Supplemental Materials

*Stage IV versus distant stage definitions*

MBC was defined as stage IV using the SEER adjusted American Joint Committee on Cancer (AJCC) 6th edition staging classification (1). This stage definition uses extent of disease (EOD) information for cases diagnosed in 1988-2003 and collaborative staging (CS) from cases diagnosed in 2004-2012 collapsed into the same AJCC 6th edition definition. The AJCC 6 stage IV definition was used because it only includes tumor in which one or more distant metastasis are identified. Stage IV from previous AJCC editions as well as distant stage from SEER historical summary staging classification include some locally advanced diagnosis without distant metastasis. For example, tumors with positive supraclavicular lymph nodes involvement without distant metastasis were classified as stage IV in previous AJCC editions and as distant in SEER historical summary stage. SEER historical summary stage also classifies inflammatory carcinomas, including diffuse (beyond that directly overlying the tumor) dermal lymphatic permeation or infiltration as distant disease. The Summary stage 2000 <http://seer.cancer.gov/tools/ssm/breast_femgen.pdf>, a summary stage variable used for cases diagnosed January 1, 2001 and forward, classifies both of those diagnoses, involvement of supraclavicular lymph nodes and inflammatory carcinoma as regional disease being in more alignment with AJCC 6th ed. stage IV. In current practice, a negative clinical history and examination are deemed sufficient to designate a case as having no distant metastasis; extensive imaging or other testing is not required.

## Modeling survival time from de novo stage IV BC to cancer death

To extrapolate survival beyond the observed data, as required by the back-calculation method, we fit mixture cure survival models to MBC relative survival data. The mixture cure survival model assumes that a proportion of cancer patients is cured from cancer and dies from other causes with rates similar to the general population while the remaining patients will die following a Weibull survival distribution. Even though most stage IV BC patients die of their cancer, this model is used because it allows for a better modeling of long term survivors and a better extrapolation of survival beyond the observed data. The relative survival data is calculated for breast cancer cases diagnosed with stage IV between years 1992 and 2012 in the SEER-11 areas and age at diagnosis is grouped in 5 year intervals 15-49, 45-64, 65-74, 75-84, 85-99 and year at diagnosis 3 year overlapping average groupings (1992-1994, 1993-1995, ….,2010-2012) to estimate trends in survival. We fit a separate model to each of the 5 age groups (15-49, 45-64, 65-74, 75-84, 85-99) and used calendar as covariate in the model. Thus, if for each age group is the relative survival to time  then the mixture cure model is specified as

 [1]

where is the fraction cured, is a Weibull survival function for those not cured, and  a hazard ratio that adjusts for trends in survival by year at diagnosis.

*BC mortality by stage at diagnosis to estimate breast cancer deaths attributable to recurrence*

Mortality data obtained from death certificate does not contain information on the tumor diagnosis such as stage. To estimate the proportion of breast cancer deaths originated from women with earlier stage BC we used incidence-based breast cancer mortality rates by stage. Incidence-based mortality is obtained from linking SEER-9 incidence cases to mortality data, and allows for mortality rates to be analyzed by tumor characteristics (e.g. stage) (2). Appendix Figure A displays breast cancer mortality rates for women diagnosed with breast cancer by stage in the SEER-9 areas per 100,000. A number of years (burn-out period) is necessary so that the rates stabilizes and can be analyzed. For example, Figure A shows that breast cancer incidence-based mortality rates are very low in 1975 since it only includes deaths for women diagnosed with breast cancer in 1975, it increases as more diagnosis years are added, and is very similar to the US breast cancer mortality rates after 1995. Using this data we estimate that in the most recent years (2010-2012) 20% of BC deaths in a given year originate from women diagnosed with *de novo* MBC, while 80% are deaths from women diagnosed with earlier stage BC who progressed to recurrent MBC.

*The MIAMOD back-calculation method*

Letbe age (in years),  be period (calendar year) and  be the birth cohort. Let be the probability of having MBC (MBC prevalence) and  be the probability of being diagnosed with MBC (MBC incidence), the two quantities we are interested in estimating. The MIAMOD method assumes that the incidence function is a parametric age, period and cohort model, given as



where , and are either polynomial functions or restricted cubic splines with tails that are linearly restricted and do not containing a constant term and , and  are the vector of coefficients to be estimated. Due to the collinearity between age, period and cohort, the linear term for period is eliminated. If a polynomial function is chosen for one of the components the user has to specify the order of the polynomial. For cubic spline function the user has to specify the number and location of knots.

Without loss of generality consider a birth cohort . We assume that before a breast cancer (BC) death there must be a diagnosis of metastatic breast cancer (MBC) either *de novo* or recurrence. The hazard rate of dying of BC, , can be estimated from national mortality statistics as the number of deaths divided by the population respective population. The hazard of dying of BC at age  being diagnosed with MBC at age  ,  , and its associated probability of surviving to age the extra risk of death associated with a MBC diagnosis at age , , can be estimated from cancer registry data and/or epidemiology, cohort studies with MBC information. Because cancer registries in the US do not collect recurrence information, we used an adjustment to the SEER *de novo* MBC survival based on study that included women diagnosed with recurrent and de novo MBC (3). For individuals born in year , the MIAMOD(4, 5) specifies two equations relating incidence, prevalence, mortality and survival(4),

, [1]

. [2]

The first equation defines the probability of dying of BC at age  as being MBC-free and progressing to a MBC diagnosis at age , surviving from age  to age and then dying at age , summing for ages between 0 and . The second equation specifies prevalence at age  in a similar way. From these two equations it is possible to write the mortality rate as a function of the parameters to be estimated. Thus by assuming that the observed number of BC deaths,  follows a Poisson distribution with mean  where is the US female population size in year  and age , we can estimate the coefficients  and its standard errors from the MIAMOD program. The order of the polynomials or the location and number of spline knots that best fit the data can be found in a systematic way, similar to backward or forward regression procedure, by fitting models with different incidence functions and using the log-likelihood ratio test. Once the incidence parameters are estimated, prevalence is directly calculated from the incidence function and survival, using equation [2]. Projections of prevalence beyond the last year of data, 2013, are based on the assumptions that incidence and relative survival levels are the same as the one estimated in the last year with data.

*Calibration of the back-calculation method*

In order to calibrate the back-calculation method to data inconsistencies, such as multiple tumors, deaths misclassification and underreporting, we compared the prevalence of *de novo* MBC in the SEER-9 areas obtained from the MIAMOD and counting methods.

The prevalence of *de novo* MBC is calculated using the counting method. The method counts all women alive at 1/1/2013 with a previous diagnosis of stage IV breast cancer (1988-2012) in the SEER-9 areas (6, 7). The method also estimates the probability of cases lost to follow-up to survive to the prevalence date by using overall survival of similar cases.

To estimate prevalence of de novo MBC using the backcalculation method we used the incidence-based mortality database and extracted all breast cancer deaths (1995-2013) of women diagnosed *de novo* stage IV breast cancer in the SEER-9 areas from 1988 onwards. Survival is the modeled *de novo* SEER stage IV survival, since we are only including breast cancer deaths from women diagnosed with *de novo* stage IV. The counting method estimates that the prevalence of *de novo* MBC in the SEER-9 areas is 3,498 at 1/1/2013. Prevalence of *de novo* MBC from the back-calculation method was 3,127 and found to be 11% lower. This type of underestimation has been reported elsewhere (8) and may be due to misclassification of deaths (deaths attributed to the metastasis site), or difficulties in determining cause of death for people with multiple tumors. After adjusting survival by a factor of 0.92=exp(-0.08) , where prevalence of de novo stage IV was 3,488.

Thus the final survival from MBC including *de novo* and recurrence is



Figures B1 and B2 in the appendix compare the incidence and prevalence estimates from MIAMOD with the reported incidence and prevalence at the SEER-9 areas.

*Trends in age-adjusted breast cancer incidence by stage at diagnosis*

In order to provide some insight in survival trends for de novo stage IV BC we looked at age-adjusted breast cancer incidence rates trends by age and by stage. We selected cases diagnosed with breast cancer in the SEER-11 areas between 1992 and 2012. We used the Adjusted AJCC 6th Stage (1988+) <http://seer.cancer.gov/seerstat/variables/seer/ajcc-stage/6th>. We selected malignant breast cancers, women with known age at diagnosis, and first matching record for each person. For women diagnosed with stage IV breast cancer we further stratified by age groups. We fit Joinpoint models to age-adjusted rates incidence trends to estimate years at which trends changed (9, 10). Figure C1 shows that age-adjusted incidence trends for BC stage II are increasing while for stage II and unstaged are decreasing in the most recent years. Figure C2 shows that age-adjusted incidence rates of women diagnosed with stage IV breast cancer in SEER-11 areas are increasing for all age groups but more significantly for younger women diagnosed at ages 15-49 years.

Supplemental Figure 1. Age-adjusted SEER incidence-based (IB) mortality rates for cases diagnosed in the SEER-9 areas between 1975-2012 by stage at diagnosis. These rates are calculated using breast cancer deaths among women with a breast cancer diagnosis reported to SEER-9 registries, the denominator being the female SEER-9 population. The black curve represents US breast cancer mortality rates. Since the IB mortality rates are composed of deaths among cases diagnosed in previous years, the rates require a long series of incidence data. In this example, the SEER-9 IB breast cancer mortality mimics the US breast cancer mortality only after 1995.



Supplemental Figure 2. Results from the calibration of the MIAMOD method. Comparison of observed and estimated (MIAMOD) prevalence counts of de *novo* stage IV breast cancer in the SEER-9 areas at 1/1/2013 by age. Observed are directly calculated, by counting all women alive at 31st, December 2012 with a previous (1988-2012) diagnosis of stage IV breast cancer in the SEER-9 areas and represent 25-year limited duration prevalence.

Supplemental Figure 3. Comparison of observed and estimated (MIAMOD) new incidence cases of de *novo* stage IV breast cancer in the SEER-9 areas in 2012 by age. Observed are stage IV breast cancer cases reported to SEER-9 registries in 2012.

Supplemental Figure 4. Trend in age-adjusted incidence rates of women diagnosed with stage IV breast cancer in SEER-11 areas by age group.



Supplemental Figure 5. Trend in breast cancer age-adjusted incidence rates in SEER-11 areas by stage.



References

1. SEER Adjusted AJCC 6th Edition Stage Classification, <http://seer.cancer.gov/seerstat/variables/seer/ajcc-stage/6th/#stage>. <http://web2.facs.org/cstage0204/breast/Breast_qad.html>.

2. Chu KC, Miller BA, Feuer EJ, Hankey BF**.** A Method for Partitioning Cancer Mortality Trends by Factors Associated with Diagnosis - an Application to Female Breast-Cancer. Journal of Clinical Epidemiology. 1994;47(12):1451-61.

3. Dawood S, Broglio K, Ensor J, Hortobagyi GN, Giordano SH**.** Survival differences among women with de novo stage IV and relapsed breast cancer. Ann Oncol. 2010;21(11):2169-74.

4. 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):201-16.

5. 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.

6. Feldman AR, Kessler L, Myers MH, Naughton MD**.** The prevalence of cancer: estimates based on the Connecticut Tumor Registry. New England Journal of Medicine. 1986;315(22):1394-7.

7. Mariotto A, Gigli A, Capocaccia R, Tavilla A, Clegg LX, Depry M, et al. Complete and Limited Duration Cancer Prevalence Estimates. In: Ries LAG, Eisner MP, Kosary CL, et al., eds. SEER Cancer Statistics Review: 1973-1999, Natinal Cancer Insitute, Benetsda MD. (<http://seer.cancer.gov/csr/1973_1999/> ); 2002.

8. De Angelis R, Tavilla A, Verdecchia A, Scoppa S, Hachey M, Feuer EJ, et al.Breast cancer survivors in the United States: geographic Variability and Time Trends, 2005-2015. Cancer. 2009;115(9):1954-66.

9. Kim HJ, Fay MP, Feuer EJ, Midthune DN**.** Permutation tests for joinpoint regression with applications to cancer rates. Statistics in Medicine. 2000;19(3):335-51.

10. Joinpoint Regression Program, Version 4.3.1.0 - April 2016; Statistical Methodology and Applications Branch, Surveillance Research Program, National Cancer Institute.