Estimation of net survival for cancer patients: relative survival setting more robust to some assumption violations than cause-specific setting, a sensitivity analysis on empirical data

Robin Schaffar[1,2] Bernard Rachet[2], Aurélien Belot[2], Laura M Woods[2]

1. Geneva Cancer Registry, Global Health Institute, University of Geneva, Geneva, Switzerland.
2. Cancer Research UK Cancer Survival Group, Department of Non-Communicable Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom

Corresponding author
Mr Robin Schaffar
Geneva Cancer Registry, Global Health Institute, University of Geneva,
55 Boulevard de la Cluse, 1205 Geneva, Switzerland.
+ 41.22.379.49.57
(robin.schaffar@unige.ch)

Key words
Net survival, Cause-specific, Relative survival, Informative censoring.

Highlights.
- Net survival is estimated using either cause-specific or relative survival setting
- A carefully revised cause of death is used for the cause-specific setting
- The Pohar-Perme estimator is used for the relative survival setting
- Bias of informative censoring is taken into account for both settings
- Enables to provide an unbiased comparison of the two settings

Abstract

Net survival is the survival that would be observed if the only possible underlying cause of death was the disease under study. It can be estimated with either cause-specific or relative survival data settings, if the informative censoring is properly considered. However, net survival estimators are prone to specific biases related to the data setting itself. We examined which data setting was the most robust against violation of key assumptions (erroneous cause of death and inappropriate life tables).

We identified 4,285 women in the Geneva Cancer Registry, diagnosed with breast, colorectal, lung cancer and melanoma between 1981 and 1991, and estimated net survival up to 20-years using cause-specific and relative survival settings. We used weights to tackle informative censoring in both settings, and performed sensitivity analyses to evaluate the impact of misclassification of cause of death in the cause-specific setting or of using inappropriate life tables on net survival estimates in the relative survival setting.

For all four cancers, net survival was highest when using the cause-specific setting and the absolute difference between the two estimators increased with time since diagnosis. The sensitivity analysis showed that (i) the use of different life tables did not compromise net survival estimation in the relative-survival setting while (ii) a small level of misclassification for the cause of death led to a large change in the net survival estimate in the cause-specific setting.

The relative survival setting was more robust to the above assumptions violations and is therefore recommended for estimation of net survival.

Introduction

Net survival measures the survival that would be observed if the only possible cause of death was the disease of interest [1]. It is the most defensible method of estimating survival from cancer. Two main settings have been described for its estimation: the relative survival setting and the cause-specific setting. The latter requires information on underlying cause of death so that deaths due to causes other than the cancer of interest can be censored. Such information is not needed in the relative survival setting. Here, the overall survival of the cancer patients is compared to the survival they would have experienced if they had had the mortality of the general population from which they were drawn [2].

In both settings, net survival estimation is susceptible to bias due to informative censoring. Informative censoring occurs when patients are removed from the risk set (censored) under a non-random way: these patients would experience a different mortality hazard compared to those that remain in the risk set. In the cause-specific setting, when the interest is in estimating the cancer-specific mortality hazard, patients who died due to other cause are censored (and so removed from the risk set). It means that patients with higher risk of dying from causes other than cancer (for example, elderly compare to young patients) are more likely to be removed from the risk set. However, because age is also an important prognostic factor for cancer, censoring these patients is informative for the cancer-survival estimation. In the relative survival setting, this mechanism of informative censoring is less easy to conceptualise (because the cause of death is unknown and/or not used); any variable with an effect on both cancer-specific and other-cause mortality hazards induces informative censoring. Demographic variables which define the life tables may lead to informative censoring and need to be accounted for. A new estimator has been described by Pohar-Perme which is able to take account of this bias within the relative survival setting [3] and its performances have been assessed in an extensive simulation study [4]. We have recently proposed a similar strategy for the estimation of net survival in the cause-specific setting [5].

If informative censoring is accounted for, estimates of net survival derived in each of these settings are theoretically unbiased. However, biases relating to the data setting itself may still occur. In the relative survival setting, bias can originate from the non-comparability between the cohort and the general population from which rates of expected mortality are drawn, due to unmeasured variable(s) affecting both expected and excess hazard rates (this latter being the rate from which the net survival is derived). In the cause-specific setting, bias can arise from the misclassification of the underlying cause of death. Our previous analyses of patients diagnosed with breast cancer in Geneva

showed that the estimation of net survival using the cause-specific setting was very sensitive to the codification of underlying cause of death, but, in contrast, the relative survival setting was robust to non-comparability in the estimation of background mortality [5].

Breast cancer may, however, represent a special case. Survival amongst breast cancer patients is high, but deaths directly caused by the original cancer still occur into the second and third decades following diagnosis: a pattern of excess mortality which is seen for very few other anatomic sites. As such, our previous conclusion may not hold for every cancer type. Here we extend our analysis of breast cancer patients to patients diagnosed with cancers of three other anatomic sites (according to the international classification of disease, 10th version, ICD-10) in order to establish whether the same conclusions hold for other malignancies.

Material and Methods

The Geneva Cancer Registry records underlying cause of death for all cancer patients. More unusual, the registry also validates the accuracy of this variable by reviewing all clinical information available for each patient. The overall agreement between the variables (revised cause of death vs. cause of death based on death certificates) was high. However, several subgroups presented a lower concordance, suggesting differences in calendar time and less attention given to older patients and more advanced diseases [6]. This context thus represents a unique opportunity to compare relative survival and cause-specific settings when estimating net survival, because the registry holds more accurate information on the underlying cause of death.

We selected women diagnosed between 1981 and 1991 at ages 15 to 99 with invasive colorectal (C18-20), lung (C34), melanoma (C44.1) or breast (C50) cancer. These malignancies afforded us tumours with a wide range of aggressiveness as well as very different incident age distributions. All patients were followed up to the end of 2012.

Our approach has been described previously [5]. Briefly, we used the Pohar-Perme estimator in the relative-survival setting and our own derived estimator for the cause-specific setting to estimate net survival. Both estimators take into account informative censoring, that is the fact that the number of patients we observe to be at risk is smaller than the number of patients that would be at risk in the hypothetical world, were people could die of the cancer of interest only. Because the same is true for the number of deaths as well weights are used to correct the net survival estimates for this bias [7], [8]. We define these estimates as the “baseline situation”.

We then examined two sets of scenarios to evaluate the extent of biases arising from the data setting. The aim of scenarios A1 and A2 was to evaluate the impact of the life tables upon net survival estimation within the relative survival setting. In Geneva, general population mortality rates are available by year of age and calendar year. However, other socio-demographic variables are known to have a strong influence on the probability of death. In scenario A1, we consider a simulated stratification of the expected age-, sex- and period-specific mortality rates by deprivation. In scenario A2, we artificially increase the expected mortality of all patients, well above what would ordinarily be expected, by attributing the mortality of the most deprived patients to the whole cohort. We computed the difference between the baseline estimator (using the cause-specific setting) and the net survival estimates derived under scenarios A1 and A2. Differences were smoothed with a weighted non-parametric regression on time since diagnosis [9].

Scenario B aimed to evaluate the impact of misclassifying the cause of death on net survival estimation in the cause-specific setting. Here we randomly reattributed non-cancer deaths to cancer deaths for 10, 15, 20 and 25% of the deceased patients (scenarios B1, B2, B3 and B4, respectively). This was performed 100 times to derive a mean cause-specific net survival for each scenario. We derived the difference between the baseline estimator (using the relative survival setting) and the cause-specific estimates in scenarios B1, B2, B3 and B4. Differences were smoothed with a weighted non-parametric regression on time since diagnosis [9].

Results

The final cohort was composed of 996 women diagnosed with colorectal cancer, 500 women diagnosed with lung cancer, 300 women diagnosed with melanoma and 2,489 women diagnosed with breast cancer.

Table 1 describes the age distribution and aggressiveness of each disease. Patients diagnosed with colorectal, lung, melanoma and breast cancer presented a mean age of 72.1, 67.6, 53.1, and 62.1 years, respectively. For colorectal cancers, 87~~.1~~% of the patients died, 511 of their cancer (59~~8.9~~%). Among women diagnosed with lung cancer, 97~~96.6~~% died. There were 483 deaths, 419 from lung cancer (~~86.7~~87%). Among patients with melanoma, ~~45.6~~46% died, 40 due to melanoma (29~~.2~~%). For breast cancer 1,700 patients died (68~~.3~~%) 844 from their cancer (~~49.6~~50%).

Baseline estimators of net survival are presented in Figure 1 for each cancer site. We observed, consistent with our previous analyses, that net survival estimates using the cause-specific setting are higher than the estimates using the relative survival setting for every localisation. The absolute difference between the two estimators increased with time since diagnosis for all four cancers. For

colorectal cancer, the difference widened from almost 2% at one year to over 7% at 20 years. For lung cancer, the difference increased sharply within the first two years after diagnosis (~~2.5~~3% at two years) and moderately afterwards. There was no detectable difference during the first three years after diagnosis for melanoma, but it subsequently increased to more than 8% at 20 years after diagnosis. For breast cancer, the absolute difference between the two estimators increased with time since diagnosis from 1% at one year to ~~10.8~~11% at 20 years.

Where deprivation-specific life tables were used (scenario A1), the difference between both net survival estimators was fairly constant across all four cancers (Figure 2). When we used the life tables of the most deprived population (scenario A2), we still observed a small but substantial difference for all cancer sites.

By contrast, increasing the proportion of deaths classified as being due to cancer led to a decreasing difference between the cause-specific estimate and the baseline estimate (derived within the relative survival setting) for all anatomic sites, even turning negative for colon cancer and melanoma (Scenarios B1- B4). For colorectal cancer, the effect was dependent on time since diagnosis: early in follow-up the difference was eliminated only with more than 25% of deaths reallocated. However, after 5 years, the reallocation of 10-15% resulted in no difference in the two estimators. In contrast, for lung cancer, 25% re-allocation did not eliminate the difference between the two net-survival estimators. From 6 years after diagnosis onwards, the difference was almost eliminated for melanoma if 25% of deaths were reallocated. For breast cancer, the difference decreased as the proportion of re-allocation increased, even turning negative. When 15–20% of the deaths were reallocated, the difference was close to zero.

Discussion

This study has evaluated whether our previous conclusions relating to the nature and size of modifications to the data in each setting are similar for breast cancer [5] and other cancers with very different patterns of incidence and excess mortality. We account for informative censoring in all analyses, allowing an accurate comparison of two unbiased net survival estimators. Our analyses are therefore not comparable with previous studies comparing cause-specific survival and relative survival [10] [11]. Indeed, the estimators used for these studies were not estimating net survival as they did not account for informative censoring [3].

Differences between the two net survival estimators varied with time since diagnosis and with anatomic site. However, the cause-specific approach generally resulted in higher estimates of net survival. The first sensitivity analysis (scenarios A1 and A2) showed that the use of different life

tables did not compromise net survival estimation in the relative-survival setting. Even with a very large modification in the expected mortality rate (by the application of mortality rates for the most deprived population to all patients), estimates of net survival were fairly stable. Net survival estimation in the relative survival setting appeared, therefore, to be relatively robust to non-comparability of the underlying mortality rates to the patient population, irrespective of the anatomic site. By contrast, the second sensitivity analyses showed a greater impact in net survival estimates within the cause-specific setting: a relatively small level of misclassification for the underlying cause of death led to a large change in the net survival estimate. This was true for all cancer sites.

After longer periods of follow up, the Pohar-Perme estimator tended to produce erratic results when the number of deaths was small. Net survival can increase within the relative-survival setting because the observed mortality of the cancer patient group can be lower than their expected mortality. With increasing time, the few remaining patients need to represent more and more of their counterparts; such patients, especially among elderly, are more likely to survive better than the general population, resulting in overall hazard lower than expected hazard. The excess hazard therefore becomes negative and the survival function increases. The erratic curves in the context of net survival derived with the relative survival setting therefore originates from the fact that we are asking questions about the hypothetical world that are not supported by sufficient information in the real world. We estimated net survival at 20 years for comparative purpose but being interested in a “25-year net survival of a 90-year old patient” implies asking what would happen in 25 years to a patient who is 90 at the time of diagnosis if they could not die from other reasons than cancer. Such a question of course makes no sense. Therefore, the length of analytical follow-up time should be restricted so that the population survival probability for all patients in the cohort is large enough. Another difficulty in long-term (net) survival is due to the increasing probability of multiple tumors. In our study, for patients having several tumours with the same ICDO-code (same or paired organ), we considered only the first tumour for the estimation of net survival.

Even with those limits aforementioned, ~~we think that~~ the relative survival setting should be the preferred approach when estimating net survival with population-based data, regardless of the cancer site, because it is less sensitive to inappropriate data changes ~~potential data biases~~ in comparison to the cause-specific setting. Parametric approaches using flexible regression models for the excess mortality hazard [Remontet 2007 StatMed, Charvat 2016 StatMed] could be considered in

the case of long follow-up time and few cancer deaths, where the Pohar-Perme estimator produces more erratic results.

Funding

This work was supported by the Swiss Cancer League [BIL KFS-3274-08-2013].

Conflict of interest

The authors have declared no conflict of interest.

Authorship contribution

All authors contributed to the manuscript. RS conducted the analysis and the writing under the supervision of LW,BR and AB. BR, LW and AB all reviewed the paper and made final corrections. All authors read and approved the final version of the manuscript.

Acknowledgements

This collaborative work was supported by the Geneva Cancer Registry and the Cancer Survival Group at London School of Hygiene and Tropical Medicine. We should like to thank Elisabetta Rapiti, Christine Bouchardy, Massimo Usel, Gerald Fioretta, Isabelle Neyroud and all collaborators at the Geneva Cancer Registry as well as Michel Coleman and all collaborators at the London School of Hygiene and Tropical Medicine for their support and encouragements.

Table 1. Deaths distribution by age groups, years after diagnosis and cancer sites for women diagnosed with an invasive tumour between 1981 and 1991.

Figure 1. Estimation of net survival for the four localisations using both cause-specific and relative survival setting.

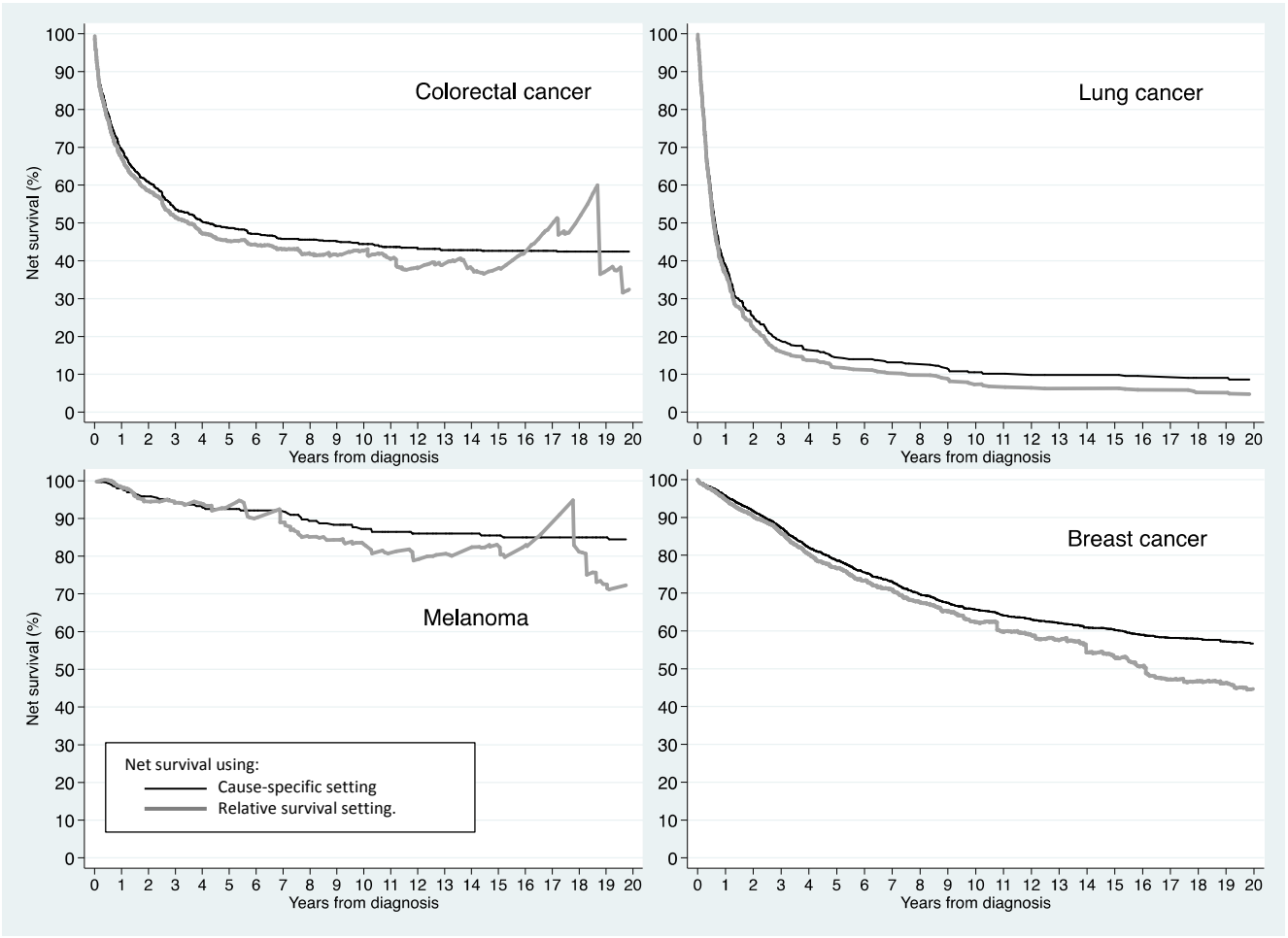
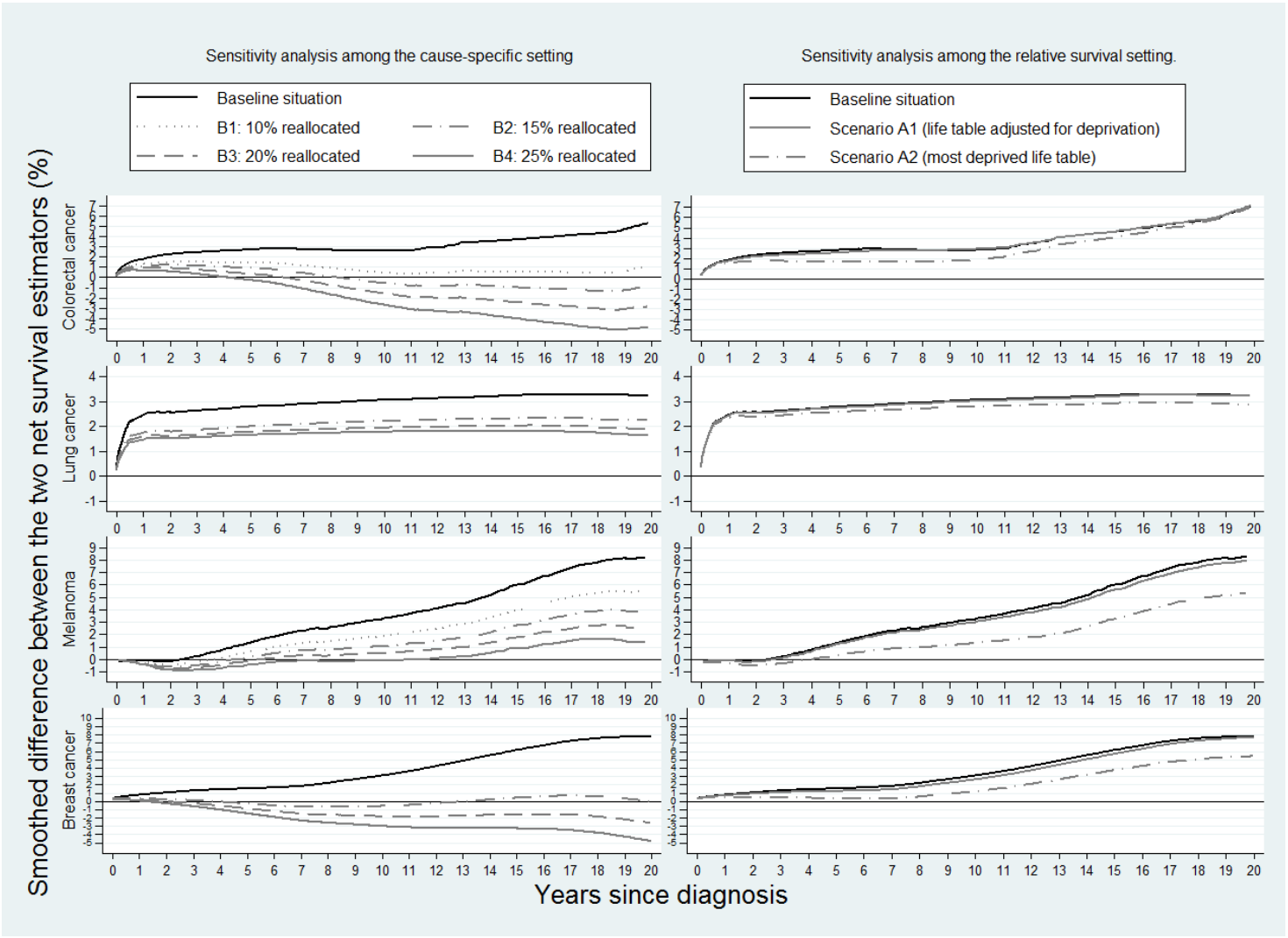


Figure 2: Sensitivity analysis among both cause-specific and relative survival setting.



References

[1] J. Berkson and R. P. Gage, “Calculation of survival rates for cancer,” 1952.

[2] F. Ederer and S. J. Cutler, “The relative survival rate: A statistical methodology,” in *National cancer institute monograph no. 6*, 1961, pp. 101–121.

[3] M. P. Perme, J. Stare, and J. Estève, “On estimation in relative survival.,” *Biometrics*, vol. 68, no. 1, pp. 113–20, Mar. 2012.

[4] C. Danieli, L. Remontet, N. Bossard, L. Roche, and A. Belot, “Estimating net survival: the importance of allowing for informative censoring.,” *Stat. Med.*, vol. 31, no. 8, pp. 775–86, Apr. 2012.

[5] R. Schaffar, B. Rachet, A. Belot, and L. Woods, “Cause-specific or relative survival setting to estimate population-based net survival from cancer? An empirical evaluation using women diagnosed with breast cancer in Geneva between 1981 and 1991 and followed for 20 years after diagnosis.,” *Cancer Epidemiol.*, vol. 39, no. 3, pp. 465–72, Jun. 2015.

[6] R. Schaffar, E. Rapiti, B. Rachet, and L. Woods, “Accuracy of cause of death data routinely recorded in a population-based cancer registry: impact on cause-specific survival and validation using the Geneva Cancer Registry.,” *BMC Cancer*, vol. 13, no. 1, p. 609, Jan. 2013.

[7] J. Robins, “Information recovery and bias adjustemnt in proportional hazards regression analysis of randomized trials using surrogate markers.,” *Proc. Biopharmmaceutical Sect. Am. Stat. Assoc.*, 1993.

[8] G. Satten, “Estimating the marginal survival function in the presence of time dependent covariates,” *Stat. Probab. Lett.*, vol. 54, no. 4, pp. 397–403, Oct. 2001.

[9] W. Cleveland, “Robust locally weighted regression and smoothing scatterplots,” *J. Am. Stat. Assoc.*, vol. 74, pp. 829–836, 1979.

[10] K. D. Skyrud, F. Bray, and B. Møller, “A comparison of relative and cause-specific survival by cancer site, age and time since diagnosis.,” *Int. J. Cancer*, Dec. 2013.

[11] N. Howlader, L. a G. Ries, A. B. Mariotto, M. E. Reichman, J. Ruhl, and K. a Cronin, “Improved estimates of cancer-specific survival rates from population-based data.,” *J. Natl. Cancer Inst.*, vol. 102, no. 20, pp. 1584–98, Oct. 2010.