

# Statistical Methods for Population-Based Cancer Survival in Registry Data

Bin Huang<sup>1</sup>, Jing Guo<sup>1</sup> and Richard Charnigo<sup>1,2\*</sup>

<sup>1</sup>Department of Biostatistics, University of Kentucky, USA

<sup>2</sup>Department of Statistics, University of Kentucky, USA

## Background

Population-based cancer registries play an important role in improving patient care programs and cancer care policies nationally and internationally. Accurate data from population-based cancer registries make it possible to monitor cancer trends, examine patterns of care, estimate survival and provide evidence-based outcomes for clinicians, public health administrators and policy makers.

A general interest in population-based cancer survival is net cancer specific survival which is the probability of surviving cancer in the absence of other causes of death. Because it is not influenced by changes in mortality from other causes, net survival provides a useful measure for tracking survival over time resulting from treatments and screening interventions, and making comparisons between different racial groups and regions. Two general approaches have been used to estimate net survival: cause-specific survival and relative survival.

Cause-specific survival is calculated by specifying the cancer related cause of death as the event. Cause-specific survival has been commonly used in clinical trials to examine efficacy of treatments. Clinical trials include a detailed review of the medical records to ascertain the causes of death, while population-based registries depend on death certificates which have inherent inaccuracies. This makes cause-specific survival a much less appealing approach in population-based cancer survival.

To circumvent the problems inherent in using causes of death from death certificates, relative survival has been the more popular approach, although its utility is not restricted to studying cancer. Assuming that the general population dies of causes other than the cancer of interest with the same probability as the cancer population, population life tables can be used to statistically factor out the probability of death from cancer and other causes. Relative survival is estimated as the ratio of the observed survival of the patients where all deaths are considered events to the expected survival from the background population life tables.

## Relative Survival

In the last thirty years, rapid and significant progress has been made in the development of statistical methodology in relative survival and its utilizations. Many articles using relative survival have reported on clinical and population-based outcomes research [1-5]. Relative survival, which estimates the net mortality associated with diagnosis of cancer in terms of excess mortality, is the difference between the total mortality experienced by the patients and the expected mortality of a comparable group from the general population, matched to the patients with respect to the main factors affecting patient survival but assumed to be practically free of the cancer of interest. The major advantages of relative survival are information on cause of death is not required and that it provides a measure of the excess mortality experienced by patients diagnosed with cancer irrespective of whether the excess mortality is directly or indirectly attributable to the cancer.

Relative survival is commonly defined as the ratio between

observed survival of the patients and expected survival with the matched background population:  $S_R(t) = S_o(t) / S_p(t)$ , where  $S_o(t)$  is the observed survival and  $S_p(t)$  is the expected survival. Three widely used methods are Ederer I, Ederer II and Hakulinen. The Ederer I approach was proposed in 1961, in which the cumulative expected survival estimation is derived by the average of expected probabilities for each individual at the beginning of follow-up and patients were considered at risk until the end of study [6]. Although the Ederer I method provides unbiased estimation for expected survival, it tends to overestimate the relative survival ratio since observed survival estimation is potentially biased; the Ederer I method does not account for the potentially heterogeneous withdrawal pattern of the patients. The Ederer II method estimates expected survival by multiplication of interval specific expected survival [7]. In this method, the expected survival estimates for patients are calculated at each point of follow-up so the matched individuals are considered to be at risk until the corresponding cancer patients die or are censored. Expected survival estimation from the Ederer II approach is biased, and relative survival may be underestimated. Since old patients are more likely to be censored than young patients by deaths from other causes, the censoring process becomes informative and leads to biased estimation of the net survival. To take into account informative censoring, the Hakulinen method produces the expected survival rates where the follow up times have been censored when the patients cannot be followed any longer [8]. In essence, the Hakulinen method introduces a biased estimator for the expected survival rates with bias similar to that in estimating observed survival by calculating potential follow-up times of patients. Properly estimating potential follow-up times is difficult, however, especially for diseased patients.

These three methods differ only in the estimation of expected survival regarding how long the matched individuals are considered to be at risk. For short term estimation, for example 5 year relative survival, there are minor differences among the three methods. In the Ederer I and Hakulinen methods, matched individuals are considered to be at risk for the entire follow-up. Because of that, estimates from the two methods tend to increase in the long term which reflect better health conditions of the long-term cancer survivors than those of the general population [9]. The Ederer II has been identified as preferable to control for increasing relative survival in long term estimation [9,10].

**\*Corresponding author:** Richard Charnigo, Department of Statistics, University of Kentucky, Room 203, Multidisciplinary Science Building, Lexington, Kentucky 40536, USA; Tel: 859-218-2072; E-mail: [RJCharn2@aol.com](mailto:RJCharn2@aol.com)

Received May 12, 2014; Accepted May 13, 2014; Published May 19, 2014

**Citation:** Huang B, Guo J, Charnigo R (2014) Statistical Methods for Population-Based Cancer Survival in Registry Data. J Biomet Biostat 5: e129. doi:10.472/2155-6180.1000e129

**Copyright:** © 2014 Huang B, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.



A more recent development in estimating net survival is the Pohar Perme approach [11]. If the risk of dying from cancer and the risk of dying from other causes are dependent, the patients with relatively high risks of dying from other causes will be removed early from the at-risk group whereas they may also have high risk of dying from cancer. If not taking into account the informative censoring mechanism, survival estimation will be biased towards the survival of the group with the lowest risk of dying from other causes. The Pohar Perme approach is unbiased for net survival even with informative censoring induced by variables in life tables as long as non-cancer deaths are correctly evaluated by the life tables. The Pohar Perme estimator is a weighted version of the Ederer II estimator which modifies the cumulative population hazard with inverse probability weighting. Estimates of Pohar Perme are similar to estimates from the Ederer I & II and Hakulinen methods for five year survival but the differences were quite large for 15 year survival [12].

## Cause-Specific Survival

In cause-specific survival, individuals who die of cause other than those specified are considered to be censored. The most common approach of ascertaining cause-specific mortality is by obtaining information from death certificates. In general, the standardized approach with the International Classification of Disease (ICD) for coding cause of death and associated conditions was used to record underlying causes of death. However, information is often inaccurately captured because of the flawed collection process, as death certificates are completed by persons with different backgrounds, such as physicians, medical examiners, coroners, and funeral directors [13].

Although relative survival has been the more popular approach to estimate net survival, difficulties obtaining accurate life tables motivated researchers to once again look at cause-specific survival as an alternative approach. In a recent study published by researchers from the National Cancer Institute (NCI), cause-specific survival estimates were calculated based on modified causes of death [14]. In the study, a classification scheme for the cause-specific death classification variable was developed by taking into account cause of death in conjunction with tumor sequence, site of original cancer diagnosis, and comorbidity factors (see <http://seer.cancer.gov/causespecific/index.html>).

Estimates from the cause-specific methods were compared with estimates from relative survival methods. For most cancers, the cause-specific survival approach produced accurate estimates similar to the relative survival. However, for several cancer sites, large differences were identified [14]. For example, the relative survival estimates for early stage breast cancer were higher than the cause-specific survival estimates. This is primarily because of the “healthy screener effect”—patients diagnosed with early-stage breast cancer through a screening examination tend to be healthier than the general population. Therefore, the expected survivals based on the general life tables were underestimated which led to overestimated relative survival.

## Estimation of Net Survival in Practice

Net survival is an unobservable measure and is rather difficult to estimate. Inaccurate or missing cause of death limited the value of estimates from cause-specific survival. Although the modified cause-specific approach with reclassifications of cause of death provided reliable estimates, the reclassification scheme was somewhat subjective and may not be valid for populations from different regions or counties. International comparisons are also difficult because of the variation in coding practices among countries [15]. Cause-specific survival

does have an advantage in estimating survival as compared to relative survival for special populations such as heavily screened populations and minority racial subgroups for which accurate life tables are not available. Another advantage for cause-specific survival is traditional modeling approaches can be easily utilized and interpreted.

Relative survival does not require cause of death but life tables of the background population are needed. A common assumption in relative survival analysis is that the excess and the population hazard are not affected by any common covariates. In practice, excess hazard almost always depends on demographic variables. Hence, the Pohar Perme estimates, which take into account informative censoring, are more accurate compared to estimates from the Ederer I, Ederer II and Hakulinen methods. In practice, there are few differences in short term survival estimates among these four methods [12].

The key disadvantage in relative survival is its dependence on accurate and representative life tables. Standard life tables usually include variables such as age, gender, calendar year and possibly race. Standard life tables are often only available at the national level. Developing reliable regional or characteristic specific life tables is not easy and may not be possible due to quality of population and mortality data and small numbers. That is why it is not surprising to find that relative survival ratios for early stage breast cancer or prostate cancer were greater than 100% because the matched population in the life table was not representative of the patients [16]. Life tables for the background population often include death from the cancer of interest, violating the assumption that life tables are based on the background population free of death from the cancer of interest. This is not a problem when cancer death is rare among all causes of death. The violation may not be ignorable for an elderly population when cancer is one of the primary causes of death, particularly when calculating survival for all cancers combined. Smoking related cancers have the most obvious bias in estimating net survival. Patients with these cancers are more likely to be smokers hence are more likely to have a higher background mortality rate than the general population. The relative survival for smoking-related cancers is likely to be underestimated because general population life-tables are used to estimate the background mortality for these patients. So far, there is no good solution to resolve the issue other than using smoker specific life tables which are hard to produce.

Since relative survival estimates likely depend on the age distribution in the population of interest, age standardization is necessary to compare relative survival across different populations. Age standardized relative survival estimates are weighted averages of age specific relative survival estimation within subgroups of patients defined by age at diagnosis, with weights equal to the proportions of patients in those subgroups in some standard population [17,18]. International Cancer Survival Standard Weights were developed for different ranges of cancer sites and comprise three sets of age weights [19]. However, when data is too sparse for either the total number of patients at risk or the number of patients that died, age standardization of relative survival could not be implemented.

## Future Research

Although the Pohar Perme approach provides unbiased estimation for net survival, the Pohar Perme estimator has larger variability than other methods [11]. This issue is more evident when dealing with long term follow-up or when age-standardization is utilized [20]. Having an unbiased estimator with reduced variation or a slightly biased estimator with small variation would be desirable, particularly for an elderly population and long term follow-up.



## Acknowledgements

This work was supported in part by grants from the Centers of Disease Control and Prevention National Center (5U48DP001932 SIP11-040) and the National Center for Advancing Translational Sciences, National Institutes of Health (UL1TR000117). The content is solely the responsibility of the authors and does not necessarily represent the official views of the CDC and NIH.

## References

- Baldwin LA, Huang B, Miller RW, Tucker T, Goodrich ST, et al. (2012) Ten-year relative survival for epithelial ovarian cancer. *Obstet Gynecol* 120: 612-618.
- Coleman MP, Quaresma M, Berrino F, Lutz JM, De Angelis R, et al. (2008) Cancer survival in five continents: a worldwide population-based study (CONCORD). *Lancet Oncol* 9: 730-756.
- Dialla PO, Dabakuyo TS, Marilier S, Gentil J, Roignot P, et al. (2012) Population-based study of breast cancer in older women: prognostic factors of relative survival and predictors of treatment. *BMC Cancer* 12: 472.
- Minicozzi P, Bella F, Toss A, Giacomini A, Fusco M, et al. (2013) Relative and disease-free survival for breast cancer in relation to subtype: a population-based study. *J Cancer Res Clin Oncol* 139: 1569-1577.
- Rutter CM, Johnson EA, Feuer EJ, Knudsen AB, Kuntz KM, et al. (2013) Secular trends in colon and rectal cancer relative survival. *J Natl Cancer Inst* 105: 1806-1813.
- Ederer F, Axtell LM, Cutler SJ (1961) The relative survival rate: a statistical methodology. *Natl Cancer Inst Monogr* 6: 101-121.
- Ederer F, Heise H (1959) Instructions to IBM 650 programmers in processing survival computations. *Methodological note*.
- Hakulinen T (1982) Cancer survival corrected for heterogeneity in patient withdrawal. *Biometrics* 38: 933-942.
- Hakulinen T, Seppä K, Lambert PC (2011) Choosing the relative survival method for cancer survival estimation. *Eur J Cancer* 47: 2202-2210.
- Cho H, Howlader N, Mariotto AB, Cronin KA (2011) Estimating relative survival for cancer patients from the SEER Program using expected rates based on Ederer I versus Ederer II method. *Surveillance Research Program, NCI, Technical Report #2011-01*.
- Perme MP, Stare J, Estève J (2012) On estimation in relative survival. *Biometrics* 68: 113-120.
- Roche L, Danieli C, Belot A, Grosclaude P, Bouvier AM, et al. (2013) Cancer net survival on registry data: use of the new unbiased Pohar-Perme estimator and magnitude of the bias with the classical methods. *Int J Cancer* 132: 2359-2369.
- Smith Sehdev AE, Hutchins GM (2001) Problems with proper completion and accuracy of the cause-of-death statement. *Arch Intern Med* 161: 277-284.
- Howlader N, Ries LA, Mariotto AB, Reichman ME, Ruhl J, et al. (2010) Improved estimates of cancer-specific survival rates from population-based data. *J Natl Cancer Inst* 102: 1584-1598.
- Bhalla K, Harrison JE, Shahraz S, Fingerhut LA; Global Burden of Disease Injury Expert Group (2010) Availability and quality of cause-of-death data for estimating the global burden of injuries. *Bull World Health Organ* 88: 831-838C.
- Brenner H, Arndt V (2005) Long-term survival rates of patients with prostate cancer in the prostate-specific antigen screening era: population-based estimates for the year 2000 by period analysis. *J Clin Oncol* 23: 441-447.
- Coebergh JXW, Sant M, Berrino F, Verdecchia A (1998) Special issue: Survival of adult cancer patients in Europe diagnosed from 1978-1989: The EURO CARE II study-Foreword. *Eur J Cancer* 34: 2137-2138.
- Berrino F, Gatta G, Sant M, Capocaccia R (2001) The EURO CARE study of survival of cancer patients in Europe: aims, current status, strengths and weaknesses. *Eur J Cancer* 37: 673-677.
- Corazziari I, Quinn M, Capocaccia R (2004) Standard cancer patient population for age standardising survival ratios. *Eur J Cancer* 40: 2307-2316.
- Dickman PW, Lambert PC, Coviello E, Rutherford MJ (2013) Estimating net survival in population-based cancer studies. *Int J Cancer* 133: 519-521.