Statistical Research and Applications Branch, NCI, Technical Report # 2008-01

The Method to Estimate Breast Cancer Prevalence at State Level: 2005-2015.

Angela B Mariotto[1] and Roberta De Angelis[2]

[1]Statistical Research and Applications Branch, Division of Cancer Control and Population Sciences, NCI,

[2]National Centre of Epidemiology, Istituto Superiore di Sanità, Rome, Italy

This technical report describes in further detail the method used to estimate and project breast cancer prevalence at state level.

1. Introduction

Calculation of complete cancer prevalence requires several years of incidence data and accurate vital status information at end of follow-up. Five states from the Surveillance, Epidemiology, and End Results (SEER) program(1) (http://seer.cancer.gov) have cases diagnosed from 1975 and allow estimation of prevalence including 30 years of diagnosis that is close to complete. Fewer years of cancer incidence are available from states funded by the Centers for Disease Control and Prevention (CDC) National Program of Cancer Registries (NPCR) (2), however most of the registries have incidence cases from 1995 which do not allow for estimates of complete prevalence. Gaps in data collection prevent direct calculation of cancer prevalence for many states. We use a statistical model that predicts cancer prevalence(3) from state specific cancer mortality data from National Center of Health Statistics (NCHS) and cancer survival model adjusted to represent cancer survival in each respective state. The method has been validated against reported incidence cases for 39 states and the District of Columbia from either SEER or NPCR.

2. The MIAMOD Method

The MIAMOD (Mortality, Incidence Approach MODel) method(3) uses as input state specific mortality (breast cancer and all causes), populations and state-specific modeled breast cancer survival. It back-calculates breast cancer incidence and prevalence from breast cancer mortality and survival.

2.1. *Input: Mortality and population data*

Single age and year state specific female mortality data for breast cancer and all causes of death from the National Center of Health Statistics (NCHS), and respective populations from the US Census Bureau, are available for calendar years 1969-2005 from the SEER*Stat Data Bases and software,(1;4). The 2006-2015 state population projections were obtained from the "State Interim Population Projections by Age and Sex: 2000– 2050" consistent with Census 2000 from the U.S Census Bureau(5).

2.2. *Input: State specific breast cancer survival model*

2.2.1. *SEER baseline model*

Data from 1975 to 2004 from the initial 9 SEER registries (1) are used to calculate female breast cancer relative survival rates by 3-year period of diagnosis (1975-1977, 1978-1980,…, 2002-2004), age at diagnosis (15-44, 45-54, 55-64, 65-74, 75-99, 85+). In MIAMOD estimation cancer survival information is needed before 1975 to include all past diagnosis, so we fit a parametric Weibull cure model (6)'(7) to the SEER survival data. These models assume that a hypothetical fraction of the patients will not die of cancer, and will experience the same mortality risk as the general population, while the complementary fraction will die of cancer, and their survival time follows a Weibull distribution. Fitting a parametric model to survival data allows survival extrapolation beyond the range of the empirical data.

The SEER baseline model is specified as the probability of surviving t years from diagnosis, for people diagnosed at year y and at the i^{th} age class (cumulative survival),

$$S_i(y,t) = \left\{ P_i + (1-P_i)\exp[-((\lambda_i t)^{\gamma_i})] \right\}^{\exp[\beta_i (y-\bar{y})]} \qquad [1]$$

where P represents the proportion of cured cases, and λ are γ, respectively, the scale and shape Weibull's parameters and β is the log relative risk of being diagnosed one year later than an arbitrary reference year $\bar{y}$ (e.g. the central value in the period 1975-2004).

2.2.2. *State Specific Relative Survival: Relative Risk (RR) Adjustment*

We adjusted the SEER baseline survival, to represent state specific survival, by applying state specific relative risks that reflect a greater or smaller risk of breast cancer death in a specific state relative to the SEER-9 areas. The method consists of regressing 5-year breast cancer survival on socio-demographic variables for all counties in the SEER-9 areas. The method and cross-validation is described in detail in Mariotto et al (2002)(8). A brief description is given below.

Sociodemographic variables at county level were constructed from different sources. They include urban/rural status, income, education, race and ethnicity from the 1970, 1980 and 1990 Census data (Geolytics Inc. 1998)(9) all causes and breast cancer mortality from NCHS, percent of county residents who ever smoked, who were at risk of obesity and who had had a mammogram in the last two years from the Behavioral Risk Factor Surveillance System (BRFSS) surveys (http://www.cdc.gov/brfss). The socioeconomic variables for the counties in the SEER areas were linked to 5-year breast cancer relative survival in the respective period.

The model consists of fitting a proportional hazards regression model to 5 year relative survival on socio-demographics variables at county level(8;10).

Once the model parameters are estimated, breast cancer 5-year survival rates and breast cancer mortality risks are projected to larger areas than counties, e.g. states, by calculating a weighted average of the estimated survival rates and risks, respectively, for the counties belonging to the area. The weights used are the respective county population size. The relative risk of breast cancer death among cancer patients in each state relative to SEER patients are calculated by dividing the estimated area specific 5-year breast cancer death hazard by the estimated SEER 5-year death hazard.

Let R_k be the relative risk of state k with respect to SEER. The state specific survival is given by

$$S_{ik}(y,t) = S_i(y,t)^{R_k} \qquad [2]$$

where $S_i(y,t)$ is the SEER baseline survival given by [1].

In this specific application we gave considered two different ecological regression approaches. In the first we considered three models one for each period of diagnosis (i.e. 1975-1980, 1981-1991, and 1992-2002) and for one period only (1975-2002). The choice between the two sets of RRs was driven by the validation procedure of MIAMOD estimates. For each state the set of RRs (one or three) that gave breast cancer estimates fitting more closely the reported incidence data was selected as the optimal.

2.3. *The MIAMOD (Mortality-Incidence Analysis MODel)[2] method.*

The statistical method MIAMOD (Mortality-Incidence Analysis MODel) (3) is based on the mathematical relationships relating mortality (M) and prevalence (P), for a given cancer, to incidence (μ) and relative survival probabilities (S).

For a birth cohort, the age-specific proportion of prevalent cases at age x, P_x, is given by:

$$P_x = \sum_{i=0}^{x-1} (1-P_i)\mu_i \; S_{i,x} \qquad [3]$$

and the age specific probability of dying for the given cancer is:

$$M_x = \sum_{i=0}^{x} (1-P_i)\; \mu_i \;\; S_{i,x} \; d_{i,x} \qquad [4]$$

where (1-P_i) represents the probability to be free from cancer at age *i*, μ_i is the probability of being diagnosed with cancer at age *i* (incidence), and $S_{i,x}$ is the probability to survive to cancer up to age *x* for patients diagnosed at age *i* (relative survival) and $d_{i,x}$ is the probability to dye from cancer at age *x* for patients diagnosed at age *i.*

Assuming P_0=0 in [3], prevalence at any age *x* is a function of incidence and survival only. Relative survival probability is assumed to be known, while incidence probability is modeled in the logistic scale as a polynomial function of age x , period t and cohort $c = t - x$ covariates:

$$\text{logit}(\mu_{x,t}(\alpha)) = \alpha_0 + \sum_{i=1}^{A} \alpha_i (x)^i + \sum_{i=1}^{P} \alpha_{A+i} (t)^i + \sum_{i=1}^{C} \alpha_{A+P+i} (c)^i \qquad [5]$$

To avoid collinearity problems, the linear period coefficient, α_{A+1}, is excluded when polynomial degrees A,P,C are all different from 0.

The vector of incidence coefficients $\theta = (\alpha_0, \alpha_1, \ldots, \alpha_A, \alpha_{A+2}, \ldots, \alpha_{A+P}, \alpha_{A+P+1}, \ldots, \alpha_{A+P+C})$ is back-calculated by regressing *observed cancer deaths* with *estimated deaths* from equation [4]. The regression parameters are derived with the maximum likelihood method, assuming a Poisson distribution for cancer deaths. A stepwise-like procedure allows to determine the polynomial degrees and the parameters values $\hat{\theta}$. Once the incidence function has been estimated, prevalence and mortality are derived from equation [3] and [4] respectively. An ad hoc developed software(11) was used to produce these estimates.

2.4. Prevalence projections 2006-2015

MIAMOD prevalence projections from 2006 to 2015 are based on assumptions of future trends of survival, incidence, population and other causes mortality. *Survival* was assumed to be constant with rates equal to those estimated for the last period of data, 2004. *Incidence* was projected using the previously estimated age and cohort incidence model. This model describes slow changes in incidence, mostly the effect of risk factors, but no period changes. The *population* projections are based on the general assumption that recent state-specific trends in fertility, mortality, domestic migration, and international migration will continue(5). *Other causes mortality* is assumed to be constant as observed in the last years of observed data.

Population denominators are not available for annual ages after age 84. Prevalence for the age group 85+ was estimated by applying prevalence proportions for the age group 80-84 to the 85+ female populations. The age-adjusted rates are based on the US 2000 standard population.

3. Validation of MIAMOD estimates

MIAMOD estimates of breast cancer incidence cases are compared to reported cases where data is available. For nine states participating in the SEER Program, data was available through 2005, with different starting years. An incidence database was obtained through an agreement with the North American Association of Central Cancer Registries (NAACCR). US cancer registries reporting data to NAACCR participate in the SEER program or the Centers for Disease Control and Prevention (CDC) National Program of Cancer Registries (NPCR), or both, and receive support from the state, province, or territory where they are located. The NPCR states that participate in this study were those who met NAACCR registry certification standards as providing complete, accurate, and timely data for at least 3 consecutive years during 1995 to 2003 and agreed to release incidence data for this project (30 states and DC)(2). The start and end years of available data varied and the first breast cancer were calculated from the respective database. **Table 1** displays, for each state, data source (SEER, NPCR and Ecological) and range of years compared. For the remaining 11 states no reported cases were available, so we compared MIAMOD estimates with different and independent incidence estimates from ecological regression analysis at county-level(12).

Because the MIAMOD estimates represent person counts rather than tumor counts, we compared with incidence of *first breast cancers* reported in the respective database. The determination of the first breast cancer depends on the length of the registration period. For example, a women diagnosed with two breast cancers in 1992 and 1998 will be a case in 1992 if she resided one of the SEER states and in 1998 if she lived in the NPCR regions.

For each state we calculated the mean absolute percent difference (MAPD) between the MIAMOD estimated e_y and the reported o_y number of breast cancer cases diagnosed up to age 84 years over years $y = y_1, \ldots, y_m$:

$$MAPD = \frac{100 \sum_{y=y1}^{yn} \left| e_y - o_y \right| / o_y}{(y_n - y_1 + 1)}$$

For the five SEER states (Connecticut, New Mexico, Hawaii, Iowa, and Utah) 30-year limited duration prevalent cases on July 1th 2005 from MIAMOD are compared to the corresponding reported values.

Table 1: Validation of MIAMOD estimates against reported data from: SEER[1], NPCR[2] and ecological regression incidence estimation[11]. Age 0-84 yrs. *Incidence validation*: mean absolute percent difference (MAPD in %) between estimated and reported number of first breast cancer cases over varying time periods, and percent proportion of years in which absolute percent difference exceeds MAPD value (% > MAPD). *Prevalence validation*: absolute percent difference (APD) between estimated and reported number of first breast cancer survivors on July 1th 2005[1].

State	Data source for comparison	Incident Cases				Survivors at July 20005		
		Period of diagnosis	Estimated	Reported	MAPD in % (%>MAPD)	Estimated	Reported	APD (%)
Alabama	NPCR	1998-2003	15,979	16,645	6.8(50)			
Alaska	NPCR	1996-2003	2,319	2,489	7.8(38)			
Arizona	NPCR	1995-2002	23,637	23,742	3.5(50)			
Arkansas	Ecological	1995-2003	13,453	15,485	13.2(63)			
California	SEER	1988-2005	311,625	321,348	7.0(44)			
Colorado	NPCR	1995-2003	22,487	22,721	3.8(44)			
Connecticut	SEER	1975-2005	63,174	65,781	7.2(45)	28,310	30,613	7.5
Delaware	NPCR	1995-2003	4,933	4,828	5.4(33)			
DC	NPCR	1999-2003	2,148	1,999	7.8(40)			
Florida	NPCR	1995-2003	110,642	105,347	6.3(33)			
Georgia	NPCR	1999-2003	22,780	23,974	6.0(60)			
Hawaii	SEER	1975-2005	15,558	16,670	9.3(45)	9,466	9,794	3.3
Idaho	NPCR	1995-2003	6,237	6,700	7.5(56)			
Illinois	NPCR	1995-2003	70,436	71,893	5.0(44)			
Indiana	NPCR	1998-2003	23,010	23,389	6.3(50)			
Iowa	SEER	1975-2005	51,355	51,549	5.3(26)	22,692	22,984	1.3
Kansas	Ecological	1995-2003	15,577	15,595	3.4(38)			
Kentucky	SEER	1995-2005	26,550	27,887	8.0(45)			
Louisiana	SEER	1995-2005	26,843	28,453	8.9(55)			
Maine	NPCR	1995-2003	8,041	8,317	4.0(44)			
Maryland	Ecological	1995-2003	32,745	32,555	4.1(38)			
Massachusetts	NPCR	1997-2003	33,570	33,495	4.6(43)			
Michigan	NPCR	1995-2003	59,631	59,320	6.0(33)			
Minnesota	NPCR	1995-2003	27,149	28,545	6.1(56)			
Mississippi	Ecological	1995-2003	13,343	14,205	7.4(63)			
Missouri	NPCR	1998-2003	22,595	22,561	4.2(50)			
Montana	NPCR	1996-2003	4,996	4,962	6.9(38)			
Nebraska	NPCR	1995-2003	10,133	9,979	2.8(33)			
Nevada	Ecological	1995-2003	9,508	9,530	4.1(38)			
New Hampshire	NPCR	1999-2003	4,412	4,395	4.1(40)			
New Jersey	SEER	1979-2005	139,468	143,487	7.6(48)			
New Mexico	SEER	1975-2005	22,342	21,739	6.7(39)	12,052	11,180	7.8
New York	NPCR	1995-2003	110,326	115,968	6.2(56)			
North Carolina	NPCR	2001-2003	16,495	15,389	7.6(33)			
North Dakota	Ecological	1995-2003	3,847	4,010	4.7(63)			
Ohio	Ecological	1995-2003	71,478	68,853	3.9(38)			
Oklahoma	NPCR	1997-2003	16,107	16,189	4.6(43)			
Oregon	NPCR	1996-2003	19,662	19,987	5.5(50)			
Pennsylvania	Ecological	1995-2003	82,289	82,407	3.9(50)			
Rhode Island	NPCR	1995-2003	6,752	6,695	3.4(33)			
South Carolina	NPCR	1997-2003	17,323	18,103	6.2(43)			
South Dakota	NPCR	2001-2003	1,428	1,507	7.9(33)			
Tennessee	Ecological	1995-2003	31,118	31,522	4.4(50)			
Texas	NPCR	1995-2003	90,909	96,524	7.2(56)			
Utah	SEER	1975-2005	22,854	20,532	10.8(35)	12,816	10,557	21.4
Vermont	Ecological	1995-2003	3,629	3,703	3.5(63)			
Virginia	Ecological	1995-2003	41,314	40,287	3.9(25)			
Washington	NPCR	1995-2003	35,988	36,524	3.6(33)			
West Virginia	NPCR	1995-2003	10,557	11,223	7.9(56)			
Wisconsin	NPCR	1995-2003	32,952	32,319	4.0(22)			
Wyoming	NPCR	1995-2001	2,262	1,913	18.9(43)			

Reference List

(1) Surveillance, Epidemiology, and End Results (SEER) Program - SEER 9 Registries, Nov 2007 Submission (1975-2005). 2008. (www.seer.cancer.gov). National Cancer Institute, released April 2008, based on the November 2007 submission.

(2) North American Association of central Cancer Registries, NAACCR. Incidence-CINA Analytic File, 1995-2003.

(3) Verdecchia A, Capocaccia R, Egidi V, Golini A. A method for the estimation of chronic disease morbidity and trends from mortality data. Stat Med 1989; 8(2):201-216.

(4) Surveillance Research Program, National Cancer Institute SEER*Stat software version 6.4.4. (www.seer.cancer.gov/seerstat) 2008.

(5) U.S.Census Bureau PD. Interim projections consistent with Census 2000 (released March 2004). http://www.census.gov/population/www/projections/usinterimproj/ . 2008.
Ref Type: Electronic Citation

(6) Verdecchia A. The cure for colon cancer: results from the EUROCARE study. International journal of cancer 1998; 77(3):322-329.

(7) De Angelis R. Mixture models for cancer survival analysis: application to population-based data with covariates. Statistics in medicine 1999; 18(4):441-454.

(8) Mariotto A, Capocaccia R, Verdecchia A, Micheli A, Feuer E, Pickle L et al. Projecting SEER cancer survival rates to the US: An ecological regression approach. Cancer Causes and Control 2002; 11:101-111.

(9) GeoLytics, Inc. census CD + Maps. East Brunswick, NJ, 1998.

(10) Hakulinen T, Tenkanen L. Regression Analysis of Relative Survival Rates. Applied Statistics 1987; 36(3):309-317.

(11) De Angelis G, De Angelis R, Frova L, Verdecchia A. MIAMOD: a computer package to estimate chronic disease morbidity using mortality and survival data. Computer Methods and Programs in Biomedicine 1994; 44(2):99-107.

(12) Pickle LW, Hao Y, Jemal A, Zou Z, Tiwari RC, Ward E et al. A new method of estimating United States and state-level cancer incidence counts for the current calendar year. CA Cancer J Clin 2007; 57(1):30-42.