International Journal of Epidemiology
© International Epidemiological Association 1997

Vol. 26, No. 2
Printed in Great Britain

A Simple Program to Create Exact Person-Time Data in Cohort Analyses

JOY WOOD, DAVID RICHARDSON AND STEVE WING

Wood J (Department of Epidemiology, School of Public Health, University of North Carolina at Chapel Hill, NC 27599-8050 USA), Richardson D and Wing S. A simple program to create exact person-time data in cohort analyses. *International Journal of Epidemiology* 1997; **26:** 395–399.

Background. Before disease rates can be calculated a tabulation of the length of follow-up for each person in the cohort has to be made. In complicated analyses such tabulations are often stratified by many characteristics, some which show no change with time, such as gender or year of birth, and some which do change with time, such as age or cumulative exposure. Available computer programs often restrict the way these tables can be made, particularly when handling time-dependent variables.

Methods. This paper presents a simple computer program which calculates the length of follow-up for each person in a study.

Results. Person-time data can be tabulated by a large number of variables using this method. This program is extremely flexible in the way that time-dependent variables can be created, can categorize observations by any unit of person-time, and will run on a range of platforms including a personal computer.

Conclusions. This method should simplify the task of creating person-time data for analyses of disease rates in epidemiological studies.

Keywords: occupational epidemiology, computing prospective studies, person-time computer program, rates

Cohort studies, unlike other epidemiological studies, allow consideration of disease rates. The calculation of rates can be used for simple comparisons of disease occurrence between groups, for example, when used for standardized rate ratios, as well as for more complicated analyses of disease rates with methods such as Poisson regression.[1]

In order to perform these analyses, data must be generated to describe the person-time of follow-up within the study cohort. Unfortunately, the available software for creation of person-time data is often complicated to understand and difficult to modify to suit the needs of researchers. Consequently, while analyses of disease rates are fundamental to cohort studies, the necessary creation of person-time data is often a sizeable obstacle.

The National Institute for Occupational Safety and Health's (NIOSH) Life Table program, for example, will only generate person-time data stratified on a limited number of variables, such as age and calendar time.[2] To consider other time-dependent variables significant changes have to be made to the program, and the boundaries for these new variables have to be considered in terms of age-at-follow-up. Furthermore, the program requires that each time unit of observation be assigned a level of exposure greater than zero; in many cases, this may simply not conform to the data. Finally, a researcher may be interested in complicated, time-dependent classifications of exposure, such as time-windows of exposure. However, NIOSH's Life Table program does not have the flexibility to allow these types of time-dependent exposure categorizations.

Pearce proposed a program to create person-year tables which resolved many of these problems.[3] Pearce's program creates a dataset in which each person-year of follow-up, rather than each person, is an observation. Each person-year of follow-up can be assigned characteristics such as the subject's gender, age, and cumulative exposure. This file is summarized to describe the distribution of person-years and events.

Pearce's program is extremely flexible; however, since an observation must be created for each unit of follow-up, the program is best suited to counting person-years rather than person-days. Consequently, the program, as Pearce proposed it, rounds the length of follow-up for each subject to the nearest whole number of person-years. Despite the utility of Pearce's program, it generated two criticisms: first, rounding years of follow-up is less preferable than counting person-days, since it could lead to systematic misclassification of

Department of Epidemiology, School of Public Health, CB # 8050, Nationsbank Plaza, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-8050, USA.

time-related variables; second, rounding years of follow-up could lead to deaths occurring in cells in which no person-years were counted.[4] Pearce has suggested that by carrying out any necessary rounding in the first year of observation, the latter problem may be avoided (personal correspondence).

In some circumstances, however, rounding person-years of follow-up may be a concern. A person who was followed for 5 years 182 days would contribute 5 person-years of observation, while a person followed for 5 years 183 days would contribute 6 person-years of follow-up. Overall, this rounding has little impact on the total number of person-years in a large cohort. However, using Pearce's program, each subject gets a full person-year for their last calendar year of follow-up; consequently, rounding leads to fewer person-years of observation in early calendar years of follow-up, and among people at younger ages. This could be a concern, for example, in analyses examining modification by age at exposure. Furthermore, since partial person-years cannot be counted, it is still possible to observe a death without a person-year if the subject was followed for less than a half year.

In a recent paper Macaluzo presented an alternative method for classifying person-time.[4] Macaluzo's method is best suited to the classification of time-dependent variables which change at a constant rate over time (for example, age or calendar year), and the efficiency of the program is greatest if all time factors are classified in categories of the same width (for example, 5-year intervals). However, in order to create person-time tables stratified by time-dependent variables that do not change at a constant rate over time (such as employment status or cumulative exposure) Macaluzo's program must separately count the person-time within each level of these strata. Such an approach quickly becomes complicated, and Macaluzo acknowledged that such cases require substantial programming effort.

The Epicure software package includes a program, DATAB, for the generation of person-time tables.[5] One important limitation of the DATAB program is that time-dependent variables which do not change at a constant rate over time can only be classified in terms of the midpoint of the strata for time-dependent factors which do change at a constant rate over time. While information is often available on the exact dates of employment, or on monthly exposure levels, associated time-dependent variables can be classified only as precisely as age or calendar time is categorized. In a table in which age was considered in 5-year intervals, for example, an indicator for date of last employment would only be evaluated at the midpoint of each 5-year interval.

METHODS

This paper presents a program which counts each day from the date of first observation to the date of last observation for each person in the cohort. After the last day of observation, a person's vital status is assessed; if the person has died a death is counted, classified by the same levels of the variables observed at the last day of observation.

Our method for making stratified person-time tables retains the strengths of Pearce's approach, but, by taking advantage of the increasing speed of personal computers, does not suffer the previously described limitations. In contrast to Pearce's program, which creates an observation for each person-year of follow-up, this program generates a temporary, multidimensional array in which follow-up is counted. Since this is a temporary array, a large number of cells can be considered in a quickly constructed table.[6] Within this array, an ongoing count is kept of the number of person-days of observation, classifying follow-up by variables of interest.

A large number of stratifying variables can be considered, and the classification of time-related variables remains flexible. This program can categorize follow-up time and events by any unit of person-time. However, since the input data for occupational cohort studies are generally recorded as dates of hire, birth, and last observation, counting individual person-days of follow-up often allows the most refined classification of person-time.

MATERIALS

This program requires only a simple data file in order to generate a person-time table. A dataset which includes the date of first observation, date of last observation, and an indicator of outcome status for each person is all that is needed for the simplest person-time tables.

Given complete ascertainment of outcome status, the date of last observation is either the end of follow-up, or the date at which an outcome occurs. In occupational cohorts, the date of first observation is often considered as some period of time (for example, one month) after the date of hire, since workers who have died shortly after hire are often excluded from analyses.[7]

Other variables, either fixed or time-dependent, can be included as needed. Since this program does not round follow-up time, the inclusion of additional time-dependent variables is simple—extra dimensions are added to the array, and the level of the time-dependent variable is calculated for each unit of observation. In order to assess age at risk, for example, the number of days from the subject's date of birth would be calculated at each unit of observation. In cohorts which

```
LIBNAME IN1  'C:\';
FILENAME OUT1 'C:\YYYY.RAW';

DATA PYWORK  (KEEP=PERYRS DTHS);
SET IN1.XXXX   END=EOF ;

**** DEFINE ARRAY ***;
ARRAY C(2) _TEMPORARY_;

**** GENERATE PERSON-DAYS ***;
DO DAY = HIREDATE TO DLO;
C(1)+1;
END;

**** GENERATE DEATHS  ***;
IF DEAD=1 THEN C(2)+1;

**** SUMMARIZE ARRAY ***;
IF EOF THEN DO;
  IF C(1) > 0 THEN DO;
          PERYRS=C(1)/365.25;
          DUM2=C(2); DTHS=SUM(DUM2,0);
  OUTPUT PYWORK;
  END; END;
RUN;

**** WRITE RAW FILE ***;
DATA T ;
SET PYWORK;
FILE OUT1;
PUT PERYRS DTHS;
RUN;
```

FIGURE 1 *Sample SAS program for calculation of data for crude rates*

```
OPEN "C:\XXX.DAT" FOR INPUT AS #1
DO WHILE NOT EOF(1)
INPUT #1, HIREDATE, DLO, DEAD

REM **DEFINE ARRAY**
DIM C(2)

REM **GENERATE PERSON-DAYS**
FOR I = HIREDATE TO DLO
LET C(1) = C(1) + 1
NEXT I

REM **GENERATE DEATHS**
IF DEAD = 1 THEN
        C(2) = C(2) + 1
END IF

REM **SUMMARIZE ARRAY**
LOOP
CLOSE #1

IF C(1) > 0 THEN
        LET PERYRS = C(1) / 365.25
        LET DTHS = C(2)
END IF

IF DTHS < 0 THEN
        DTHS = 0
END IF

REM **WRITE RAW FILE**
OPEN "C:\YYY.RAW" FOR OUTPUT AS #2
WRITE #2, DTHS, PERYRS
CLOSE #2
```

FIGURE 2 *Sample BASIC program for calculation of data for crude rates*

have detailed exposure monitoring data, cumulative exposure can also be calculated as a time-dependent variable (summing routine exposure measurements). Race, gender, birth cohort, and other fixed variables simply add further dimensions to the array; since no calculations are required to classify these variables at each day of observation their inclusion adds very little running time to the program.

In Figure 1, an example of our person-time program, written for use with the SAS system, is presented.[6] This simple program counts person-days and converts the results to person-years, to allow the calculation of crude rates. The source data includes the following: date of hire as a SAS date (HIREDATE); date of last observation as a SAS date (DLO); and, a numeric indicator of outcome status (DEAD).

In this sample program person-days and deaths are counted in a one-dimensional array. The statement ‘C(1)+1’ increments the count of person-time by one, while the statement ‘C(2)+1’ increments the count of deaths by one. Often one is interested in studying multiple outcomes, such as different causes of death, in a cohort. Multiple outcomes can be counted by this program in one pass of the data by allowing the indicator for outcome status to include more than two levels, and counting deaths due to each cause separately.

The program creates a raw dataset in which the number of person-years and events at each level of the stratifying variables are recorded; this file may be easily read into standard statistical packages.[5,8] The dataset includes observations only for those cells of the person-time table in which days of follow-up were counted. For cells in which no deaths occurred, a zero is entered for the number of deaths using the statements ‘DUM2=C(2); DTHS=SUM(DUM2,0)’.

This method can easily be adapted to other programming languages. For example, Figure 2 presents a BASIC language program which performs the same

```
LIBNAME IN1  'C:\';
FILENAME OUT1 'C:\YYYY.RAW';

DATA PYWORK  (KEEP=PERYRS DTHS AGERISK  RACE  GENDER  COHRISK  DOSE  PERIOD);
     SET IN1.XXXX   END=EOF ;

**** DEFINE ARRAY:  TYPE, AGERISK, RACE, GENDER, COHORT, DOSE, PERIOD ***;
ARRAY C(2,15,2,2,5,9,11) _TEMPORARY_;
ARRAY CUM_DOSE(43) CUM43-CUM85;
LAGYEARS=20;     DOSE=1;    PERIOD=1;  MID42=MDY(7,1,42);

**** GENERATE PERSON-DAYS ***;
DO DAY = HIREDATE TO DLO;
         AGERISK=FLOOR((DAY-BIRTHDTE)/1826.25)-3;
         IF AGERISK <1 THEN AGERISK= 1;  ELSE IF  AGERISK >15  THEN  AGERISK= 15;
         YR=YEAR(DAY);    PERIOD=YR-1979;  IF  PERIOD < 1  THEN  PERIOD= 1;

**DOSE WAS ASSESSED ONLY ONCE A YEAR.  FOLLOWING CODE IS DONE ONCE A YEAR*;
IF  YR_IND  NE FLOOR((DAY-MID42)/365.25)  THEN  DO;
                  YR_IND = FLOOR((DAY-MID42)/365.25);
                  IF YR_IND-LAGYEARS > 0 THEN DO;
                           CUMDOSE1 = CUM_DOSE(YR_IND-LAGYEARS);
                           IF  CUMDOSE1 =  0 THEN  DOSE=1;
                            ELSE IF  0 < CUMDOSE1< 500 THEN  DOSE = 2;
                            ELSE DO;
                                    DOSE = (FLOOR(LOG2(CUMDOSE1 / 500)))+3;
                                    DOSE =MIN(DOSE,9);
         END ;  END;   END;
C(1, AGERISK,RACE,GENDER,COHRISK,DOSE,PERIOD)+1;
END;

**** GENERATE DEATHS  ***;
IF DEAD2=1 THEN C(2, AGERISK,RACE,GENDER,COHRISK,DOSE,PERIOD)+1;

**** SUMMARIZE ARRAY ***;
IF EOF THEN DO  A=1 to 15; DO B=1 to 2; DO D=1 to 2; DO E=1 to 5; DO F=1 to 9; DO G=1 to 11;
AGERISK=A; RACE=B; GENDER=D; COHRISK=E; DOSE=F; PERIOD = G;
  IF C(1,A,B,D,E,F,G) > 0 THEN DO;
 PERYRS=C(1, A,B,D,E,F,G)/365.25;
DUM2=C(2, A,B,D,E,F,G); DTHS=SUM(DUM2,0);
OUTPUT PYWORK;
END; END; END; END;END;END;END;
RUN;
DATA T ;  SET PYWORK;
FILE OUT2;
PUT PERYRS DTHS AGERISK RACE GENDER COHRISK DOSE PERIOD;
RUN;
```

FIGURE 3 *Sample SAS program for the creation of person-year tables stratified by age, race, gender, birth cohort, cumulative dose, and calendar year*

tabulation of person-time and events. Again, the source data includes the following: the date of hire, counted in days from 1/1/1960 (HIREDATE); the date of last observation, also in days from 1/1/1960 (DLO); and a numeric indicator of outcome status (DEAD).

Figure 3 presents a sample SAS program in which person-time and events, stratified by six factors (age at risk, race, gender, birth cohort, 20-year lagged cumulative dose, and calendar year), are allocated in a table with 59 400 cells. The number of levels for each time-dependent variable must be decided upon ahead of time, and these categories must be exhaustive. For example, in the sample program, age at risk is categorized into 15 levels with upper and lower bounds of <25 years and >90 years. Similarly, cumulative dose is grouped into nine categories with an upper bound of >3200 mSv and a lower bound of 0 mSv.

When disease rates are examined in relation to an exposure measured on a continuous scale it is often desirable to associate a specific exposure value with each cell of the person-time table.[9] These values are useful for graphical display of the data, and for regression analyses that consider exposure as a continuous variable. Cell-specific mean doses can be calculated by summing not only person-days and deaths, but cumulative dose counted within each cell as well. The mean dose in a cell, then, is the sum of the cumulative dose divided by the number of person-days of observation.

DISCUSSION

This paper presents a simple way to generate person-time data in which time-dependent variables are accurately classified at each interval of observation. Importantly, this program does not require any specialized software; the programs we present were written for a personal computer using the SAS system, as well as an example written in BASIC. Consequently, this method is well suited to helping a wide range of researchers accomplish one of the fundamental tasks of cohort analyses–creating person-time data in order to examine disease and mortality rates.

One of the most useful aspects of this program is its flexibility in creating time-related variables. The sample program in Figure 3 shows a method of categorizing person-time using lagged cumulative exposure. More complicated exposure classifications, such as exposure time-windows,[10] are often also of interest, and can easily be accommodated.

One limitation of this program, in contrast to Pearce's method, is the additional running time necessary to count person-days of follow-up. Since the program counts follow-up time within an array, however, there is no need to construct a large file or sort the resulting dataset. The program in Figure 3 counts 75 000 person-years in less than 40 minutes on a personal computer with a Pentium processor.

Many of the limitations we identified with previous methods used for counting person-time are minor when considering simple crude or age-adjusted rates. However, in more complicated analyses, for example investigations of the role of time-dependent factors in exposure-response relationships, problems of rounding and misclassification of person-time are potentially important.

Given the advances in the processing speed of personal computers it is no longer necessary to sacrifice precision in classifying person-time in order to conveniently create person-time data for analysis. This program provides a simple way to create tables using whatever level of precision is available for classifying time-dependent; furthermore, it can be easily adapted to examine complicated functions for creating time-dependent variables.

ACKNOWLEDGEMENTS

This work was supported in part by CDC grant 1 RO3 OH03343–01. Special thanks to Neil Pearce for his work on this problem and his comments on this program. Also thanks to Vic Rhodes for programming assistance.

REFERENCES

[1] Frome E L, Checkoway H. Use of Poisson regression models in estimating incidence rates and ratios. *Am J Epidemiol* 1985; **121:** 309–23.

[2] Waxweiler R J, Beaumont J J, Henry J A *et al*. A modified life-table analysis system for cohort studies. *J Occup Med* 1983; **25:** 115–23.

[3] Pearce N, Checkoway H. A simple computer program for generating person-time data in cohort studies involving time-related factors. *Am J Epidemiol* 1987; **125:** 1085–91.

[4] Macaluzo, M. Exact stratification of person-years. *Epidemiology* 1992; **3:** 441–48.

[5] Preston D L, Lubin J H, Pierce D A, McConney M E. *Epicure User's Guide*. Seattle: Hirosoft International Corp., 1993.

[6] SAS Institute Inc. *SAS Language: Reference, Version 6, First Edition*. Cary, NC: SAS Institute Inc., 1990.

[7] Arrighi H M, Hertz-Piccioto I. The evolving concept of the healthy worker survivor effect. *Epidemiology* 1994; **5:** 189–96.

[8] Royal Statistical Society, *The GLIM System, Release 4 Manual*. Oxford: Oxford University Press, 1993.

[9] National Research Council, Committee on the Biological Effects of Ionizing Radiations. *Health Risks of Radon and Other Internally Deposited Alpha-Emitters (BEIR IV)*. Washington, DC: National Academy Press, 1988.

[10] Rothman K. Induction and latent periods. *Am J Epidemiol* 1981; **114:** 253–59.

(Revised version received September 1996)