Recent developments in methods for the analysis of population-based cancer survival

Paul W. Dickman
Department of Medical Epidemiology
Karolinska Institutet
Stockholm, Sweden
paul.dickman@mep.ki.se

Department of Medical Epidemiology
January 14, 2003

Overview

- Two 'new' developments are causing excitement among practitioners of population-based cancer survival
  - Estimating relative survival using a period (as opposed to cohort) approach
  - New approaches to modelling relative survival (modelling excess mortality)
- These 'new' developments have been familiar to epidemiologists for decades.

Key concepts in population-based cancer survival analysis

- Net survival — the proportion of patients who would have survived t years or more following diagnosis in the hypothetical situation where the disease of interest were the only possible cause of death.
  - Net survival is a hypothetical quantity which can be estimated using, e.g., the cause-specific survival proportion or the relative survival ratio.
- Cause-specific survival is the analog of cause-specific mortality — only those deaths which can be attributed to the cancer in question are considered to be events, while all other deaths are considered censorings.
- Relative survival is the analog of excess mortality — the relative survival ratio (RSR) is defined as the observed survival in the patient group divided by the expected survival of a comparable group from the general population.
- The concept of excess mortality and the use of relative excess risk as a measure of association are familiar to epidemiologists.

Cervical cancer diagnosed in New Zealand 1994 – 2001 Survival estimates stratified by age at diagnosis

		1-year			5-year		
Age	N	Obs.	Exp.	Relative	Obs.	Exp.	Relative
0–44	708	93.6	99.9	93.7	84.0	99.5	84.4
45–54	315	87.6	99.7	87.9	66.6	98.2	67.8
55–64	205	84.4	99.3	85.0	57.9	95.6	60.5
65–74	197	80.7	98.2	82.2	45.6	89.8	50.7
75+	134	59.0	93.3	63.2	28.4	68.9	41.2
All	1559	86.6	99.0	87.5	67.2	95.8	70.2

Cervical cancer diagnosed in New Zealand 1994 – 2001 Life table estimates of survival

Women diagnosed Jan 1994 - June 2001 with follow-up to June 2002

					Interval-specific	Cumulative		Interval-specific	Cumulative
				Effective number	observed	observed	Cumulative expected	relative	relative
I	N	D	W	at risk	survival	survival	survival	survival	survival
1	1559	209	0	1559.0	0.86594	0.86594	0.98996	0.87472	0.87472
2	1350	125	177	1261.5	0.90091	0.78014	0.98192	0.90829	0.79450
3	1048	58	172	962.0	0.93971	0.73310	0.97362	0.94772	0.75296
4	818	32	155	740.5	0.95679	0.70142	0.96574	0.96459	0.72630
5	631	23	148	557.0	0.95871	0.67246	0.95766	0.96679	0.70218
6	460	10	130	395.0	0.97468	0.65543	0.94972	0.98284	0.69013
7	320	5	129	255.5	0.98043	0.64261	0.94198	0.98848	0.68219
8	186	3	134	119.0	0.97479	0.62641	0.93312	0.98405	0.67130
9	49	1	48	25.0	0.96000	0.60135	0.91869	0.97508	0.65457

Estimation using a period approach

- The life tables typically used in cancer survival analysis are what demographers refer to a cohort life tables.
- In demography, a cohort life table is constructed by following all individuals born during one time period until all have died and keeping track of how many have died at different ages.
- A period life table (or static life table) uses information on individuals alive in each age class at one cross-section of time.
- For example, the expectation of life is calculated as if one person lived through all ages with age-specific survival probabilities of the current year.
- In cancer epidemiology, the incidence proportion is calculated in the same way from cross-sectional age-specific incidence rates.

Cervical cancer diagnosed in New Zealand 1994 – 2001 Period estimates of survival for Jan. 2000 – Dec. 2001

Interval	N	D	W	Interval-specific relative survival	Cumulative relative survival
0.0 - 1.0	510	41	91	0.90347	0.90347
1.0 - 2.0	542	25	180	0.94052	0.84973
2.0 - 3.0	490	16	180	0.95929	0.81513
3.0 - 4.0	442	12	139	0.96896	0.78983
4.0 - 5.0	425	11	142	0.96980	0.76598
5.0 - 6.0	398	7	138	0.98151	0.75182
6.0 - 7.0	253	4	126	0.98665	0.74178
7.0 - 8.0	123	1	122	0.99388	0.73724

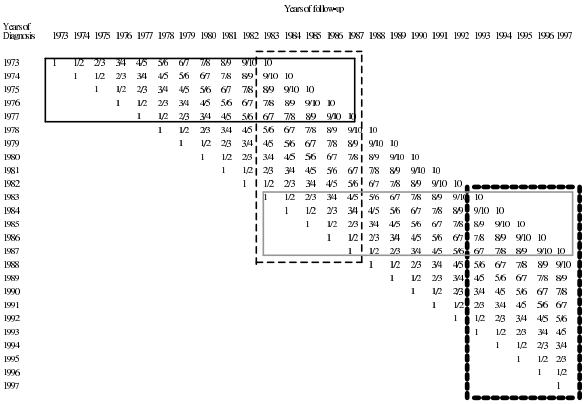
The approach was heavily criticised when first suggested

- A potential problem with the period method is that estimates may be too optimistic if, for example, survival improves during the first year following diagnosis but this is offset by a decrease in survival during subsequent years.

Table 1: Hypothetical data illustrating a potential problem with the period method. The table shows interval-specific survival estimates.

Interval	cohort 1991–1995 (fu to 2000)	cohort 1996–2000 (fu to 2005)	period 1996–2000 (fu to 2000)
1	0.54	0.73	0.73
2	0.82	0.77	0.78
3	0.88	0.85	0.87
4	0.97	0.89	0.95
5	0.98	0.95	0.98
II	0.37	0.40	0.46

- Studies have shown that this is not a concern in practice [1, 2, 3].
- Some people find it difficult to characterize the set of patients to whom the rates refer — demographic (index) thinking is required.
- The period approach detects trends in patient survival earlier than the cohort approach.
- Brenner and Hakulinen [3] demonstrated that, at least for Finnish data, period analysis is a useful tool for predicting long-term survival of cancer patients using the most up-to-date data available.
- Relative survival estimates made using the period approach can be modelled in the usual manner.



Modelling relative survival

- The hazard at time since diagnosis t for persons diagnosed with cancer (with covariate vector $\mathbf{z}$) is modelled as the sum of the known baseline hazard, $\lambda^*(t;\mathbf{z})$, and the excess hazard due to a diagnosis of cancer, $\nu(t;\mathbf{z})$. That is,

$$\lambda(t;\mathbf{z}) = \lambda^*(t;\mathbf{z}) + \nu(t;\mathbf{z}). \quad (1)$$

- We are interested in modelling the excess mortality. Excess mortality is estimated as the difference between the observed and expected mortality.
- The excess hazards are assumed to be constant within subintervals (bands) of follow-up time.
- It is generally assumed that the excess hazard component, ν, is a multiplicative function of the covariates, written as $\exp(\mathbf{x}\beta)$.

$$\lambda(\mathbf{x}) = \lambda^*(\mathbf{x}) + \exp(\mathbf{x}\beta). \quad (2)$$

- In Equation 2, follow-up time, t, is incorporated into the covariate vector $\mathbf{x}$.
- Implicit in Equation 2 is the assumption that the excess hazards for any two patient subgroups are proportional over follow-up time.
- Non-proportional excess hazards can, however, be incorporated by introducing follow-up by covariate interaction terms.
- Two approaches to estimating the model have previously been described by Hakulinen and Tenkanen [4] and Estève *et al.* [5].

Survival of women diagnosed with cervical cancer in New Zealand 1994–2001

		RER	95% CI	
FU	1	1.000	(reference)	
	2	0.940	0.740	1.195
	3	0.567	0.405	0.793
	4	0.443	0.291	0.674
	5	0.390	0.228	0.669
	6	0.216	0.087	0.536
	7	0.141	0.036	0.547
AGE	0-44	1.000	(reference)	
	45-54	1.540	1.149	2.064
	55-64	1.759	1.286	2.405
	65-74	1.889	1.380	2.585
	75+	2.835	1.986	4.046

		RER	95% CI	
STAGE	distant	174.40	70.993	428.42
	regional	30.092	11.935	75.876
	local spread	15.849	5.758	43.620
	localised	1.000	(reference)	
	not stated	39.785	16.495	95.955
YYDX	1994-1997	1.399	1.121	1.744
	1998+	1.000	(reference)	
MORPH	adenocarcinoma	1.340	1.048	1.713
	other	2.030	1.439	2.864
	squamous cell	1.000	(reference)	

		RER	95% CI	
ETHN	Maori	2.213	1.721	2.847
	Pacific Island	1.653	1.054	2.594
	other	0.671	0.370	1.217
	European	1.000	(reference)	
NZDEP	>=1060	0.844	0.613	1.163
	>=1000 <1060	0.919	0.681	1.241
	>=950 <1000	1.035	0.752	1.424
	<950	1.000	(reference)	

The Estève *et al.* full likelihood approach

- Estève *et al.* [5] described a method for estimating the model in Equation 2 directly from individual-level data using a maximum likelihood approach.
- Although they use a slightly different parameterisation in their paper, the underlying model is identical to Equation 2.
- The likelihood function is

$$L = \prod_{i=1}^{n} \exp(-\int_0^{t_i} \lambda(s)\,ds)[\lambda(t_i)]^{\delta_i}, \quad (3)$$

where t_i is the survival time and δ_i the failure variable (1 if t_i is the time of death; 0 if the survival time is censored at t_i) for each of the $i = 1, \ldots, n$ individuals.

- Writing the total hazard as the product of the expected hazard and the excess hazard, the log-likelihood function is

$$l(\beta) = -\sum_{i=1}^{n} \int_0^{t_i} \lambda^*(s)\, ds - \sum_{i=1}^{n} \int_0^{t_i} \nu(s)\, ds + \sum_{i=1}^{n} \delta_i \ln[\lambda^*(t_i) + \nu(t_i)] \quad (4)$$

  where $\nu = \exp(\mathbf{x}\beta)$.

- The first component of the log likelihood does not depend on β leading to the attractive feature from a computational viewpoint that, for each individual, only one value need be read from the excess hazards file, the expected hazard at t_i.
- Their approach is implemented in special-purpose software which runs under DOS [6].
- A major problem in applying the Estève *et al.* approach in practice is that it is not possible, using the accompanying software [6], to model time varying covariates.

- This means that there is no way to control for non-proportional excess hazards, which are very common with cancer registry data.
- Regression diagnostics are not available for the Estève *et al.* model and there is no way of assessing goodness-of-fit.

Full likelihood based on multiple observations per subject

- We first split the data to obtain separate observations for each subject-band, as is often done when analysing epidemiological cohort studies using Poisson regression.
- Each subject-band observation includes variables representing the time at risk (y), death indicator (δ), expected hazard (λ^*), and indicator variables for each of the components of β (including follow-up band).
- The log likelihood function, expressed in terms of the J subject-band observations, is

$$l(\beta) = \sum_{j=1}^{J} \left[d_j \ln[\lambda^*(\mathbf{x}_j) + \exp(\mathbf{x}_j\beta)] - y_j \exp(\mathbf{x}_j\beta) \right]. \quad (5)$$

- The log likelihood in Equation 4 is written as the sum of the contributions from each subject

- We are simply partitioning this into the sum of the the contributions from each subject-band.
- The model can be estimated using procedures available in standard statistical software packages for maximum likelihood estimation, such as the Stata `ml` command or SAS PROC NLP (part of SAS/OR).
- The estimates presented in the table on slide 27, for example, were obtained using the following SAS commands

```
proc nlp data=individual cov=2 vardef=n;
max loglike;
parms int fu_2 fu_3 fu_4 fu_5 female year2 age2 age3 age4;
theta = int+fu_2*fu2+fu_3*fu3+fu_4*fu4+fu_5*fu5+year2*year8594
        +age2*age_gr2+age3*age_gr3+age4*age_gr4+female*sex2;
loglike = d*log(-log(p)+exp(theta))-y*exp(theta);
run;
```


- In Stata, the model is defined using the following ado file (`esteve.ado`)

```
program define esteve
version 7
args lnf theta
qui replace `lnf'=-exp(`theta')*y if $ML_y1==0
qui replace `lnf'=ln(r+exp(`theta'))-exp(`theta')*y if $ML_y1==1
end
```

- Then to fit the model

```
ml model lf esteve (d=fu2-fu5 sex2 year2 agegrp2-agegrp4)
ml maximize, eform("RER")
```


- The likelihood shown in Equation 5 is identical to the likelihood for grouped Poisson data.
- This is because the underlying model assumes piecewise constant hazards (sometimes called a piecewise exponential model) meaning the number of deaths in each interval can be described by a Poisson distribution (Andersen et al. [7, pp. 409], Breslow and Day [8, Section 4.2]).
- As such, we could estimate the model in the framework of generalised linear models with a Poisson error structure.

A generalised linear model where the observed number of deaths is assumed Poisson

- We assume that the number of deaths for observation j can be described by a Poisson distribution, $d_j \sim \text{Poisson}(\mu_j)$ where $\mu_j = \lambda_j y_j$ and y_j represents person-time at risk.
- The observations can represent either life table intervals (in which case there can be multiple deaths per observation), individual patients, or subject-bands.

- Equation 2 ($\lambda = \lambda^* + \nu$) is then written as

$$\mu_j/y_j = \lambda_j^* + \exp(\mathbf{x}\beta), \quad (6)$$

  which (writing $\lambda_j^* = d_j^*/y_j$) can be written as

$$\ln(\mu_j - d_j^*) = \ln(y_j) + \mathbf{x}\beta, \quad (7)$$

  where d_j^* is the expected number of deaths (due to causes other than the cancer of interest).

- This implies a generalised linear model with outcome d_j, Poisson error structure, link $\ln(\mu_j - d_j^*)$, and offset $\ln(y_j)$.
- Breslow and Day [8, pp. 173–176] discuss similar (identical) models with application to an occupational cohort study.

- The model can be estimated using any software package which supports the estimation of generalised linear models with user-specified links (e.g. SAS (from version 6.10), Stata (from version 7), S-PLUS, R, and GLIM).
- To fit the model in SAS

```
proc genmod data=mydata;
fwdlink link = log(_MEAN_-dstar);
invlink ilink= exp(_XBETA_)+dstar;
class fu sex age dgnyear;
model d = fu sex age dgnyear / error=poisson offset=ln_y;
run;
```

- The model is defined in Stata using an ado file and then fitted using the standard Stata `glm` statement.

```
xi: glm d i.fu i.sex i.dgnyear i.age,
          family(pois) link(rs dstar) offset(log_y)
```


Empirical comparison of the models

- Two data sets from Finland are used, localised skin melanoma (5318) and localised colon carcinoma (6274 patients).
- The data sets contain all cases diagnosed in Finland (population 5.1 million) during 1975–94 with follow-up to the end of 1995.
- The localised skin melanoma data were chosen since this was one of the few cancer sites for which a main effects model provides a reasonable fit to the data (i.e. an assumption of proportional excess hazards is appropriate).
- The localised colon carcinoma data provide a typical example of data exhibiting non-proportional hazards with respect to age.

- The following approaches to estimating the model were used

(1) Grouped survival times, GLM with a binomial error structure;
(2) Grouped survival times, GLM with a Poisson error structure;
(3) Exact survival times, subject-band observations, GLM with a Poisson error structure;
(4) Exact survival times, subject-band observations, full likelihood; and
(5) Exact survival times, collapsed data, GLM with a Poisson error structure.

- SAS code for fitting these models to the Finnish colon and melanoma data is in the files `c:\data\toronto2002\sas_colon\models.sas` and `c:\data\toronto2002\sas_melanoma\models.sas`.

	Skin melanoma					Colon carcinoma				
	Grouped		Exact times			Grouped		Exact times		
	Bin.	Poi.	GLM	Full	Coll.	Bin.	Poi.	GLM	Full	Coll.
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
Deviance	76	73			76	120	113			131
Residual df	70	70			70	70	70			70
Estimated excess hazard ratios (i.e. $\exp(\beta)$)										
Follow-up 2/1	6.69	6.64	6.79	6.79	6.76	0.84	0.85	0.83	0.83	0.80
Follow-up 3/1	7.11	7.07	7.13	7.13	7.24	0.65	0.66	0.68	0.68	0.62
Follow-up 4/1	5.33	5.30	5.36	5.36	5.42	0.52	0.53	0.54	0.54	0.50
Follow-up 5/1	4.59	4.56	4.73	4.73	4.66	0.45	0.46	0.46	0.46	0.43
Female / Male	0.56	0.57	0.55	0.55	0.56	0.96	0.98	0.95	0.95	0.96
Year 85-94/75-84	0.63	0.63	0.63	0.63	0.63	0.73	0.73	0.73	0.73	0.73
Age 45-59/0-44	1.38	1.38	1.38	1.38	1.38	0.86	0.86	0.87	0.87	0.86
Age 60-74/0-44	1.90	1.86	1.92	1.92	1.89	1.07	1.05	1.06	1.06	1.07
Age 75+/0-44	3.19	2.99	3.14	3.14	3.24	1.37	1.29	1.34	1.34	1.44
SEs of the log hazard ratio (i.e. of the parameter estimates)										
Follow-up 2/1	.298	.301	.297	.297	.301	.093	.095	.094	.094	.092
Follow-up 3/1	.299	.301	.298	.298	.301	.109	.111	.108	.108	.108
Follow-up 4/1	.307	.310	.306	.306	.309	.131	.133	.128	.128	.130
Follow-up 5/1	.315	.317	.313	.313	.317	.151	.153	.150	.150	.150
Female / Male	.097	.098	.097	.097	.097	.077	.079	.077	.077	.076
Year 85-94/75-84	.098	.099	.097	.097	.098	.075	.076	.075	.075	.074
Age 45-59/0-44	.125	.125	.125	.125	.125	.156	.157	.156	.156	.157
Age 60-74/0-44	.128	.129	.127	.127	.128	.143	.144	.143	.143	.143
Age 75+/0-44	.173	.181	.173	.173	.172	.151	.153	.151	.151	.150

Comparison of the various relative survival models

- We are estimating the same model using slightly different approaches and we wouldn't expect large differences in the estimates.
- The approaches based on exact times are theoretically superior to those based on aggregated data, although the advantages are only minor.
- The choice of approach, in practice, depends not only on theoretical considerations but also on how easy the approach is to apply using available software.
- We recommend the generalised linear model with a Poisson error structure since one can utilise the theory of generalised linear models for assessing goodness-of-fit and studying regression diagnostics.
- Such models can be fitted using standard statistical software packages (e.g. SAS (from version 6.10), Stata (from version 7), S-PLUS, R, and GLIM).

References

[1] Brenner H, Gefeller O, Stegmaier C, Ziegler H. More up-to-date monitoring of long-term survival rates by cancer registries: an empirical example. *Methods Inf Med* 2001;**40**:248–52.

[2] Brenner H, Hakulinen T. Long-term cancer patient survival achieved by the end of the 20th century: most up-to-date estimates from the nationwide Finnish cancer registry. *British Journal of Cancer* 2001;**85**:367–371.

[3] Brenner H, Hakulinen T. Up-to-date long-term survival curves of patients with cancer by period analysis. *Journal of Clinical Oncology* 2002;**20**:826–832.

[4] Hakulinen T, Tenkanen L. Regression analysis of relative survival rates. *Applied Statistics* 1987;**36**:309–317.

[5] Estève J, Benhamou E, Croasdale M, Raymond L. Relative survival and the estimation of net survival: Elements for further discussion. *Statistics in Medicine* 1990;**9**:529–538.

[6] Hédelin G. *RELSURV 2.0 a program for relative survival analysis*. Department of Epidemiology and Public Health, Faculty of Medicine, Louis Pasteur University, Strasbourg, France, 1997.

[7] Andersen PK, Borgan , Gill RD, Keiding N. *Statistical Models Based on Counting Processes*. Springer-Verlag, 1995.

[8] Breslow NE, Day NE. *Statistical Methods in Cancer Research: Volume II - The Design and*

Analysis of Cohort Studies. IARC Scientific Publications No. 82. Lyon: International Agency for Research on Cancer, 1987.