CALCULATING AND ILLUSTRATING THE PROBABILITY OF DEVELOPING CANCER USING SAS® AND SAS/GRAPH® SOFTWARE

Mike Zdeb and Matt Dairman, University at Albany-School of Public Health

BACKGROUND

Cancer is a disorder of uncontrolled, abnormal cell growth in which a cell is altered is such a manner as to continually, but inappropriately, replicate itself producing millions of similarly altered abnormal cells (1). Cancer is usually expressed as a swelling or tumor, a tumor being a mass of cells that has undergone a series of changes causing the cells to be unresponsive to normal growth controlling mechanisms. Of the three types of tumors (i.e. benign, in-situ, and malignant), malignant tumors are of the most concern, given their ability to destroy local, normal cells and then spread to other parts of the body to continue their invasion and destruction of normal cells.

Various statistics are used to estimate the public health impact of cancer. One of these is the cancer mortality rate, defined as the number of deaths to due cancer in a given population over a specified time period. The yearly cancer mortality rate in the United States is approximately 200 deaths per 100,000 population, with one-in-four deaths a result of cancer. Another measure is cancer incidence, or the number of cases of cancer that develop (or are detected) in a given population over a specified time period. Cancer incidence in the United States is approximately 475 cases detected per 100,000 population per year.

Mortality and incidence are usually based upon data from a given calendar year. They both provide a 'snap shot' of cancer risk. A third measure that combines cancer mortality and incidence with the competing risk of dying of diseases other than cancer is cancer probability, or the chance of developing cancer within a given age interval, e.g. lifetime probability is the chance at birth of developing cancer over one's lifetime. This paper shows how SAS data step programming and SAS/GRAPH procedures are used to calculate then illustrate the probability of developing cancer.

METHODS AND MATERIALS

In 1956, Goldberg used life table methodology to calculate the probability of developing cancer (2). Zdeb used Goldberg's methods to update cancer probability in a paper published in 1977 (3). Though other methods have been proposed since Zdeb's 1977 paper, many researchers have based subsequent cancer probability estimates on the life table methodology used by Goldberg and Zdeb. The method starts with the construction of a current life table (4,5). Then, a series of equations are applied to cancer mortality and incidence data, resulting in the final data used to compute cancer probabilities - a hypothetical cancer-free population at five-year age intervals starting at birth, and the number of cases of cancer expected to develop in this population within a given number of years.

To construct a current life table, 1990 census population estimates for upstate New York (i.e. New York State exclusive of New York City) were combined with mortality data obtained from death certificates for upstate New York residents for the period 1989 through 1991. Probability estimates were made using the life table together with cancer mortality data, again for upstate New York residents based on death certificates, and cancer incidence data from the New York State cancer registry. In New York State, cancer is, by law, reported to the registry upon discovery by physicians, hospitals, and/or the laboratory where cancer specimens are examined. Just as with the mortality data, cancer incidence data for 1989 through 1991 for upstate New York residents were used. Populations, deaths, and cancer cases were all used within five year age intervals, starting at the interval birth to four years of age, then five through nine, ten through fourteen, etc., through a final age interval of eighty-five plus years.

The SAS code is used to construct a current life table is shown in the appendix. It is based on the equations derived by Chiang (4,5). A current life table starts with a hypothetical population (commonly 100,000) and then uses the mortality experience of a given population to compute the proportion of people that die in each given age interval. Normally, the output of interest from a current life table is 'e,' the expected years of life to be lived either at birth or at some subsequent starting age. However, when using the life table as a step in calculating the probability of developing cancer, there are five other quantities of interest from the life table: L, l, q, d, and M.

(Note: The meaning of each of these symbols can be found in Chiang (4,5), as can the set of equations that are implemented via the SAS code in the appendix).

The method devised by Goldberg uses the five life table statistics and data on cancer incidence and mortality to compute two new values at each age interval. The first is a hypothetical population that is cancer-free at the start of any given age interval. This is a value similar to the starting value in a current life table. In a current life table, the starting value for the population is decreased based on the mortality experience of the population, derived from death rates. In the Goldberg methodology, the cancer-free population is decreased by a combination of cancer cases, cancer mortality, and non-cancer mortality. The second value is the number of cancer cases expected to occur in the hypothetical population. Given the combination of cases and a cancer free population, one can compute the probability of developing cancer, i.e. cancer cases per cancer-free population that develop in a given age interval. The SAS code used to implement the Goldberg methodology is shown in the appendix, as is the probability matrix produced by the SAS code.

Once the cancer-free population at each age interval is calculated, it can be used together with site-specific cancer incidence to compute the probability of developing various

cancers, e.g. lung, breast, prostate, etc. The number of cancer cases at any given age interval is merely the product of the life table quantity L and the cancer incidence.

(Note: The complete equations for probability calculation that are implemented with the SAS code in the appendix can be found in either the original paper by Goldberg (2) or in the update by Zdeb (3). As with the equations associated with life table computation, a complete presentation of the equations was deemed too detailed for this paper).

RESULTS

Probabilities...

Using the method just described and data from the time period 1989-1991, the probability (x 100) at birth of a male developing cancer during his lifetime is 52.9, and 48.7 for a female. These probabilities assume a long life, i.e. one extending past 85 years. The probability at birth of a male developing cancer by age 65 drops to 16.1, and to 18.7 for a female. The female probability through age 65 is higher than that of males since the highest probability site for females is breast cancer, a cancer that is detected earlier than the highest probability male site, i.e. prostate cancer.

Illustrating Probabilities...

The probability matrix shown in the appendix contains a lot of information that may be difficult to assimilate. If one wants to compare male to female probabilities, or sex-specific probabilities calculated during two different time periods, charts produced using SAS/GRAPH offer an easier option than comparing the numbers shown in probability matrices.

The appendix contains SAS code that uses PROC GPLOT to produce a chart that shows the cancer probabilities. All the data contained in the probability matrix is shown in the chart. The only data necessary to produce the a chart with the SAS code is the cancer-free population and the number of cases of cancer that develop at various age intervals. The SAS code contains a number of data steps, some of which set up the annotate data sets that are used to draw reference lines and label the charts, others that are used to scan the data and eliminate from plotting any data points (probabilities x 100) that fall below a value of .02. Not only are points eliminated, but also any line that contains a point that falls below .02.

A plot is shown in the appendix that compares male and female probabilities for the period 1989-1991. It is labeled 'TOTAL (1990)' to indicate that the plots represent the probability of developing any type of cancer (i.e. all sites combined). The side-by-side plot was produced using PROC GREPLAY to rescale and reposition the results of two separate plots produced with the SAS code shown in the appendix.

REFERENCES

1 Doll R, Peto R: The causes of cancer: Quantitative estimates of avoidable risks of cancer in the United States today. J Natl Cancer Inst 1981, 66(6):1192-1308

2 Goldberg ID, et al: The probability of developing cancer. J Natl Cancer Inst 1956, 17(2):155-173

3 Zdeb MS: The probability of developing cancer. Am J Epid 1977, 106(1):6-16

4 Chiang CL: *Introduction to Stochastic Processes in Biostatistics*, New York, Wiley, 1968

5 Chiang CL: On constructing current life tables. J Am Stat Assn 1972, 67(339):538-541

(Note: All of the SAS code shown in the appendix is available upon request from Mike Zdeb. Just send a note to msz03@health.state.ny.us or a FAX to 518/473-2015. In addition to the SAS code on a diskette (or it can be sent via e-mail), you will be provided with an explanation as to how the SAS code for the life table and probability computation relate to the equations in the original papers by Chiang and Goldberg.)

APPENDIX

Life Table Calculation...

The following SAS code uses the male deaths and male populations (see the data step that creates dataset DFILE) to compute a current life table for male residents of upstate New York for the period 1989-1991. The naming conventions used for the variables corresponds to the variable names used by Chiang (4,5). Where the prefix "lg" is used, it corresponds to LARGE, e.g. the variable "lgl" is large l from the Chiang equations, while variable "l" is small l...

```
*format used to label PROC PRINT output,
the 18 age groups used;
proc format;
value $age
"01" = "  0-4"  "02" = "  5-9"
"03" = "10-14"  "04" = "15-19"
"05" = "20-24"  "06" = "25-29"
"07" = "30-34"  "08" = "35-39"
"09" = "40-44"  "10" = "45-49"
"11" = "50-54"  "12" = "55-59"
"13" = "60-64"  "14" = "65-69"
"15" = "70-74"  "16" = "75-79"
"17" = "80-84"  "18" = "  85+";
run;

*male deaths and population plus
values for 'a' - a variable used in the
computation of the life table;
data dfile;
input agegrp $ deaths pop a;
label
agegrp = "age interval (x to x+1)"
deaths = "total number of deaths (D)"
pop    = "total population (P)"
a      = "fraction (a)";
datalines;
01  2535 1170597 0.35
02   229 1106193 0.46
03   267 1063398 0.54
04   866 1179306 0.57
05  1449 1283256 0.49
06  1712 1310862 0.50
07  2144 1342959 0.52
08  2730 1231836 0.54
09  3188 1136298 0.54
```

```
10  3811  904464 0.54
11  4796  743877 0.53
12  7498  707178 0.52
13 12416  710469 0.52
14 17057  617040 0.52
15 20007  456948 0.51
16 21086  312015 0.51
17 18210  176460 0.48
18 21036  110823 0.98
;
run;

*use the mortality dataset to compute the
life table;
data lifetab;
retain n 5;
set dfile end=last;
M = round((deaths/(pop)),.000001);
q = (n*M)/(1+((1-a)*n*M));
if last then q=1;
label M = "total death rate (M)"
      q = "total proportion dying (q)";
run;

*compute the life table - make extensive
use of temporary arrays;
data life (keep=age deaths pop sml smd smq
sma bgm bgl bgt sme rename=(age=agegrp
smq=q sml=l smd=d sma=a bgm=m bgl=lgl
bgt=lgt sme=e));
array agegrp(18) $ _temporary_;
array l(18)       _temporary_;
array d(18)       _temporary_;
array q(18)       _temporary_;
array a(18)       _temporary_;
array m(18)       _temporary_;
array n(18)       _temporary_;
array lgl(18)     _temporary_;
array lgt(18)     _temporary_;
array e(18)       _temporary_;

retain nrec 0 agegrp l d q a m lgl lgt e;
set lifetab (rename=(agegrp=ageold q=qold
a=aold m=mold n=nold))end=last;
nrec+1;
agegrp(nrec)=ageold;
q(nrec)=round(qold,.00001);
a(nrec)=aold;
m(nrec)=mold;
n(nrec)=nold;
lgl(_n_)=1;

if last then do;
 l(1)=100000;
 do j=1 to 17;
  d(j)=round(q(j)*l(j),1.);
  l(j+1)=round(l(j)-d(j),1.);
  lgl(j)=int((n(j)*(l(j)-d(j)))
             +(a(j)*n(j)*d(j)));
 end;

*at this point, j is equal to 18;
 l(j)   = l(j-1) - d(j-1);
 d(j)   = l(j) * q(j);
 lgl(j) = (n(j)*(l(j)-d(j)))+
          (a(j)*n(j)*d(j));

 do j=1 to 18;
  lgt(1)+lgl(j);
 end;
 e(1)=round(lgt(1)/l(1),.01);
 do j=2 to 18;
  lgt(j)=round(lgt(j-1) - lgl(j-1),1.);
  e(j)=round(lgt(j)/l(j),.01);
 end;
 do j=1 to 18;
  age = agegrp(j);
  smq = q(j);   sml = l(j);
  smd = d(j);   sma = a(j);
  bgm = m(j);   bgl = lgl(j);
  bgt = lgt(j); sme = e(j);
  output;
 end;
stop;
end;
run;

title "1989-1991 ABRIDGED LIFE TABLES FOR
UPSTATE NEW YORK - MALES";
proc print data=life noobs;
var agegrp q l d a lgl lgt e;
format agegrp $age. lgt lgl 12.;
run;
```

Cancer Probability Calculation...

The following SAS code uses values computed during the completion of the life table, plus data on total cancer incidence and mortality to compute and print a probability matrix that shows the probability (x 100) of developing cancer for male residents of upstate New York for the period 1989-1991. In addition to the probability matrix, the hypothetical 'cancer-free' population at each age interval is computed and stored in dataset CANCFREE..

```
*used to label probability matrix;
proc format;
value i2age
 1 = " 0 "   2 = " 5 "   3 = "10 "
 4 = "15 "   5 = "20 "   6 = "25 "
 7 = "30 "   8 = "35 "   9 = "40 "
10 = "45 "  11 = "50 "  12 = "55 "
13 = "60 "  14 = "65 "  15 = "70 "
16 = "75 "  17 = "80 "  18 = "85+";
value j2age
 1 = "  5"   2 = " 10"   3 = " 15"
 4 = " 20"   5 = " 25"   6 = " 30"
 7 = " 35"   8 = " 40"   9 = " 45"
10 = " 50"  11 = " 55"  12 = " 60"
13 = " 65"  14 = " 70"  15 = " 75"
16 = " 80"  17 = " 85"  18 = "85+";
run;

*data from current life table;
data life;
input agegrp $ m q l d lgl;
datalines;
01 .002166 0.01075 100000  1075 496506
02 .000207 0.00103  98925   102 494349
03 .000251 0.00125  98823   124 493829
04 .000734 0.00366  98699   361 492718
05 .001129 0.00563  98338   554 490277
06 .001306 0.00651  97784   637 487327
07 .001596 0.00795  97147   772 483882
08 .002216 0.01102  96375  1062 479432
09 .002806 0.01394  95313  1329 473508
10 .004214 0.02087  93984  1961 465409
11 .006447 0.03175  92023  2922 453248
12 .010603 0.05170  89101  4607 434448
13 .017476 0.08386  84494  7086 405463
14 .027643 0.12962  77408 10034 362958
15 .043784 0.19771  67374 13321 304233
16 .067580 0.28990  54053 15670 231873
17 .103196 0.40682  38383 15615 151316
18 .189816 1.00000  22768 22768 111563
;
run;

*cancer incidence and mortality;
```

```
data cancer;
input agegrp $ deaths cases pop;
*cancer incidence;
ci = cases / pop;
*cancer mortality;
cm = deaths / pop;

datalines;
01   34   362  1170597
02   27   168  1106193
03   35   167  1063398
04   49   270  1179306
05   77   466  1283256
06  125   742  1310862
07  208  1023  1342959
08  322  1359  1231836
09  558  1912  1136298
10 1014  2608   904464
11 1539  3872   743877
12 2631  6261   707178
13 4445 10450   710469
14 5801 13400   617040
15 6089 13779   456948
16 5569 11300   312015
17 3977  7377   176460
18 3086  5039   110823
;
run;

data combined;
merge life cancer;
by agegrp;
numcase = ci * lgl;
run;

data cancfree (keep=age2 slp
rename=(age2=agegrp slp=lp));
length title $ 100;
array age(18) $  _temporary_;
array md(18)     _temporary_;
array tp(18)     _temporary_;
array dt(18)     _temporary_;
array dp(18)     _temporary_;
array dc(18)     _temporary_;
array qt(18)     _temporary_;
array qp(18)     _temporary_;
array qc(18)     _temporary_;
array lt(18)     _temporary_;
array lp(18)     _temporary_;
array lc(18)     _temporary_;
array ll(18)     _temporary_;
array mt(18)     _temporary_;
array mc(18)     _temporary_;
array k(18)      _temporary_;
array c(18)      _temporary_;
array dpcorr(18) _temporary_;
array dccorr(18) _temporary_;

do j=1 to 18;
set combined;
age(j)= agegrp; md(j) = deaths;
tp(j) = pop;    lt(j) = l;
dt(j) = d;      ll(j) = lgl;
qt(j) = q;      mt(j) = m;
c(j)  = numcase;
end;

lp(1) = 100000;
lc(1) = 0;

do j=1 to 17;
 mc(j) = md(j) / tp(j);
 dc(j) = dt(j) * (mc(j) / mt(j));
 dp(j) = dt(j) - dc(j);
 qc(j) = dc(j) / lt(j);
 qp(j) = dp(j) / lt(j);
 k(j)  = qp(j)*(lc(j)
         + (c(j)/2) - (dc(j)/2));
 dpcorr(j) = dp(j) - k(j);
 dccorr(j) = dc(j) + k(j);
 lp(j+1) = lp(j) - c(j) - dpcorr(j);
 lc(j+1) = lt(j) - lp(j);
end;

mc(18) = md(18) / tp(18);
dc(18) = dt(18) * (mc(18) / mt(18));
dp(18) = dt(18) - dc(18);
qc(18) = dc(18) / lt(18);
qp(18) = dp(18) / lt(18);

k(18)  = qp(18)*(lc(18)
         + (c(18)/2) - (dc(18)/2));
dpcorr(18) = dp(18) - k(18);
dccorr(18) = dc(18) + k(18);

file "c:\mprob90.all";

put "Probabilities (x100) of Developing
Cancer in New York";
put "Males - 1989-1991" /;
put @1 "Current" @61 "Develop Cancer by
Age" / @1 "Age" / +6 @;
do j=1 to 18;
 put +5 j j2age. @;
end;
put / 150*"-";
do i=1 to 18;
 over=((i-1)*8);
 put i i2age. +3 +over @;
 do j=i to 18;
  cases = 0;
 do h=i to j;
  cases + c(h);
 end;
  prob = 100 * cases / lp(i);
  put prob 8.2 @;
 end;
 put;
end;

do j=1 to 18;
  age2 = age(j);
  slp  = round(lp(j),1.);
  output;
end;
stop;
run;
```

Illustrating Probabilities...

The following SAS code uses PROC GPLOT and annotate data sets to illustrate the cancer probabilities computed using the Goldberg method and shown in the matrix at the end of the paper. Two quantities that were determined in the previous code, i.e. the cancer free population in each age group and the cancer cases expected to develop within each age group, are input as variables mpop and m0 in the data step that creates dataset males. These two variables are then used to compute the various probabilities that are illustrated in the plot. The plot allows one to pick any starting age (on the x-axis) and any ending age (one of the curved lines), and then read the cancer probability for the specified start/end age interval from the y-axis...

```
symbol I=j r=18 c=black;

axis1
logbase=10 logstyle=expand
```

```
order = (.010 .100 1.000 10.000 100.00)
label = (a=90 r=0 "PROBABILITY x 100");

axis2
label=("AGE AT START OF INTERVAL")
value=(tick=1  "0"  tick=2  " "
       tick=3  "10" tick=4  " "
       tick=5  "20" tick=6  " "
       tick=7  "30" tick=8  " "
       tick=9  "40" tick=10 " "
       tick=11 "50" tick=12 " "
       tick=13 "60" tick=14 " "
       tick=15 "70" tick=16 " "
       tick=17 "80" tick=18 " ")
minor=none;

*this data step uses one of the ANNOMAC set
of macros that are supplied with SAS/GRAPH
- the %line macro allows one to draw a line
by specifying a start and end position;
data lines;
retain xsys ysys '2' hsys '3' when 'b';
%line(1,000.010,18,000.010,black,2,.05);
%line(1,000.050,18,000.050,black,2,.05);
%line(1,000.100,18,000.100,black,2,.05);
%line(1,000.500,18,000.500,black,2,.05);
%line(1,001.000,18,001.000,black,2,.05);
%line(1,005.000,18,005.000,black,2,.05);
%line(1,010.000,18,010.000,black,2,.05);
%line(1,050.000,18,050.000,black,2,.05);
%line(1,100.000,18,100.000,black,2,.05);
run;

data males;
input mpop m0;
datalines;
100000     153.1
 98787      75.1
 98622      77.6
 98437     112.8
 97984     178.0
 97284     275.8
 96420     368.6
 95359     528.9
 93902     796.3
 92025    1342.0
 89271    2359.2
 84985    3847.6
 78277    5966.6
 68075    7883.4
 54293    9173.3
 37489    8398.3
 20804    6325.8
  8004    5071.6
;
run;

data new (keep=I j prob);
array lp(18) _temporary_;
array  c(18) _temporary_;
array  p(18) _temporary_;
do j=1 to 18;
  set males0;
  lp(j) = mpop;
  c(j)  = m0;
end;
do i=1 to 18;
do j=I to 18;
 cases = 0;
do h=i to j;
 cases + c(h);
end;
 prob = 100 * cases / lp(I);
 output;
end;
end;
run;

proc sort data=new;
by j;
run;

data new (keep=I j prob);
array p(18) _temporary_;
do j=1 to 18;
 check=0;
do i=1 to j;
 set new;
 if prob le .020 then check=1;
 p(i)=prob;
end;
if check=0 then do i=1 to j;
 prob=p(i);
 output;
end;
end;


data xnew;
set new;
by j;
if i ne 1;
if last.j then output;
run;

*determine where to write the age labels on
the curved lines and the text 'AGE AT
....;';
data labels;
length text $ 50;
retain xsys ysys '2'
       function 'label'
       when 'a'
       position 'D'
set xnew end=last;
if i ne 18 then text = put((5*(i-1)),3.);
else            text = "85+";
x = i;
y = prob;
output;
if i eq 12 then do;
 x = i + .5;
 prob = prob - (prob/5);
 position = "9";
 text = "AGE AT END OF INTERVAL";
 output;
 position = "d";
end;
if last then do;
 position = "6";
 x = 1;
 y = 85;
 size = 1.5;
 text = "MALES - TOTAL";
 output;
end;
run;

data labels;
 set labels lines;
run;

proc gplot data=new;
 plot prob*I=j/vaxis=axis1 haxis=axis2
                nolegend annotate=labels;
run;
quit;
```

Cancer Probability Matrix and Probability Plot...

The following matrix and plot show the probability (x 100) of a male resident of upstate New York developing cancer during the period 1989-1991. The lifetime probability is 52.93, i.e. the probability at birth (age 0) through age 85+. The matrix (or plot) can be used to find the probability in any given age interval, e.g. the probability of developing cancer by age 65 if you are already 45 years of age is 14.68. The plot also shows the female probabilities for various start/end age combinations. The lifetime probability at birth (age 0 to age 85+) for females can be read on the left axis at the intersection of the topmost line as approximately 49.

Probabilities (x100) of Developing Cancer in New York
Males - 1989-1991

Current Age	Develop Cancer by Age																	
	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	85+
0	0.15	0.23	0.31	0.42	0.60	0.87	1.24	1.77	2.57	3.91	6.27	10.11	16.08	23.96	33.13	41.53	47.86	52.93
5		0.08	0.15	0.27	0.45	0.73	1.10	1.64	2.44	3.80	6.19	10.08	16.12	24.10	33.39	41.89	48.29	53.43
10			0.08	0.19	0.37	0.65	1.03	1.56	2.37	3.73	6.12	10.02	16.07	24.06	33.37	41.88	48.29	53.44
15				0.11	0.30	0.58	0.95	1.49	2.30	3.66	6.06	9.96	16.02	24.03	33.35	41.88	48.31	53.46
20					0.18	0.46	0.84	1.38	2.19	3.56	5.97	9.89	15.98	24.03	33.39	41.96	48.41	53.59
25						0.28	0.66	1.21	2.03	3.40	5.83	9.78	15.91	24.02	33.45	42.08	48.58	53.80
30							0.38	0.93	1.76	3.15	5.60	9.58	15.77	23.95	33.46	42.17	48.73	53.99
35								0.55	1.39	2.80	5.27	9.30	15.56	23.82	33.45	42.25	48.89	54.20
40									0.85	2.28	4.79	8.89	15.24	23.63	33.40	42.34	49.08	54.48
45										1.46	4.02	8.20	14.68	23.25	33.22	42.34	49.22	54.73
50											2.64	6.95	13.63	22.46	32.74	42.14	49.23	54.91
55												4.53	11.54	20.82	31.61	41.49	48.94	54.91
60													7.62	17.69	29.41	40.14	48.22	54.70
65														11.58	25.05	37.39	46.68	54.13
70															16.90	32.36	44.01	53.35
75																22.40	39.27	52.80
80																	30.40	54.78
85+																		63.36