PERSONYRS: A SAS© PROCEDURE FOR PERSON YEAR ANALYSES

BY

Erik J. Bergstralh, Kenneth P. Offord,
Jon L. Kosanke and Glenn A. Augustine

Technical Report Series, No. 31
April 1986

PERSONYRS: A SAS® PROCEDURE FOR PERSON YEAR ANALYSES

Erik J. Bergstralh, Kenneth P. Offord, Jon L. Kosanke, Glenn A. Augustine
Mayo Clinic

Introduction

In epidemiologic studies, one often wishes to estimate rates of certain events in a cohort of individuals observed over a period of time. For example, what is the rate of cancer following the diagnosis of rheumatoid arthritis and is it higher than expected? A common method of estimating such rates is to divide the total number of observed events by the total person years (sum of the individual follow-up times in years) at risk. It is also often useful to look at event rates based on person years classified by sex, age, calendar year and follow-up year. Table 1 below contains an example of how an individual's follow-up time can be broken down by age, calendar year and follow-up year.

PROC PERSONYRS is a SAS procedure which can calculate sex-, age-, calendar year- and follow-up year-specific person years, events and rates. The expected number of events can also be calculated by applying user-supplied event rates to sex-, age- and calendar year-specific person years. In addition, rates adjusted for age or age and sex can be calculated. This is accomplished by having the user supply the number of persons in sex and age categories for the population to which the rates are to be adjusted.

Specifications

The statements used to control PERSONYRS are:

```
PROC PERSONYRS options,
   ID identifying variable,
   VARIABLES [keyword=variable(s)],
   TABLES requests/options,
   AGEYRINT specify age intervals,
   CALYRINT specify calendar year intervals;
   FUYRINT  specify follow-up year
            intervals,
   BY variables,
```

The PROC, VARIABLES and TABLES statements are required.

PROC PERSONYRS statement

PROC PERSONYRS options;

The options available are:

DATA=dataset
names the SAS data set containing the data to be analyzed. If DATA= is omitted, the most recently created SAS data set is used.

NOPRINT supresses all printed output.

RATEMULT=integer
integer is the constant multiplier for calculated rates. Rates will be printed as events per "integer" person years. The default value is 100,000.

MALES / FEMALES
indicates that input data set contains only MALES or only FEMALES. If omitted, it is assumed that the data set is not restricted to a single sex.

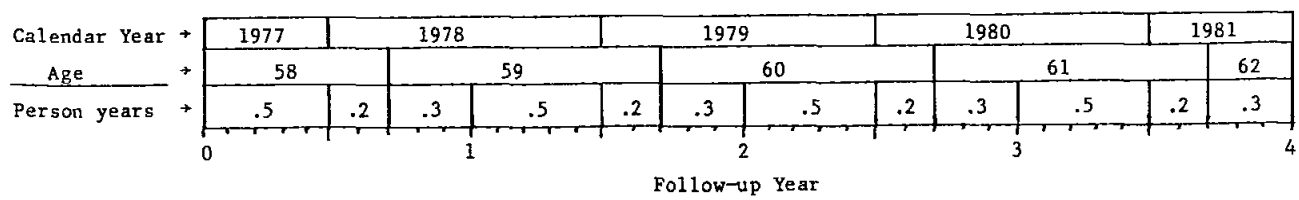
RATESDATA=dataset
names the SAS data set containing expected event rates In this data set, there must be one observation for every age interval for every calendar period. The data set must be sorted by calendar period and age. Required variables are:

CALYRB 4-digit integer numeric variable containing the first calendar year (19xx) to which the rate applies. The unique values of CALYRB must be identical to those defined below in the CALYRINT statement. Set this variable to missing if there are no calendar time restrictions, i.e., this age-specific rate applies to all calendar periods.

AGEB integer numeric variable containing the first year of age to which the rate applies. These values must be identical to those defined below in the AGEYRINT statement.

Table 1

A person begins follow-up at age 58.3 halfway through 1977 (1977.5) and is under observation through 1981.5 for a total of 4.0 person years of follow-up. The person years can be subdivided as follows:



MRATE
FRATE
these two numeric variables contain the expected annual event rate per RATEMULT for males (_MRATE_) and females (_FRATE_). The opposite sex variable is not needed if one of the only MALES or only FEMALES options is used.

RATESLABEL='character string'
character string is a description of the data set named in RATESDATA which will be included in the PERSONYRS output. A maximum of 40 characters is allowed.

TOFIRSTEVT / TOLASTFU
TOFIRSTEVT requests that each individual's person years be calculated from the ZERODT (see VARNAMES statement) to the date of the first event (if any events) or to the date of last follow-up (if no events). TOLASTFU requests that each individual's person years be calculated from the ZERODT to the date of last follow-up with the number of events to be determined by the DAYSEVT variables as defined below in the VARNAMES statement. Only one of these two options is permitted.

ADJPOP=dataset
provides the name of a SAS data set containing population frequencies by sex and age interval. This option is required when rates are to be age-adjusted or age- and sex-adjusted to an external population (see the TABLES statement options below for further information). This dataset must be sorted by age and must include the following variables.

AGEB numeric variable defining the beginning integer age to which frequencies apply. The values of _AGEB_ must be identical to those defined in the AGEYRINT statement.

MCOUNT
FCOUNT variables containing the population counts for males and females, respectively. The opposite sex variable not needed if one of the only MALES or only FEMALES options is used.

ADJLABEL='character string'
character string is a description of the data set named in ADJPOP which will be included in the PERSONYRS output. A maximum of 40 characters is allowed.

ID statement

ID variable ;
This statement is used to request that both the ID variable specified and the VARNAMES variables (see below) be printed in the log for observations which are deleted due to missing values or bad data.

VARNAMES statement

VARNAMES [keyword=key_word_var_name(s)];
The choices for keywords are listed below. If not specified, the variable appropriately defined with the same name as the keyword is assumed to exist on the DATA= data set referred to in the PROC statement.

SEX = variable_name_for_sex This character variable contains the code for sex. It must have a length of one and denoted M (male) or F (female). This variable is not required if the MALES/FEMALES only option of the PROC statement is used.

ZERODT = variable_name_for_date_to_begin_person_years This numeric variable is the SAS date indicating the beginning time for person years calculations.

AGEYRZ = variable_name_for_age_at_zero_date
This numeric variable contains the age in years at the ZERODT. Note that this variable is not restricted to integer ages.

DAYSEVT = days_to_event_variables
These numeric variables contain the number of days from the ZERODT to the events of interest. If one wishes to include all events for every person the number of "DAYSEVT" variables needed should equal the maximum number of events occurring to any one person. The "DAYSEVT" variables should contain numeric missing values whenever there isn't a corresponding event. This is because the number of events per person is calculated as the number of non-missing "DAYSEVT" variables. The "DAYSEVT" variables do not have to be ordered on time to events.

DAYSLFU = days_to_last_follow-up_variable This numeric variable contains the number of days from the ZERODT to the date of last follow-up. The value of DAYSLFU must be > the maximum value of the "DAYSEVT" variables.

TABLES statement

TABLES requests / options ;

The TABLES statement indicates which combinations of sex-, age-, calendar year-, and follow-up year-specific person years (also events, rates, expected rates, etc.) are to be tabled. (The cell entry options for the tables are defined below.) Requests are made in a fashion similar to PROC FREQ and are described below:

Requests The breakdown of person years, etc., is indicated by specifying one or more of the following fixed factors joined by asterisks:

Factor	Description
SEX	person years, etc., tabled separately for men and women.
AGEYR	person years, etc., tabled separately for each interval of age specified in the AGEYRINT statement as described below.

CALYR person years, etc. tabled separately for each interval of calendar years specified in the CALYRINT statement as described below.

FUYR person years, etc. tabled separately for each interval of follow-up years specified in the FUYRINT statement as described below.

For factors not joined by an asterisk, a one-way table is generated for each label name. Two-way cross-tabulations (e.g. age-and calendar year-specific person years) are generated by two factors joined with an asterisk. Three-and four-way tables are also permitted. For two or more labels joined by an asterisk, the last factor forms the columns of the table and the next-to-last factor the rows. A separate table is produced for each level (or combination of levels) of the other labels.

Each table also contains an analysis pooling rows and pooling columns. No factors other than those listed are permitted. Note that multi-way tables in combination with short intervals can produce a large amount of printed output with many sparse cells.

Any number of requests may be given on one TABLES statement and any number of TABLES statements are permitted.

Options

When no options are included, the TABLES statement in PROC PERSONYRS produces tables with cells that include the number of person years, the number of observed events and the rate per RATEMULT person years (i.e., (events/person years)xRATEMULT). The table cells referred to include both row and column totals. The options below may be used in the TABLES statement after the slash (/):

P requests that the number of persons contributing person years to a particular table cell be printed.

PY requests that the number of person years in a particular table cell be printed.

O requests that the number of observed events in a particular table cell be printed.

R requests that the event rate per RATEMULT person years be printed for each cell. R=(O/PY)*RATEMULT.

SER requests that the estimated standard error of the rate (R) be printed for each cell. Assuming that the events follow a Poisson distribution, then SER=R/$\sqrt{O}$.

R95 requests that a 95% confidence interval for the true R be printed for each cell. This is an exact confidence interval based on the assumption that the observed number of events has a Poisson distribution and the number of person years is fixed. The interval is calculated as follows:

Lower limit = $[X_L/(2*PY)]$*RATEMULT, where X_L is defined so that Pr(chi-square, d.f.=k<X_L)=.025 and k=2*(observed number of events).

Upper limit = $[X_U/(2*PY)]$*RATEMULT, where X_U is defined so that Pr(chi-square, d.f.=k+2>X_U)=.025.

OUT=dataset
sets up an output data set corresponding to the last table requested on the TABLES statement. This data set contains one observation for each cell produced by the table request. Each observation contains the cell identifier variables, plus a variable for each of the TABLES statement cell options requested. These variables have the same names as the option names with underscores prefixed (e.g. _P, _PY, _O, _E, _R, etc.).

The following cell options require that the "RATESDATA=" option of the PROC statement be in effect.

E requests that the expected number of events per cell be printed. For each cell, E = sum over sex, age and calendar year of "sex-age-calendar year-specific person years multiplied by the corresponding expected rates" (as contained in the RATESDATA data set) divided by RATEMULT.

RR requests that the ratio of observed (O) to expected (E) events be printed. This ratio is often referred to as the rate ratio (RR).

RR95 requests that a 95% confidence interval for the underlying population RR be printed. This is an exact confidence interval based on the assumption that the observed number of events has a Poisson distribution and the expected number of events is fixed. This interval is calculated as follows.

Lower limit = $X_L/(2*E)$, where X_L is as defined above.

Upper limit = $X_U/(2*E)$, where X_U is as defined above.

The following options require that the "ADJPOP=" option of the PROC statement be in effect:

AAR requests that an age-adjusted event rate per RATEMULT person years be printed for all table cells that are not age-specific. For each cell, AAR = sum over age intervals (i) of "W_i*R_i", where the W_i are the proportion of subjects (males and females combined) in each age group of the ADJPOP data set.

SEAAR requests that the estimated standard error of the age-adjusted rate (AAR) be printed for all table cells that are not age-specific.
SEAAR=[sum over age intervals (i) of "$(W_i^2 * R_i^2 / O_i)$"]$^{1/2}$.

AAR95 requests that a 95% confidence interval for the age-adjusted rate be printed for all table cells that are not age-specific. AAR95 = AAR ± 1.96*SEAAR. Lower 95% limits less than zero are printed as zero with an asterisk; i.e., "0*".

SAAR requests that a sex- and age-adjusted event rate per RATEMULT person years be printed for all table cells that are not age-or sex-specific. For each cell, SAAR = sum over age intervals(i) and sexes(j) of "$W_{ij} * R_{ij}$".
The adjusting fractions (W_{ij}, i for age interval, j for sex) are the proportion of subjects in each age-sex group of the total number in the ADJPOP data set. (Note that the sum of the W_{ij} over age and sex is 1.)

SESAAR requests that the standard error of the sex- and age-adjusted rate be printed for all table cells that are not sex- or age-specific.
SESAAR = [sum over age intervals(i) and sexes(j) of "$(W_{ij}^2 * R_{ij}^2 / O_{ij})$"]$^{1/2}$.

SAAR95 requests that a 95% confidence interval for the sex- and age-adjusted rate be printed for all table cells that are not sex- or age-specific. SAAR95 = SAAR ± 1.96*SESAAR. Lower 95% limits less than zero are printed as zero with an asterisk; i.e., "0*".

Interval Statements --- defining age, calendar year and follow-up year intervals

The three interval statements (to be described separately below) allow one to group the person years into the desired age-, calendar year- or follow-up year intervals. For any interval statement, the user must specify the beginning value (b_i) of each interval. These beginning values are used internally to define the intervals. The intervals are closed on the left and open on the right, i.e., $b_i \leq x < b_{i+1}$. For example, providing the 5 ages of "20 30 40 60 85" in the AGEYRINT statement would result in the following 5 intervals.

Interval	Ranges	Label
1)	20≤age<30	20-29
2)	30≤age<40	30-39
3)	40≤age<60	40-59
4)	60≤age<85	60-84
5)	age ≥85	85+

The above intervals could also be produced by specifying "20 TO 40 BY 10 60 85" in the AGEYRINT statement. If the data generate person years below the lowest interval specified, an error message will be printed. If the user has not specified the beginning interval values (for age and calendar period) to be identical to both those defined in the expected rates data set (RATESDATA= option) and in the adjusting population data set (ADJPOP= option), processing will stop and an error message will be printed.

AGEYRINT statement --- defining age intervals

AGEYRINT beginning_of_age_intervals
beginning_of_first_age_interval
TO beginning_of_last_age_interval BY
increment;

This statement is used to define desired age intervals. All values must be positive integers with first_age < last_age. This statement is required whenever one or more of the following are true:

1) age-specific person years are requested in the TABLES statement
2) the RATESDATA= option of the PROC statement is used
3) the ADJPOP= option of the PROC statement is used.

CALYRINT statement -- defining calendar year intervals

CALYRINT beginning_of_calendar_year_intervals
beginning_of_first_calendar_year
interval TO
beginning_of_last_calendar_year interval
BY increment;

This statement is used to define desired calendar year intervals. All beginning values must be positive 4-digit integers (e.g. 1980) with 'first calendar year' < 'last calendar year'. The values of increment must be a positive integer. This statement is required whenever one or both of the following are true:

1) calendar year-specific person years are requested in the TABLES statement
2) the RATESDATA= option of the PROC statement is used and age-sex rates differ with calendar period.

FUYRINT statement --- defining follow-up year intervals

FUYRINT beginning_of_follow-up_year_intervals
beginning_of_first_follow-up_year
interval TO beginning_of_last_follow-up
year interval BY increment;

This statement is used to define desired follow-up year intervals. All values must be non-negative numbers (decimal values are permitted) with the 'first follow-up year' < 'last follow-up year'. The first beginning value

specified should always be 0. This statement is required whenever follow-up year-specific person years are requested in the TABLES statement.

BY statement

BY variables;

A BY statement may be used with PROC PERSONYRS to obtain separate analyses on observations in groups defined by the BY variables. When a BY statement appears, the procedure expects the input data to be sorted in order of the BY variables. In addition, any BY variables must also be included in the RATESDATA and ADJPOP data sets. The values of the BY variables in these data sets are immaterial as they are needed only to circumvent a bug in the SAS system.

Details

Missing Values and Bad Data

Any observation in the input data set with missing values for any of the variables used in the VARNAMES statement (with the exception of the "DAYSEVT" variables) will be deleted. Any observation with the value of the DAYSLFU variable less than any of the "DAYSEVT" variable values will be deleted. Use the ID statement if you wish to have the observations printed in the log. Any missing values in the RATESDATA or ADJPOP data sets for the required variables will cause the observations to be printed and processing will be terminated.

Example

The following example is taken from a study by Katusic et al., 1985, and illustrates a common use of PROC PERSONYRS. The data consists of 521 incidence cases of rheumatoid arthritis (RA) diagnosed in Rochester, MN from 1950-74 inclusive. The primary question of interest was whether patients with RA have a higher risk of developing malignancy than does the general population. To answer this question we compared the total number of malignancies subsequent to the date of RA diagnosis to that expected, using age- and sex-specific incidence rates of all malignancies available for Rochester, MN for the period 1974-77. We also examined the data by follow-up year as sometimes the effect of a risk factor (i.e. RA) takes several years to surface.

i) Data set: RAPTS is the dataset of RA patients. For brevity, only the first 5 observations (one obs per patient) are included.

Variables:

RA_DX = date of RA diagnosis

AGE_RA = age at RA diagnosis

DAYS_CA1 = days from RA_DX to cancer diagnosis, missing if no ca.,

DAYS_CA2 = some pts. had 2 subsequent cancers

DAYS_LFU = days from RA_DX to date of last follow-up

Data:

CASE	SEX	RA_DX	AGE_RA	DAYS_CA1	DAYS_CA2	DAYS_LFU
1	F	09/29/65	53.7	.	.	6,040
2	F	12/19/50	39.8	8861	.	12,102
3	M	08/24/54	85.8	.	.	138
4	F	05/02/56	58.9	8225	7588	10,012
5	M	10/21/60	75.2	.	.	5,575
:						

ii) Rates Dataset: R is the dataset containing age-and sex-specific expected malignancy rates (per 100,000 per year)

Variables.
see RATESDATA option or PERSONYRS statement for definitions.

Data:

AGEB	MRATE	FRATE
0	10.1	25.4
20	27.1	42.3
30	61.7	163.7
40	209.8	219.1
50	527.3	720.7
60	1348.5	906.8
70	2058.0	1380.9
80	3606.6	2058.1

iii) SAS code to request person year analyses by follow-up year and sex

```
PROC PERSONYRS  DATA=RAPTS
                RATESDATA=R
                RATESLABEL=
                   'TOTAL MALIGNANCY RATES'
                TOLASTFU;
  VARNAMES  SEX=SEX
            ZERODT=RA_DX
            AGEYRZ=AGE_RA
            DAYSEVT=DAYS_CA1  DAYS_CA2
            DAYSLFU=DAYS_LFU,
    AGEYRINT  0 20 TO 80 BY 10;
    FUYRINT  0  4  8  12;
    TABLES  FUYR*SEX /  PY O E RR RR95;
```

iv) Output

Numbered items below are circled on the output.

1. List of PROC PERSONYRS options used
2. List of variable definitions from the VARNAMES statement
3. Distribution of number of events per person
4. Explanation of table cell entries
5. Person year analysis by sex and follow-up year with totals

PROC PERSONYRS EXAMPLE--CANCER FOLLOWING RHEUMATOID ARTHRITIS

PERSONYRS: PERSON YEARS ANALYSES FOR COHORT STUDIES

(1) PROC OPTIONS USED:

```
DATA=WORK.RAPTS
RATESDATA=WORK.R
RATESLABEL=TOTAL MALIGNANCY RATES
TOLASTFU
```

(2) VARIABLE DEFINITIONS(KEYWORD=VARIABLE NAME):

```
SEX=SEX_
ZEROUT=RA_DX
AGEYRZ=AGE_RA
DAYSLFU=DAYS_LFU
DAYSEVT=DAYS_CA1 DAYS_CA2
```

(3) OBSERVATIONS:

```
NUMBER USED=        521
NUMBER DELETED=       0
```

DISTRIBUTION OF NUMBER OF EVENTS:

EVENTS	MALES	FEMALES	TOTAL PERSONS
0	114	345	459
1	20	37	57
2	0	5	5
TOTAL PERSONS	134	387	521
TOTAL EVENTS	20	47	67

PERSONYRS: PERSON YEARS ANALYSES FOR COHORT STUDIES

TABLE CELLS FORMED BY COMBINATIONS OF FOLLOWUP YEAR AND SEX

(4)

```
CELL ENTRIES:
        PY      = NO. PERSON YEARS
        O       = NO. OBSERVED EVENTS

        E       = EXPECTED NO. EVENTS
        RR      = RISK RATIO
        RR95L   = LOWER 95% CL OF RISK RATIO
        RR95U   = UPPER 95% CL OF RISK RATIO
```

(5)

FOLLOWUP YEAR		SEX: MALE	FEMALE	TOTAL
0 TO <4	PY	498.166	1442.828	1940.994
	O	2	12	14
	E	4.520	10.246	14.765
	RR	0.442	1.171	0.948
	RR95L	0.054	0.605	0.518
	RR95U	1.593	2.045	1.592
4 TO <8	PY	443.417	1257.952	1701.369
	O	5	6	11
	E	4.732	9.863	14.595
	RR	1.057	0.608	0.754
	RR95L	0.343	0.223	0.377
	RR95U	2.462	1.323	1.350
8 TO <12	PY	349.243	1043.956	1393.199
	O	5	5	10
	E	4.148	8.795	12.942
	RR	1.205	0.569	0.773
	RR95L	0.392	0.185	0.370
	RR95U	2.809	1.325	1.422
12 +	PY	512.868	1839.945	2352.813
	O	8	24	32
	E	7.524	18.459	25.982
	RR	1.063	1.300	1.232
	RR95L	0.459	0.833	0.842
	RR95U	2.093	1.935	1.739
TOTAL	PY	1803.693	5584.682	7388.375
	O	20	47	67
	E	20.923	47.362	68.286
	RR	0.956	0.992	0.981
	RR95L	0.584	0.729	0.760
	RR95U	1.476	1.320	1.246

v) Comment - We see that the observed total number of malignancies was approximately equal to the expected number for both sexes. We also found that the risk of cancer was not changing significantly with follow-up year.

Correspondence

Inquiries may be directed to Erik Bergstralh, Medical Research Statistics, Mayo Clinic, Rochester, MN 55905.

Acknowledgements

We would like to thank Dr. Joe Melton for his suggestions in the development of PROC PERSONYRS. We also thank Ms. Marilyn Nelson for skillfully typing this paper.

Reference

Katusic, S., Beard, C.M., Kurland, L.T., Weis, J.W., Bergstralh, E.J.: Occurrence of Malignant Neoplasms in the Rochester, Minnesota Rheumatoid Arthritis Cohort. The Am. J. of Med., 78(1A), 1985.