0.31, 1.02], would indicate that the risk of having any type of recurrence is lower in the treatment thiotepa group than in the placebo group.

SAS

```
proc phreg data=Bladder covs(aggregate);
 class rx (ref='1');
 model (tstart, tstop) * event(0) = rx Size Number/ rl;
 strata enum;
 id subjid;
run;
```

R

```
PWP <- coxph(Surv(tstart, tstop, event) ~ as.factor(rx)+size+number + strata(enum) , ties="breslow",
cluster=subjid, data=bladder)
summary(PWP)
```

Table 7. Comparison of code in SAS® and R for WLW model

SAS

Analysis of Maximum Likelihood Estimates with Sandwich Variance Estimate											
Parameter		DF	Parameter Estimate	Standard Error	StdErr Ratio	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits		Label
rx	2	1	-0.57984	0.30344	1.508	3.6516	0.0560	0.560	0.309	1.015	rx 2
Size		1	-0.05093	0.09304	1.335	0.2997	0.5841	0.950	0.792	1.140	
Number		1	0.20852	0.06568	1.400	10.0811	0.0015	1.232	1.083	1.401	

R

```
                  coef exp(coef) se(coef) robust se      z Pr(>|z|)
as.factor(rx)2 -0.57986   0.55998  0.20118   0.30344 -1.911   0.0560 .
size           -0.05094   0.95034  0.06967   0.09304 -0.548   0.5840
number          0.20849   1.23182  0.04691   0.06567  3.175   0.0015 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

               exp(coef) exp(-coef) lower .95 upper .95
as.factor(rx)2    0.5600     1.7858    0.3089     1.015
size              0.9503     1.0523    0.7919     1.140
number            1.2318     0.8118    1.0830     1.401
```

Figure 5. WLW model results obtained with SAS® and R

CONCLUSION

This paper has presented four different approaches to analyse time to event with recurrent events with SAS® and R. The first step is to select the correct model based on the importance of the order of the events, type of the events and outcome of interest. Secondly, having the correct data structure for the selected model is key and as important as using the correct options during the programming. As demonstrated, the same piece of code would lead to different results due to the data structure associated with each model. Finally, the same results are obtained regardless of working with SAS® or R, as long as the correct statements and default options are carefully applied.

REFERENCES

[1] Andersen PK, Gill RD (1982) Cox's Regression Model for Counting Processes: A Large Sample Study. Ann Stat 10: 1100-1120.

[2] Prentice RL, Williams BJ, Peterson AV (1981) On the regression analysis of multivariate failure time data. Biometrika 68: 373-379.

[3] Wei LJ, Lin DY, Weissfeld L (1989) Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions. J Am Stat Assoc 84: 1065-1073.

[4] https://CRAN.R-project.org/package=admiral

[5] Leila DAF Amorim , Jianwen Cai, Modelling recurrent events: a tutorial for analysis in epidemiology, *International Journal of Epidemiology*, Volume 44, Issue 1, February 2015, Pages 324–333, https://doi.org/10.1093/ije/dyu222

[6] https://CRAN.R-project.org/package=survival

[7] Lu J, Shen D (2018) Application of Survival Analysis in Multiple Events Using SAS. PharmaSug 2018 – Paper EP -02.

[8] Yadav CP, Sreenivas V, Khan MA, Pandey RM (2018) An Overview of Statistical Models for Recurrent Events Analysis: A Review. Epidemiology (Sunnyvale) 8: 354.

[9] SAS/STAT® 9.22 User's Guide. The PHREG Procedure (Book Excerpt)

ACKNOWLEDGMENTS

I would like to thank Leanne Johnstone for her careful and thorough review of this paper, and her valuable corrections and suggestions. Additionally, my thanks to Abby Fuller for her review, feedback, and recommendations from a statistical point of view.

CONTACT INFORMATION

Miriam Amor
Veramed
Registered Office: 5th Floor Regal House, 70 London Road, Twickenham TW1 3QS,
Email: Miriam.Amor@veramed.co.uk
Web: www.veramed.co.uk