# Supplementary Methods and Materials

This document includes details on the development of the analytical and numerical solutions for the deconvolution method to estimate the survival of the time from diagnosis to recurrence, , using the estimated survival distribution for the non-cured  and survival from recurrence to death  . It also includes the simulations to evaluate the impact of departures from assumptions on the estimation of the survival distribution from diagnosis to recurrence  . The assumptions evaluated are: i) exponential distribution for the survival time from recurrence; ii) independence between the time from diagnosis to recurrence and time from recurrence to death; and iii) misspecification of the survival time from recurrence to death.

#### Analytical estimate of the time from diagnosis to recurrence (T1)

We define  as the time from diagnosis to metastatic recurrence and  the time from recurrence to breast cancer death. Assuming independence between and we can write the probability density function from diagnosis to cancer death  as 

 [E1]

An analytical solution to [E1] exists (1) in the case that the survival time from recurrence to death, , is exponential, . Differentiating both sides of the equation below



we get





Therefore, the density and the survival function for can be expressed respectively as



 [E2]

#### Numerical estimate of the time from diagnosis to recurrence (T1)

A numeric solution to equation[E1] can be obtained in the case that  is not exponential as below. Let  be uniformly assigned between 0 to  and , where  is the maximum follow-up time. The probability density function at  can be calculated from actuarial cumulative survival as. We can discretize equation [E1] by writing  equations at each 

 [E3]

 [E4]

#### Recurrence-free survival and the probability of progressing to cancer recurrence

Following estimation of the recurrence-free survival curve among those not cured, , we calculate the recurrence-free survival and probability of progressing to recurrence among all newly diagnosed cancer patients. A newly diagnosed cancer patient survives the event of recurrence at any specified time, if he/she can either be cured or belong to the not cured group but remain recurrence-free, i.e.,

 [E5]

Thus, the probability of progressing to cancer recurrence, or risk of recurrence by time *t* is calculated as

 [E6]

#### Variance-covariance matrix of G(t)

For  estimated analytically let be the vector of parameters where  is the cure fraction,  the parameters of the survival time for those not cured estimated from the mixture cure survival model, and  the hazard of the exponential density function . Let  be the covariance matrix of . The variance covariance matrix of is calculated using the delta method. In this method, we did not include variability of the survival from recurrence adjustment factor .

#### Simulations to assess methods performance and assumptions

We used simulations to evaluate the impact of departures from assumptions on the estimation of the survival distribution from diagnosis to recurrence . The assumptions were: i) exponential distribution for the survival time from recurrence ; ii) independence assumption that the time from diagnosis to recurrence does not influence the time from recurrence to death; and iii) misspecification of the survival time from recurrence, .

We generated 500 datasets. In each dataset, we randomly generated 500 times from diagnosis to recurrence and from recurrence to cancer death  for non-cured cases, using different survival distribution scenarios. The scenarios used are described below.

*Base scenario*:  ~ Weibull,  ~exponential, with  and  independent.

*Misspecified* *scenario*: The data is generated as in the base scenario but we use a value of  in equation [E2] different from the true exponential parameter used to generate the data.

*Non-exponential scenario*:  ~ Weibull,  ~ Mixture Cure Weibull model with  and  independent. The data is generated with a non-exponential distribution for  but the analytical estimate uses the exponential assumption for .

*Dependence scenario*: in this scenario is associated with  and a non-cured case is assigned with probability  to a group with shorter survival times  and with probability  to a group with longer survival times .

We used the following parameterization: Weibull specified as , exponential specified as  and mixture cure Weibull specified as . The median survival time for the Weibull, exponential and mixture cure Weibull are respectively,  and . Table A1 displays the scenarios and the parameters of the different survival distributions.

Supplementary Table 1. Scenarios and parameter values used in the simulations



*Simulation methods*

For each case, we calculate their overall survival time  considering censoring at 16, 20 and 25 years from diagnosis. For example, if , and censoring is 16 years then:  is not censored, is censored at 16 years, and  censored at 1 year. The estimand of interest is the probability of being recurrence-free for non-cured patients,  at 5, 10 and 15 years from diagnosis.

1. To estimate analytically we first fit a Weibull distribution to  and an exponential function to  using SAS Proc Lifereg. We then substitute the respective Weibull and exponential estimated parameters in equation [E2].
2. To estimate numerically we first calculate the Kaplan-Meier survival distributions of  and , namely  and , using SAS Proc Lifetest. The discretized probability density functions at are calculated as,  and used in equations [E3] and [E4].
3. In the  misclassified scenario we only calculate the analytical estimate of by fitting a Weibull distribution to . We then use a misspecified hazard  that provides a shorter or longer mean survival compared to the true generated exponential distribution for  .

For both the analytical and numerical estimates of  we calculate the bias and the empirical standard error, respectively as,



and

 ,

where =500.

#### Simulation results

Table A2 displays the bias and empirical standard error of the analytical and numerical estimates of  for the different scenarios at  = 5, 10 and 15 years from diagnosis using censoring intervals of 16, 20 and 25 years. In general, the biases were small and lower than 5% survival points. The largest bias, an overestimation of 10% survival points on the estimate of , occurred in the analytical method when the time from recurrence was misspecified as being much shorter, median survival time of 1.4 years instead of 2.1 years. Misspecifying the time from recurrence using a longer survival time, 2.8 years instead of 2.1 years, underestimated, but the biases were small except for the estimate of  and censoring interval of 16 years. Using the exponential assumption when the true is a Weibull mixture cure model, produced very small biases as did ignoring the dependence between the two survival times  and in the estimation methods. Note that the biases incurred in the estimation of have an attenuated effect on the final estimates of probability of recurrence as they are multiplied by the proportion of non-cured.

Supplementary Table 2. Estimates of bias and its empirical standard error (SE) for the analytical and numerical estimates of  in the base, misspecified, non-exponential and dependence scenarios. All the measures were calculated at  5, 10 and 15 years from diagnosis and using censoring intervals of 16, 20 and 25 years. Light and darker shaded cells represented biases 3.0% -4.9% and 5.0%-10% survival percent, respectively.



Supplementary Table 3. Estimated percent progressing to metastatic recurrence within 5, 10, 15 and 20 years and respective 95% confidence intervals for women diagnosed with breast cancer in 2000-2013 by stage and age. The results used the analytical deconvolution method and the log-logistic cure mixture model and represent probabilities in the absence of other causes of death. Survival from recurrence used an adjustment of hazard ratio r=1.35 compared to *de novo* stage IV breast cancer.



References

1. Capocaccia R**.** Relationships between incidence and mortality in non-reversible diseases. Stat Med. 1993;12(24):2395-415.