# An R package for fitting age, period and cohort models

Adriano Decarli(1), Carlo La Vecchia(1), Matteo Malvezzi(1), Rocco Micciolo(2)

## ABSTRACT

**In this paper we present the R implementation of a GLIM macro which fits age-period-cohort model following Osmond and Gardner. In addition to the estimates of the corresponding model, owing to the programming capability of R as an object oriented language, methods for printing, plotting and summarizing the results are provided. Furthermore, the researcher has fully access to the output of the main function (apc) which returns all the models fitted within the function. It is so possible to critically evaluate the goodness of fit of the resulting model.**

*Key words: Age-period-cohort models, Cohort analysis, R language, Trends*

*(1) Department of Clinical Sciences and Community Health, Sezione di Statistica Medica e Biometria "Giulio A. Maccacaro", Università degli Studi di Milano, Milano*
*(2) Department of Psychology and Cognitive Sciences, Università degli Studi di Trento, Rovereto, Italy*

*CORRESPONDING AUTHOR: Rocco Micciolo, Department of Psychology and Cognitive Sciences, Università degli Studi di Trento, Corso Bettini 31, 38068 Rovereto, Italy; Tel:+39 0464 808634; email: rocco.micciolo@unitn.it*

*DOI:* ***10.2427/9977***
*Accepted on 11 April, 2014; Published as Online First on 17 December, 2014*

## INTRODUCTION

In this short note we present an R [1] package suitable to fit age, period and cohort (APC) models. The core component of the package is the R implementation of a GLIM [2] macro [3] based on a model proposed by Osmond and Gardner [4].

The main issue making APC models complicated to treat is that age, period and cohort are not independent. Separate analyses of the three effects, as well as two-factor analyses, ignoring one of the three variables (age, period, cohort) sequentially, can produce misleading results when these variables have distinct underlying biological interpretations. Clayton and Schifflers [5 , 6] reviewed methods for modelling variation in cancer incidence and mortality rates in term of either period and/or cohort effects in the general multiplicative risk model, drawing attention to the difficulty of attributing regular trends to either period or cohort influences.

The function we present here is based on a solution proposed by Osmond and Gardner [4] that researchers or health care planners can use easily. Furthermore, since R is an object oriented programming language, it is possible to employ the same function to obtain estimates as well as diagnostic statistics for "simpler" models like the "age only", "age + period" and "age + cohort" models. We hope that, given the versatility and the diffusion of R, this function could serve as a standard base reference, which can be easily modified or integrated by statisticians.

## METHODOLOGICAL BACKGROUND

Before presenting the R package, we give a short methodological summary of the model implemented; more details can be found in [3].

The data for APC models are usually derived starting from two tables containing, respectively, the number of observed deaths and the estimates of resident population. Tables 1 and 2 show an example derived from the original article where the GLIM macro was presented [3]. Usually in this tables, the age groups are displayed in rows and the calendar periods in columns and the grouping interval is equal on both sides (5 years in our example). From these two tables a third table of age and calendar period specific rates is computed, where the incidence rates of the corresponding birth cohorts can be read from the diagonals.

Let us consider the following linear model:

$$y_{ij} = \mu + \alpha_i + \pi_j + \gamma_k + \varepsilon_{ij}$$

where α,π and $y$ represent the age ($i = 1,2,\ldots,I$), the calendar period ($j = 1,2,\ldots,J$) and the cohort period ($k = I - i + j$) effects respectively, and the dependent variable $y_{ij}$ is a function of the incidence rate. Taking the linear relationships between the three independent variables into account, the model has a general problem of identifiability. However, for the practical purposes of epidemiological interpretation, interest is mainly focused on the estimation of differences between various cohorts in relative terms.

Among the many proposed solutions to work around the above mentioned issue, the one proposed by Osmond and Gardner [4] has found wide application in the analysis of mortality data.

If $O_{ij}$ (the number of deaths in the *i*-th age group and the *j*-th calendar period) is a Poisson variable and $N_{ij}$ is the corresponding number of subjects at risk (considered non random), then $y_{ij} = \ln \frac{O_{ij}}{N_{ij}}$. Let us now consider the following log-linear model

$$\ln O_{ij} = \ln N_{ij} + \ln a_i + \ln p_i + \ln c_k$$

where the parameters (corresponding to $\alpha_i$, $\pi_j$, $\gamma_k$) can be estimated minimizing

$$f(\mathbf{a},\mathbf{p},\mathbf{c}) = \sum O_{ij} \cdot \left(\ln O_{ij} - \ln N_{ij} - \ln a_i - \ln p_i - \ln c_k\right)^2 \quad (1)$$

Due to the linear relationships between age, period and cohort, the solution set **X**(**a**,**p**,**c**) is infinite. However, the solution set can be re-parameterized in a further variable λ as **X**(**a**,**p**,**c**,λ) and a goodness of fit statistic which is independent from λ can be calculated (see references 3 and 7 for further details).

Following Osmond and Gardner, the three two-factor log-linear models

$$f(\mathbf{a}_0,\mathbf{p},\mathbf{c}) \qquad f(\mathbf{a},\mathbf{p},\mathbf{c}_0) \qquad f(\mathbf{a},\mathbf{p}_0,\mathbf{c})$$

are interpolated, minimizing, respectively, the corresponding functions (1), where $\mathbf{c}_0$ and $\mathbf{p}_0$ are unit vectors of appropriate length and $\mathbf{a}_0$ is the vector

$$a_{0j} = \exp\left[\sum_j O_{ij}\left(\ln O_{ij} - \ln N_{ij}\right) \Big/ \sum_j O_{ij}\right]$$

The three parameter estimate vectors

$$\mathbf{X}_a = \left(\mathbf{a}_0, \hat{\mathbf{p}}_a, \hat{\mathbf{c}}_a\right) \quad \mathbf{X}_c = \left(\hat{\mathbf{a}}_c, \hat{\mathbf{p}}_c, \mathbf{c}_0\right) \quad \mathbf{X}_p = \left(\hat{\mathbf{a}}_p, \mathbf{p}_0, \hat{\mathbf{c}}_p\right)$$

have corresponding goodness of fit measures $G_a^2, G_c^2, G_p^2$.

The Euclidean distances between each two factor model and the full model **X**(**a**,**p**,**c**,λ) are

$$d_a(\lambda_1) = \| \mathbf{X}_a - \mathbf{X}(\lambda_1) \| \quad d_p(\lambda_2) = \| \mathbf{X}_p - \mathbf{X}(\lambda_2) \| \quad d_c(\lambda_3) = \| \mathbf{X}_c - \mathbf{X}(\lambda_3) \|$$

TABLE 1

GASTRIC CANCER DEATH CERTIFICATIONS FOR MALES AGED 25-74 YEARS, ITALY 1955-1979

| | PERIOD OF DEATH | | | | |
|---|---|---|---|---|---|
| AGE GROUP AT DEATH | 1955 – 59 | 1960 – 64 | 1965 – 69 | 1970 – 74 | 1975 – 79 |
| 25 – 29 | 84 | 89 | 72 | 67 | 65 |
| 30 – 34 | 193 | 224 | 205 | 176 | 152 |
| 35 – 39 | 385 | 480 | 458 | 425 | 385 |
| 40 – 44 | 1047 | 835 | 946 | 849 | 701 |
| 45 – 49 | 2260 | 1873 | 1443 | 1685 | 1542 |
| 50 – 54 | 3908 | 3830 | 3014 | 2334 | 2685 |
| 55 – 59 | 5433 | 5807 | 5863 | 4552 | 3440 |
| 60 – 64 | 6713 | 7598 | 7909 | 7649 | 5800 |
| 65 – 69 | 8144 | 8426 | 8929 | 9135 | 8810 |
| 70 – 74 | 8782 | 8867 | 8873 | 9022 | 9010 |

TABLE 2

RESIDENT POPULATION FOR MALES AGED 25-74 YEARS, ITALY 1955-1979

| | PERIOD | | | | |
|---|---|---|---|---|---|
| AGE GROUP | 1955 – 59 | 1960 – 64 | 1965 – 69 | 1970 – 74 | 1975 – 79 |
| 25 – 29 | 9882353 | 9569892 | 9729730 | 9054054 | 10317460 |
| 30 – 34 | 9507389 | 9531915 | 9318182 | 9513514 | 9101796 |
| 35 – 39 | 7129630 | 9266409 | 9346939 | 9120172 | 9459459 |
| 40 – 44 | 7830965 | 6964137 | 9052632 | 9148707 | 9045161 |
| 45 – 49 | 8097456 | 7610727 | 6758782 | 8798956 | 8949507 |
| 50 – 54 | 6803621 | 7794058 | 7294288 | 6486937 | 8540076 |
| 55 – 59 | 5576884 | 6417284 | 7410263 | 6850263 | 6175943 |
| 60 – 64 | 4380138 | 5102754 | 5942149 | 6834346 | 6302978 |
| 65 – 69 | 3642870 | 3805953 | 4447821 | 5248793 | 6000545 |
| 70 – 74 | 2784224 | 2948002 | 3032260 | 3573211 | 4264080 |

The weighted sum of these distances are minimized with respect to l. In this procedure each of the three two-factor models is included with a weight inversely proportional to the corresponding model goodness of fit statistic. Hence, the solution minimizes $\hat{\mathbf{x}}(\lambda)$ the distance of the saturated model from the three two-factor models and can be considered a geometrical weighted average.

## IMPLEMENTATION OF THE AGE, PERIOD, COHORT MODEL IN THE `apc` PACKAGE

The core component of the `apc` package is the `apc` function which is invoked with two arguments: the first (`num`) is the matrix (with age groups in rows and calendar periods in columns) containing the number of the considered events, while the second (`den`) is the corresponding population at risk matrix. A third argument (`scale`) has a default value of 100,000 and represents the scale adopted for incidence rates. To properly use the `apc` function, it is fundamental that the data are grouped with equal time intervals on both age and calendar period (say 5 years). The `apc` function calculates the number of age groups, calendar periods and cohorts from the two input matrices rows and columns. The function begins estimating the age effects alone. Then it fits the three two-factor models ($a_0pc$, $ap_0c$ and $apc_0$). Finally the weighted sum of the squares of the distances is minimized. The output of the function is a list containing, among others, the R objects resulting from all the above mentioned fitted models, giving to the researcher full access for deeper analyses. Two local functions (`norm` and `pcpet`) are defined within the `apc` function. The former normalizes cohort and period effects, while the latter extracts requested values from the input vector (padding the remaining cells with zeroes).

A number of methods for printing and/or plotting results are provided. In particular, a `print` method gives the estimated age, cohort and period effects, a `predict` method gives (among others) the estimated number of events, a `summary` method shows detailed results with respect to all the fitted models and an `anova` method prints two analysis of deviance tables showing the deviances, the degrees of freedom and the values of AIC associated with each fitted model. These analysis of deviance tables can be used to compare selected models as shown in Clayton and Schifflers [5 , 6]. In addition to the above mentioned models (age, age + period, age + cohort, age + period + cohort), the deviance of an “age + drift” model is shown. Such a model, which is discussed in details in Clayton and Schifflers [5 , 6], refers to a specific type of regular trend in which the ratio of age-specific rates between two adjacent time periods is not only constant across age groups, but is constant for any pair of adjacent time periods (or, alternatively, the relative risk between adjacent birth cohorts is constant).

The modelling technique does not allow for the calculation of confidence intervals of the parameter estimated for the “full” age, period, cohort model in a conventional manner. However the package `apc` contains a `confint` method which performs a parametric bootstrap simulation. Data for each 5-year age-specific number of deaths in all time periods is obtained generating, by means of the `rpois` function, pseudo-random numbers from a Poisson distribution with an expected value equal to the observed number of deaths for that period and age-group. The resulting datasets are passed to the `apc` function and bootstrap parameter estimates are stored. The arguments `nrep` and `level` control the number of bootstrap replications (i.e. the number of bootstrap datasets) and the confidence level, respectively.

## SOME ILLUSTRATIVE EXAMPLES

### Gastric cancer certification rates in males aged 25-74 years (Italy 1955-1979)

The data shown in tables 1 and 2 are those of the example published from Decarli and La Vecchia [3] in the article where the original GLIM macro was presented. The analysis on gastric cancer was subsequently updated [7]. To illustrate how to use the function `apc`, we consider that these data are stored, without row and column labels, in two external ASCII files, named `num.txt` and `den.txt`, respectively. As a first step, data are read and stored in the data-frames `num` and `den`:

```
num <- read.table(“num.txt”,header=FALSE)
den <- read.table(“den.txt”,header=FALSE)
```

For printing purposes it is better that both `num` and `den` have labels for the age groups (rows) and calendar periods (columns). This can be easily accomplished within R. For the data presented in tables 1 and 2, the considered ages range between 25 and 74 years, (with central values between 27 and 72 years) in 5-year categories; therefore row labels can be obtained in the following way:

```
age <- seq(27,72,5); x <- cbind(age-2,age+2)
lbl <- paste(x[,1],”-“,x[,2],sep=””)
rownames(num) <- lbl; rownames(den) <- lbl
```

As far as the calendar periods are concerned, the considered years range between 1955 and 1979 (with central values between 1957 and 1977) in 5-year categories; therefore column labels can be obtained as:

```
per <- seq(1957,1977,5); x <- cbind(per-2,per+2)
lbl <- paste(x[,1],”-“,x[,2],sep=””)
colnames(num) <- lbl; colnames(den) <- lbl
```

Finally, cohort labels can be obtained considering that in this example the first central calendar year is 1957, the oldest central age is 72 years and there are a total of 14 cohorts (i.e. the number of age groups (10) minus the number of calendar periods (5) plus 1):

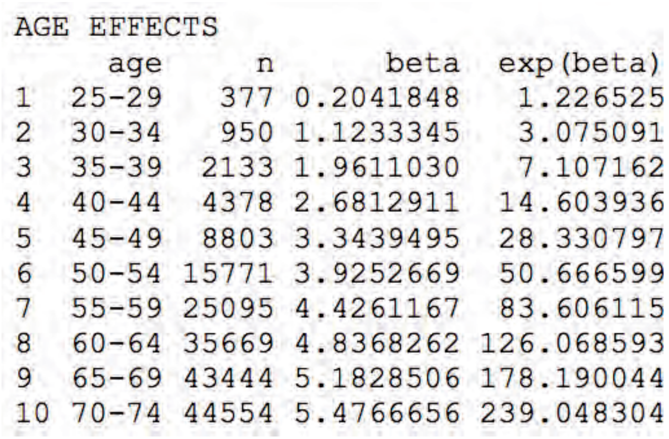
```
n <- nrow(num)+ncol(num)-1
coh <- c(1:n); tmp <- 1957-72+(coh-1)*5
x <- cbind(tmp-2,tmp+2)
```

```
clab <- paste(x[,1],”-“,x[,2],sep=””)
```

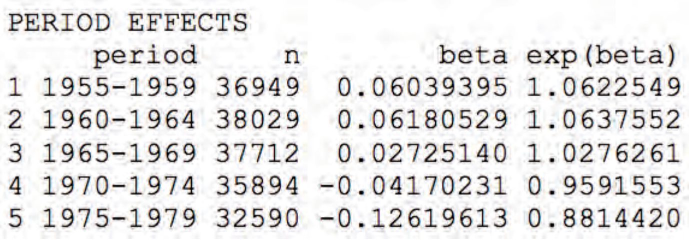
Now, the function `apc` can be invoked (passing the cohort labels using the argument `labels`); the results are returned in the object `fit`:

```
fit <- apc(num,den,labels=clab)
```

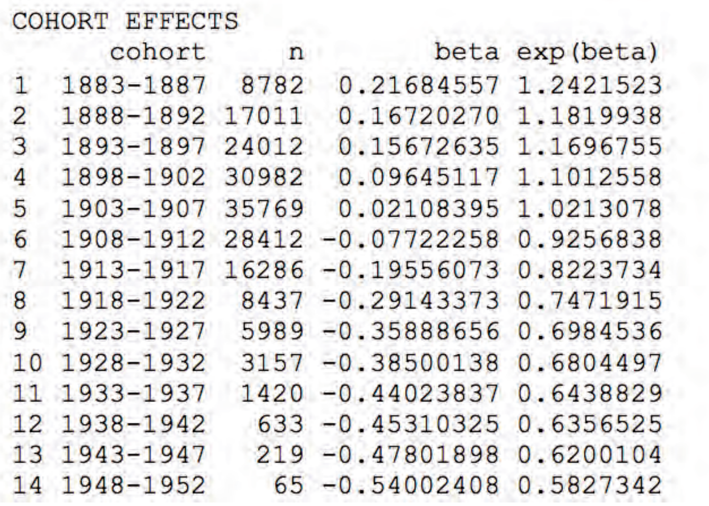
The estimated age, cohort and period effects can be immediately printed by typing the name of the returned object (`fit`):



```
AGE EFFECTS
      age     n      beta   exp(beta)
1  25-29   377 0.2041848    1.226525
2  30-34   950 1.1233345    3.075091
3  35-39  2133 1.9611030    7.107162
4  40-44  4378 2.6812911   14.603936
5  45-49  8803 3.3439495   28.330797
6  50-54 15771 3.9252669   50.666599
7  55-59 25095 4.4261167   83.606115
8  60-64 35669 4.8368262  126.068593
9  65-69 43444 5.1828506  178.190044
10 70-74 44554 5.4766656  239.048304
```



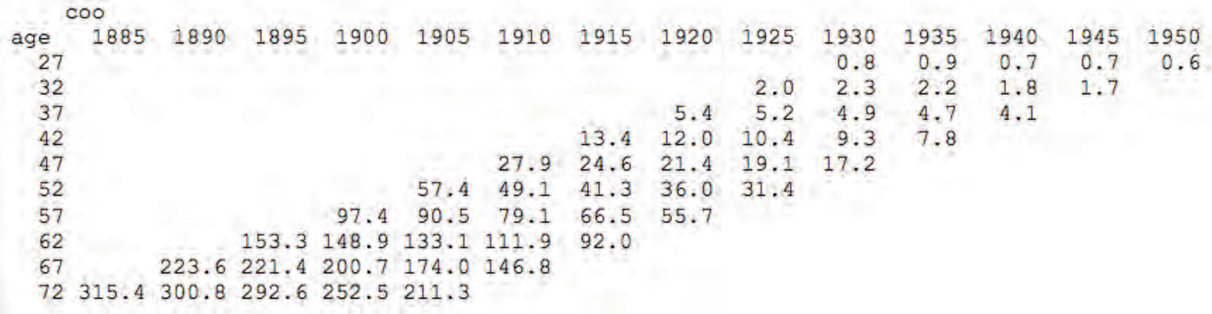
```
PERIOD EFFECTS
     period     n        beta exp(beta)
1 1955-1959 36949  0.06039395 1.0622549
2 1960-1964 38029  0.06180529 1.0637552
3 1965-1969 37712  0.02725140 1.0276261
4 1970-1974 35894 -0.04170231 0.9591553
5 1975-1979 32590 -0.12619613 0.8814420
```



```
COHORT EFFECTS
      cohort     n        beta exp(beta)
1  1883-1887  8782  0.21684557 1.2421523
2  1888-1892 17011  0.16720270 1.1819938
3  1893-1897 24012  0.15672635 1.1696755
4  1898-1902 30982  0.09645117 1.1012558
5  1903-1907 35769  0.02108395 1.0213078
6  1908-1912 28412 -0.07722258 0.9256838
7  1913-1917 16286 -0.19556073 0.8223734
8  1918-1922  8437 -0.29143373 0.7471915
9  1923-1927  5989 -0.35888656 0.6984536
10 1928-1932  3157 -0.38500138 0.6804497
11 1933-1937  1420 -0.44023837 0.6438829
12 1938-1942   633 -0.45310325 0.6356525
13 1943-1947   219 -0.47801898 0.6200104
14 1948-1952    65 -0.54002408 0.5827342
```

In addition to the estimated age, cohort and period effects, the corresponding number of events are printed. The observed age-specific rates for each considered period can be obtained as `num/den`; the age-specific rates rearranged by central date of birth are stored in the output of `apc` (within the object `actable`) and can be printed as follows:

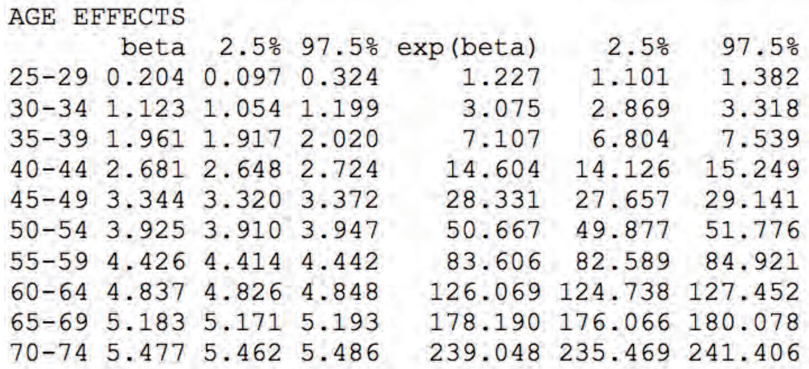
```
lbl <- seq(27,74,5); rownames(fit$actable) <- lbl
lbl <- seq(1885,1950,5); colnames(fit$actable) <- lbl
round(fit$actable,1)
```



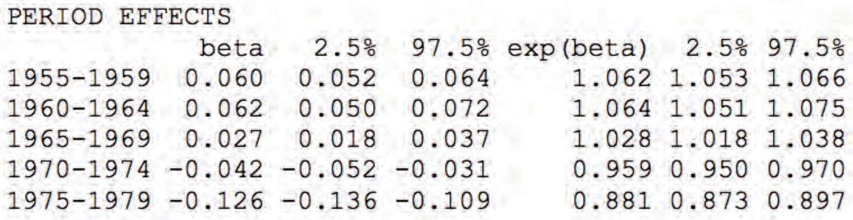
```
    coo
age  1885  1890  1895  1900  1905  1910  1915  1920  1925  1930  1935  1940  1945  1950
 27                                                     0.8   0.9   0.7   0.7   0.6
 32                                               2.0   2.3   2.2   1.8   1.7
 37                                         5.4   5.2   4.9   4.7   4.1
 42                                  13.4  12.0  10.4   9.3   7.8
 47                            27.9  24.6  21.4  19.1  17.2
 52                      57.4  49.1  41.3  36.0  31.4
 57                97.4  90.5  79.1  66.5  55.7
 62         153.3 148.9 133.1 111.9  92.0
 67   223.6 221.4 200.7 174.0 146.8
 72 315.4 300.8 292.6 252.5 211.3
```

Bootstrap confidence intervals for the parameter estimates can be obtained invoking the function `confint` specifying the number of bootstrap replications (via the argument `nrep` which has a default of 100) and the confidence level (via the argument `level` which has a default of 0.95):

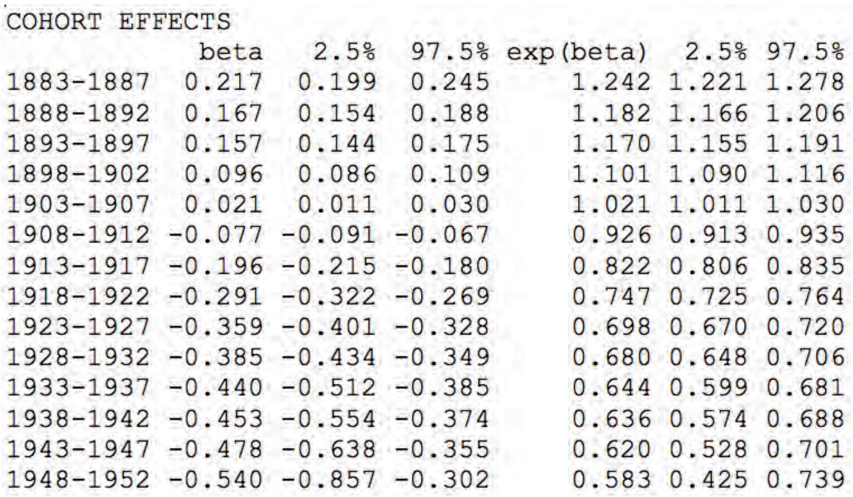
```
set.seed(123456)
tmp <- confint(fit, nrep=1000)
print(tmp,round=3)
Parameter estimates and 95% confidence interval (using 1000 replicates)
```



```
AGE EFFECTS
          beta  2.5% 97.5% exp(beta)    2.5%   97.5%
25-29 0.204 0.097 0.324     1.227   1.101   1.382
30-34 1.123 1.054 1.199     3.075   2.869   3.318
35-39 1.961 1.917 2.020     7.107   6.804   7.539
40-44 2.681 2.648 2.724    14.604  14.126  15.249
45-49 3.344 3.320 3.372    28.331  27.657  29.141
50-54 3.925 3.910 3.947    50.667  49.877  51.776
55-59 4.426 4.414 4.442    83.606  82.589  84.921
60-64 4.837 4.826 4.848   126.069 124.738 127.452
65-69 5.183 5.171 5.193   178.190 176.066 180.078
70-74 5.477 5.462 5.486   239.048 235.469 241.406
```



```
PERIOD EFFECTS
             beta   2.5%  97.5% exp(beta)  2.5% 97.5%
1955-1959  0.060  0.052  0.064     1.062 1.053 1.066
1960-1964  0.062  0.050  0.072     1.064 1.051 1.075
1965-1969  0.027  0.018  0.037     1.028 1.018 1.038
1970-1974 -0.042 -0.052 -0.031     0.959 0.950 0.970
1975-1979 -0.126 -0.136 -0.109     0.881 0.873 0.897
```

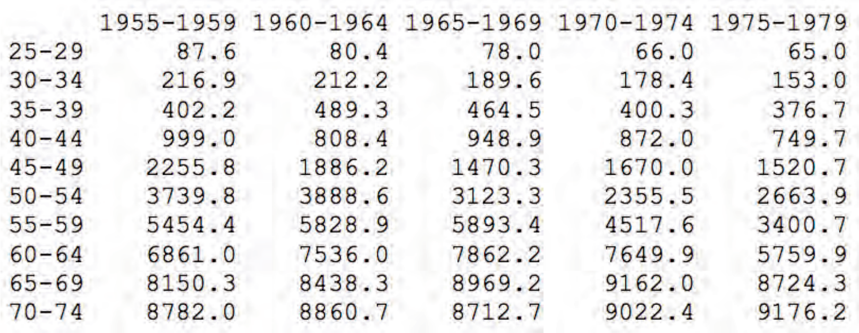


```
COHORT EFFECTS
             beta   2.5%  97.5% exp(beta)  2.5% 97.5%
1883-1887  0.217  0.199  0.245     1.242 1.221 1.278
1888-1892  0.167  0.154  0.188     1.182 1.166 1.206
1893-1897  0.157  0.144  0.175     1.170 1.155 1.191
1898-1902  0.096  0.086  0.109     1.101 1.090 1.116
1903-1907  0.021  0.011  0.030     1.021 1.011 1.030
1908-1912 -0.077 -0.091 -0.067     0.926 0.913 0.935
1913-1917 -0.196 -0.215 -0.180     0.822 0.806 0.835
1918-1922 -0.291 -0.322 -0.269     0.747 0.725 0.764
1923-1927 -0.359 -0.401 -0.328     0.698 0.670 0.720
1928-1932 -0.385 -0.434 -0.349     0.680 0.648 0.706
1933-1937 -0.440 -0.512 -0.385     0.644 0.599 0.681
1938-1942 -0.453 -0.554 -0.374     0.636 0.574 0.688
1943-1947 -0.478 -0.638 -0.355     0.620 0.528 0.701
1948-1952 -0.540 -0.857 -0.302     0.583 0.425 0.739
```

The first command (set.seed(123456)) permits to reproduce exactly the results obtained.

The predict method is useful since it returns a data-frame containing both the observed and the predicted number of events (in the variables y and fitted.values of the data-frame, respectively) in addition to the covariates of the fitted model. For example, the predicted estimates to be compared with the corresponding observed values (in Table 1) can be obtained (and printed) in the following way:

```
yhat <- predict(fit)$fitted
yhat <- matrix(yhat,nrow=nrow(num))
rownames(yhat) <- rownames(num)
colnames(yhat) <- colnames(num)
round(yhat,1)
```



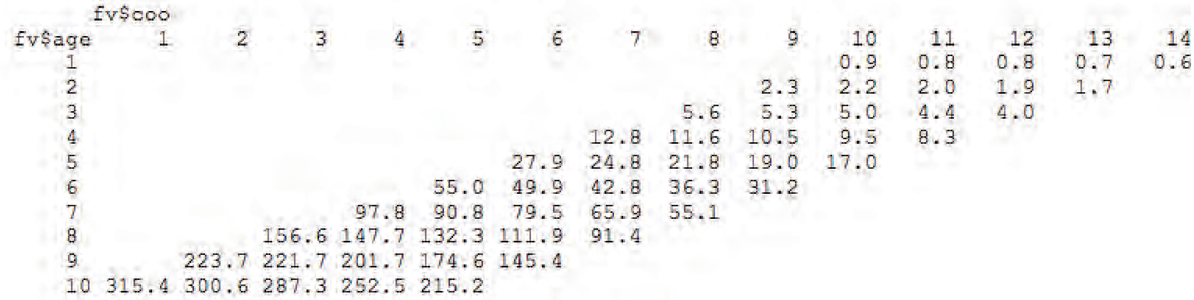
```
      1955-1959 1960-1964 1965-1969 1970-1974 1975-1979
25-29      87.6      80.4      78.0      66.0      65.0
30-34     216.9     212.2     189.6     178.4     153.0
35-39     402.2     489.3     464.5     400.3     376.7
40-44     999.0     808.4     948.9     872.0     749.7
45-49    2255.8    1886.2    1470.3    1670.0    1520.7
50-54    3739.8    3888.6    3123.3    2355.5    2663.9
55-59    5454.4    5828.9    5893.4    4517.6    3400.7
60-64    6861.0    7536.0    7862.2    7649.9    5759.9
65-69    8150.3    8438.3    8969.2    9162.0    8724.3
70-74    8782.0    8860.7    8712.7    9022.4    9176.2
```

A plot of observed vs expected events (on a logarithmic scale) can also be obtained:

```
plot(predict(fit)$y,predict(fit)$fitted,log="xy").
```

The expected rates can be printed, arranged by birth cohorts:

```
fv <- predict(fit)
rates <- fv$fitted/exp(fv$offset)
tmp <- xtabs(rates ~ fv$age+fv$coo)
tmp[tmp == 0] <- NA
round(tmp,1)
```



```
      fv$coo
fv$age     1     2     3     4     5     6     7     8     9    10    11    12    13    14
     1                                                         0.9   0.8   0.8   0.7   0.6
     2                                                   2.3   2.2   2.0   1.9   1.7
     3                                             5.6   5.3   5.0   4.4   4.0
     4                                      12.8  11.6  10.5   9.5   8.3
     5                                27.9  24.8  21.8  19.0  17.0
     6                          55.0  49.9  42.8  36.3  31.2
     7                    97.8  90.8  79.5  65.9  55.1
     8             156.6 147.7 132.3 111.9  91.4
     9       223.7 221.7 201.7 174.6 145.4
    10 315.4 300.6 287.3 252.5 215.2
```

The package apc contains a number of datasets taken from the literature. In what follows some examples will be given to display some of the functionality of the R functions included in the apc package.

## Lung cancer death certification rates in males aged 30-79 years (Italy 1970-2009)

By invoking data(lungM) the number of death certification rates for lung cancer observed between 1970 and 2009 in Italian males aged 30-79 years are available (for intervals of 5 years) in the dataset lungM.num (the corresponding denominator are in the dataset lungM.den). The age-period-cohort model is fitted by issuing the commands

```
coh <- 1972-77+(c(1:17)-1)*5
x <- cbind(coh-2,coh+2)
clab <- paste(x[,1],"-",x[,2],sep="")
fit <- apc(lungM$num,lungM$den,labels=clab)
```

where the first three rows prepare the labels for the 17 cohorts considered. The results can be displayed employing the print and/or the summary functions. Here we invoke the plot function to graphically display the estimates of the effects. This function has an argument labels which has to be a list with three named arguments (age, period, cohort) containing the central values for age, calendar year and date of birth respectively.

```
age <- seq(32,77,5)
per <- seq(1972,2007,5)
coh <- 1972-77+(c(1:17)-1)*5
xlbl <- list(age=age,period=per,cohort=coh)
plot(fit, labels=xlbl)
```

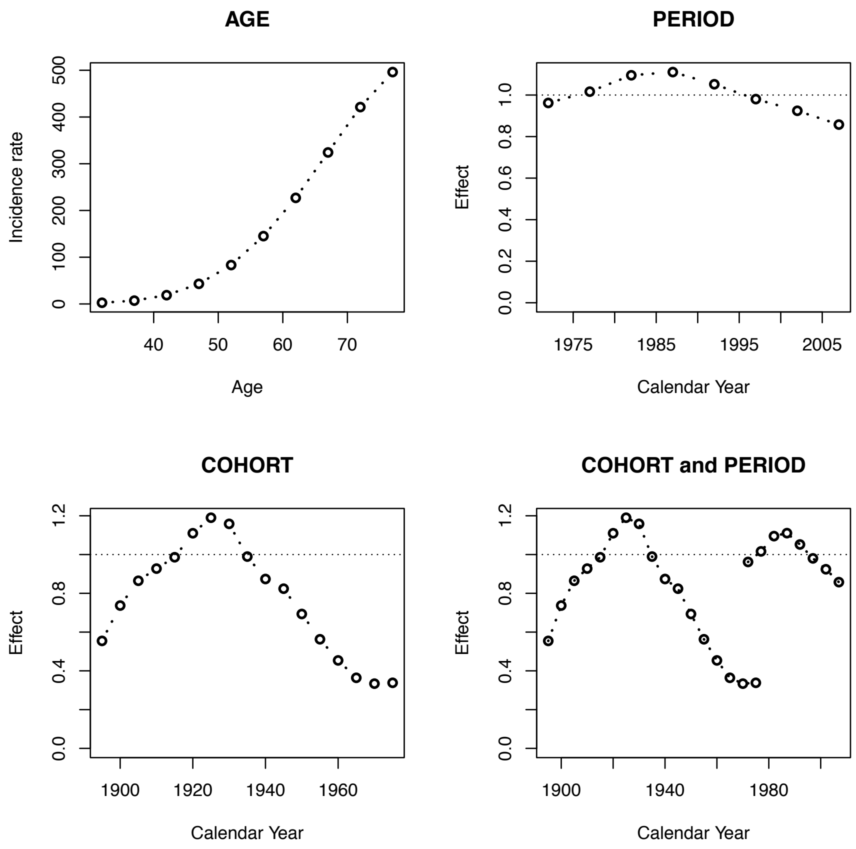
As a default the function plot produces four graphics, as shown in Figure 1.

Besides the well know age effect, common to all non hormone-related epithelial neoplasms [8], there is a strong cohort effect, with major rises for the cohorts born between 1890 and 1920, and subsequent declines up to the cohort born in 1965. This indicates that the worst affected cohort for male lung cancer is the 1920 one, and reflects the pattern of smoking initiation and cessation across subsequent generations of Italian men [9-11]. The peak period effect was registered in 1980, with subsequent declines. This confirms the observation that smoking has not only early, but also late state effects on the process of lung carcinogens, and that stopping smoking leads to reductions of lung cancer (cumulative) incidence and mortality within a few year [12].

Each of the four graphics displayed in figure 1 can be obtained by employing a mode argument with character values "a" (for age-specific rates), "p" (for period effects), "c" (for cohort effects) and "pc" (for cohort and period effects on the same graphic). In these cases, other standard

FIGURE 1

LUNG CANCER DEATH CERTIFICATION RATES IN MALES AGED 30-79 YEARS (ITALY 1970-2009). PLOT OF THE PARAMETER ESTIMATES OF THE "FULL" AGE, PERIOD, COHORT MODEL



graphical arguments (like `lty`, `lwd`, and so on) can be passed to the `plot` function to customize the resulting graphic. A fourth value for the `mode` argument is `"apc"` whose output will be illustrated in the next example.

### Coronary heart disease (CHD) death certification rates in males aged 30-79 years (Italy 1970-2009)

By invoking `data(chdM)` the number of death certification rates for coronary heart disease observed between 1970 and 2009 in Italian males aged 30-79 years are available (for intervals of 5 years) in the dataset `chdM.num` (the corresponding denominator are in the dataset `chdM.den`). The age-period-cohort model is fitted by issuing the same commands displayed above, replacing only the last row with

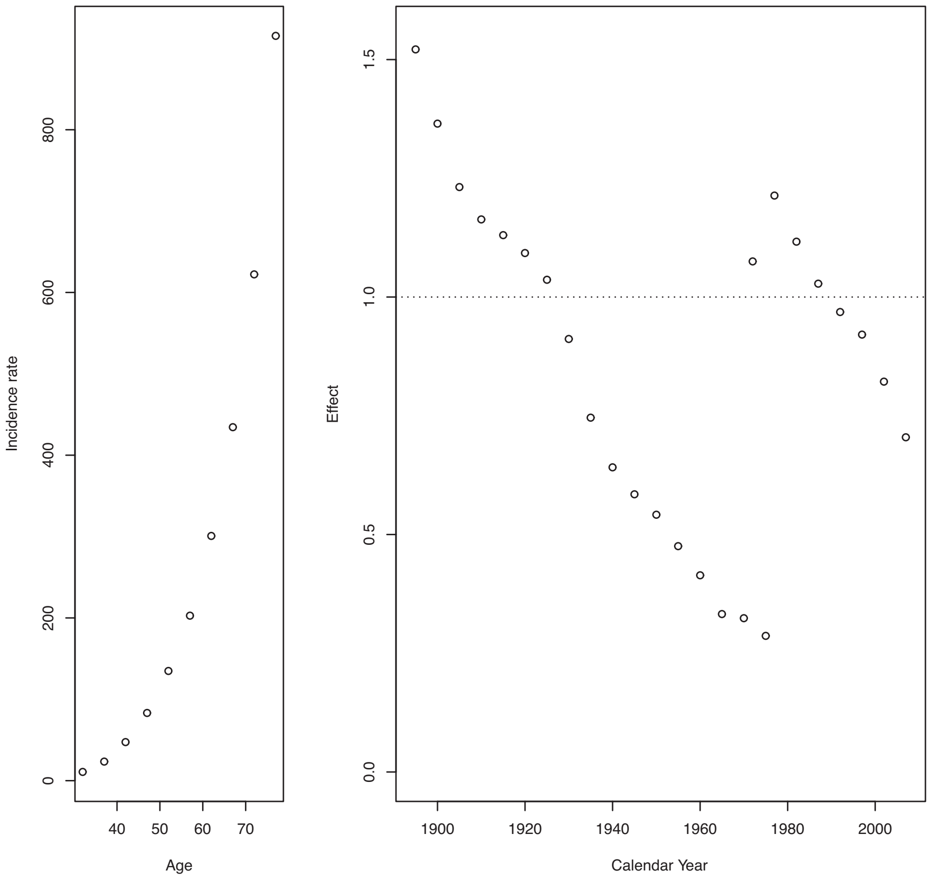
```
fit <- apc(chdM$num,chdM$den,labels=clab)
```

Figure 2 shows the results obtained by issuing the command `plot(fit, labels=xlbl, mode="apc")`.

Apart from the substantial rise of CHD mortality with age, the model shows a major decline in mortality on a cohort basis, starting from the 1900 cohort. The fall was somewhat smaller for the 1910-1920 cohorts, again likely reflecting the tobacco-related disease epidemic in those cohorts [13]. The period effect peaked in 1980, and largely declined thereafter. These data reflect both the long term impact of changing risk factor exposure on CHD mortality (cohort effect), and the improvement

FIGURE 2

CORONARY HEART DISEASE DEATH CERTIFICATION RATES IN MALES AGED 30-79 YEARS (ITALY 1970-2009). PLOT OF THE PARAMETER ESTIMATES OF THE "FULL" AGE, PERIOD, COHORT MODEL



in management and treatment of the diseases [14-16].

The `apc` package includes also the objects `lungF` and `chdF` containing, respectively, the lung cancer death certification rates in females aged 30-79 years (Italy 1970-2009) and the coronary heart disease death certification rates in females aged 30-79 years (Italy 1970-2009).

### Bladder cancer certification rates in males aged 25-79 years (Italy 1955-1979)

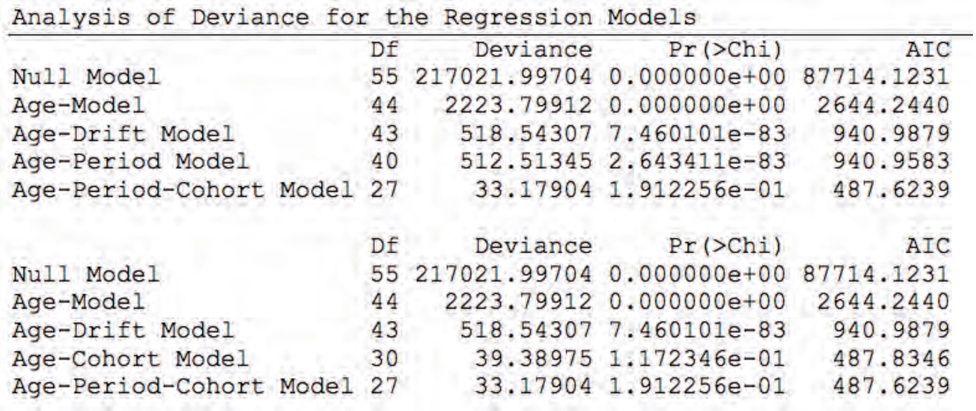
The object `Clayton` is a list containing the number of deaths for bladder cancer in Italian males during the period 1955-1979 as well as the corresponding denominator. These data were employed by Clayton and Schifflers [5] in the first of two papers where age, period and cohort models were discussed. In this example Clayton and Schifflers observed that an attempt to fit the age-period model (i.e. a model including age and period but not cohort) was not very successful. On the other hand, plotting the logarithm of mortality rates of different cohorts against ages resulted in nearly parallel cohort curves, i.e. the differences in age-specific mortality between any pair of birth cohorts was approximately constant throughout the life. In such a case, the age-cohort model (which includes age and cohort, but not period) could provide a useful description of the data.

By means of these data, we show how to replicate, employing the `apc` package, some of the analyses presented by Clayton and Schifflers (the results are not identical, since population data were extrapolated from the corresponding rates presented in the Table IV of the paper of Clayton and Schifflers). The function `apc` can be invoked after having defined the cohort labels:

```
data(Clayton)
coh <- c(1:15); x.coh <- 1957-77+(coh-1)*5
x <- cbind(x.coh-2,x.coh+2)
clab <- paste(x[,1],”-“,x[,2],sep=””)
fit <- apc(Clayton$num,Clayton$den,labels=clab)
```

The deviances associated with each of the models fitted within the function can be displayed by means of the `anova` function:

```
anova(fit)
```



```
Analysis of Deviance for the Regression Models
                          Df      Deviance      Pr(>Chi)        AIC
Null Model                55  217021.99704  0.000000e+00  87714.1231
Age-Model                 44    2223.79912  0.000000e+00   2644.2440
Age-Drift Model           43     518.54307  7.460101e-83    940.9879
Age-Period Model          40     512.51345  2.643411e-83    940.9583
Age-Period-Cohort Model   27      33.17904  1.912256e-01    487.6239

                          Df      Deviance      Pr(>Chi)        AIC
Null Model                55  217021.99704  0.000000e+00  87714.1231
Age-Model                 44    2223.79912  0.000000e+00   2644.2440
Age-Drift Model           43     518.54307  7.460101e-83    940.9879
Age-Cohort Model          30      39.38975  1.172346e-01    487.8346
Age-Period-Cohort Model   27      33.17904  1.912256e-01    487.6239
```

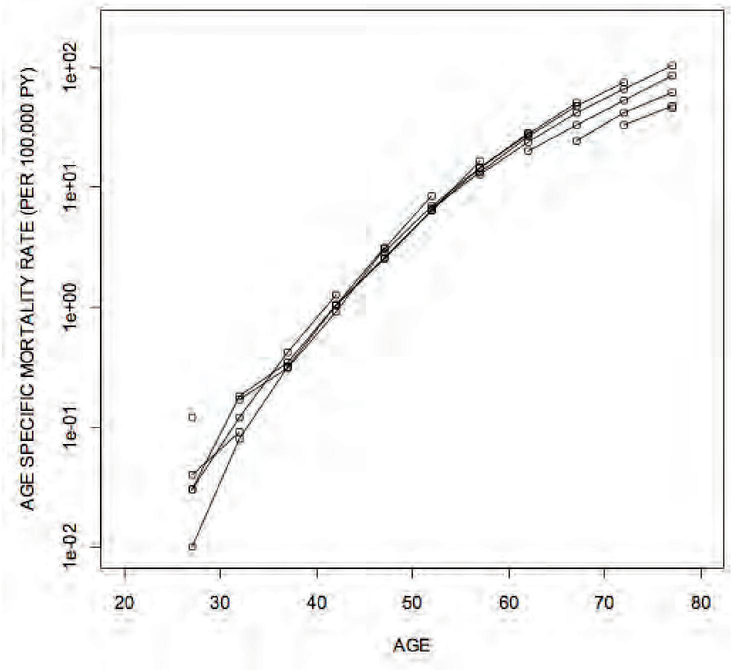
The global deviance chi-squared test of fit of the age-period model is highly significant, yielding a deviance of 513 on 40 degrees of freedom. On the other hand, the fit of the age-cohort model is much better: the global deviance chi-squared test gives 39.4 on 30 degrees of freedom (with a non-significant associated *p-value* of 0.12). As discussed in Clayton and Schifflers [5], in this example the ratio in age-specific mortality between any pair of birth cohorts is approximately constant throughout the life, as can be seen plotting the age-specific rates against age for each birth cohort. Adopting a logarithmic scale for *y* axis, the cohort curves are nearly parallel.

This can be easily accomplished within the `apc` package, since the rates for each cohort are available in the output of the function (in the object `actable`). Figure 3 shows the graphic, generated using the following commands:

```
rates <- fit$actable
```

FIGURE 3

MORTALITY RATES (1955-1979) OF BLADDER CANCER IN ITALIAN MALES AGED 25-79 YEARS BY BIRTH COHORT. RATES ARE PLOTTED USING A LOGARITHMIC SCALE



```
   plot(c(20,80),c(0.01,200),log=”y
”,type=”n”,xlab=”AGE”,
   +              ylab=”AGE  SPECIFIC
MORTALITY RATE (PER 100,000 PY)”)
   for (j in 1:ncol(rates)) {
   +       x <- seq(27,77,5);  y <-
rates[,j]
   +      points(x,y); lines(x,y)
   + }
```

The researcher has fully access to the results of the fitted age-cohort model, which are stored in the `fitac` object within the output of the `acp` function. For example, the command `fit$fitac$coefficients` will print the parameter estimates for age and cohort. In a similar manner, it is possible to perform residual analyses on each of the fitted models.

## CONCLUSIONS

In this paper the R implementation of a GLIM macro [3] which fits age-period-cohort model following Osmond and Gardner [4] was presented. As usual in chronic disease epidemiology, where proportional hazards model are employed, also the age-period model (which predicts constant ratios of age-specific rates between different periods) and the age-cohort model (which predicts constant ratios of age-specific rates between different cohorts) are fitted.

Only in the case where none of these models provides an adequate fit to the observed table of rates, both cohort and period effects can be included. In this case the researcher must be aware that there is a problem of identifiability and that there is no single unique solution, but infinite ones. As Clayton and Schifflers [6] pointed up, “identical descriptions of data may be obtained from different sets of parameter values. Also, two such indistinguishable sets of parameter values may lead to quite different interpretations.” Osmond and Gardner [4] introducing a mathematical constraint in the model, were able to identify one of these possible solutions, which found wide application for the analysis of mortality data [3]. The researcher must be aware that the `apc` function estimates the parameters of the “full” age-period-cohort model according to this solution.

Owing to the programming capability of R as an object oriented language, the researcher has fully access to the output of the main function (apc) which, in addition to the estimates of the age-period-cohort model, returns all the models fitted within the function. It is so possible to critically evaluate the goodness of fit of these models. We hope that, owing to the diffusion of R in health research, this package could be useful in the analysis of cohort studies.

## References

[1] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from: http://www.R-project.org/. 2013.

[2] Baker RJ, Nelder JA. The Glim System, Release 3., Oxford, Numerical Algorithms Group. 1978.

[3] Decarli A, La Vecchia C. Age, period and cohort models: a review of knowledge and application in GLIM. Rivista di Statistica Applicata 1987;20:392-410.

[4] Osmond C, Gardner MJ. Age, period and cohort models applied to cancer mortality rates. Stat Med 1982; 1(3): 245-59.

[5] Clayton D, Schifflers E. Models for temporal variation in cancer rates. I: Age-period and age-cohort models. Stat Med 1987; 6(4): 449-67.

[6] Clayton D, Schifflers E. Models for temporal variation in cancer rates. II: Age-period-cohort models. Stat Med 1987; 6(4): 469-81.

[7] Malvezzi M, Bertuccio P, V. E. Age period color analysis of cancer mortality data: methods and application to Italian male mortality data for gastric cancer and cancers of the oral cavity and pharynx. BioMedical Statistics and Clinical Epidemiology 2009;3:97-105.

[8] Armitage P, Doll R. The age distribution of cancer and a multi-stage theory of carcinogenesis. Br J Cancer 2004; 91(12): 1983-9.

[9] La Vecchia C, Decarli A, Pagano R. Patterns of smoking initiation in Italian males and females from 1955 to 1985. Prev Med 1995; 24(3): 293-6.

[10] Gallus S, Lugo A, La Vecchia C, et al. Pricing Policies And Control of Tobacco in Europe (PPACTE) project: cross-national comparison of smoking prevalence in 18 European countries. Eur J Cancer Prev 2014; 23(3): 177-85.

[11] Gallus S, Lugo A, Colombo P, Pacifici R, La Vecchia C. Smoking prevalence in Italy 2011 and 2012, with a focus on hand-rolled cigarettes. Prev Med 2013; 56(5): 314-8.

[12] Peto R, Darby S, Deo H, Silcocks P, Whitley E, Doll R. Smoking, smoking cessation, and lung cancer in the UK since 1950: combination of national statistics with two case-control studies. BMJ 2000;3 21(7257): 323-9.

[13] Negri E, La Vecchia C, D'Avanzo B, Nobili A, La Malfa RG. Acute myocardial infarction: association with time since stopping smoking in Italy. GISSI-EFRIM Investigators. Gruppo Italiano per lo Studio della Sopravvivenza nell'Infarto. Epidemiologia dei Fattori di Rischio dell'Infarto Miocardico. J Epidemiol Community Health 1994; 48(2): 129-33.

[14] Negri E, La Vecchia C, Franzosi MG, Tognoni G. Attributable risks for nonfatal myocardial infarction in Italy. GISSI-EFRIM investigators. Gruppo Italiano per lo Studio della Sopravvivenza nell'Infarto Miocardico. Epidemiologia dei Fattori di Rischio dell'Infarto Miocardico. Prev Med 1995; 24(6): 603-9.

[15] Yusuf S, Hawken S, Ounpuu S, et al. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. Lancet 2004; 364(9438): 937-52.

[16] Levi F, Chatenoud L, Bertuccio P, Lucchini F, Negri E, La Vecchia C. Mortality from cardiovascular and cerebrovascular diseases in Europe and other areas of the world: an update. Eur J Cardiovasc Prev Rehabil 2009; 16(3): 333-50.

*