Cancer Epidemiology 48 (2017) 16–21

Contents lists available at ScienceDirect

Cancer Epidemiology

The International Journal of Cancer Epidemiology, Detection, and Prevention

journal homepage: www.cancerepidemiology.net

Original research article

Correcting bias due to missing stage data in the non-parametric estimation of stage-specific net survival for colorectal cancer using multiple imputation

Milena Falcaro[a,*], James R. Carpenter[b,c]

[a] University College London, UK
[b] London School of Hygiene and Tropical Medicine, UK
[c] MRC Clinical Trials Unit at UCL, London, UK

ARTICLE INFO

Article history:
Received 22 June 2016
Received in revised form 30 January 2017
Accepted 17 February 2017
Available online 19 March 2017

Keywords:
Cancer
Informative censoring
Multiple imputation
Net survival
Pohar-Perme estimator
Uncongeniality

ABSTRACT

Background: Population-based net survival by tumour stage at diagnosis is a key measure in cancer surveillance. Unfortunately, data on tumour stage are often missing for a non-negligible proportion of patients and the mechanism giving rise to the missingness is usually anything but completely at random. In this setting, restricting analysis to the subset of complete records gives typically biased results. Multiple imputation is a promising practical approach to the issues raised by the missing data, but its use in conjunction with the Pohar-Perme method for estimating net survival has not been formally evaluated.
Methods: We performed a resampling study using colorectal cancer population-based registry data to evaluate the ability of multiple imputation, used along with the Pohar-Perme method, to deliver unbiased estimates of stage-specific net survival and recover missing stage information. We created 1000 independent data sets, each containing 5000 patients. Stage data were then made missing at random under two scenarios (30% and 50% missingness).
Results: Complete records analysis showed substantial bias and poor confidence interval coverage. Across both scenarios our multiple imputation strategy virtually eliminated the bias and greatly improved confidence interval coverage.
Conclusions: In the presence of missing stage data complete records analysis often gives severely biased results. We showed that combining multiple imputation with the Pohar-Perme estimator provides a valid practical approach for the estimation of stage-specific colorectal cancer net survival. As usual, when the percentage of missing data is high the results should be interpreted cautiously and sensitivity analyses are recommended.

1. Introduction

Net survival, namely the probability of survival in the hypothetical situation where patients can only die of the disease under investigation, plays a fundamental role in cancer survival studies. Its estimation poses several challenges. First, it requires the handling of competing mortality risks because death can occur for reasons other than cancer. Secondly, these competing risks are almost always mutually correlated, which results in an informative censoring mechanism that cannot be safely ignored [1]. In addition, analyses may be further complicated by the unavailability or unreliability of information on cause of death. In population-based cancer registry studies this is usually handled via a so-called relative survival approach, which consists in estimating the excess mortality experienced by the cancer patients as compared to the mortality expected in a comparable general population. The advantage of this approach is that it does not require an accurate recording of the cause of death for the cancer patients.

Various methods have been devised for the estimation of net survival in the relative survival setting [1–3]. Pohar Perme et al. [1] proposed an unbiased non-parametric estimator that adjusts for informative censoring via inverse probability weighting. Danieli et al. [4] and Roche et al. [5] recommended this method especially for routine net survival estimations by cancer registries. This estimator is particularly convenient when the analyst is not

* Correspondence to: Department of Primary Care and Population Health, UCL Medical School, Rowland Hill Street, London, NW3 2PF, UK.
E-mail address: m.falcaro@ucl.ac.uk (M. Falcaro).

http://dx.doi.org/10.1016/j.canep.2017.02.005
1877-7821/

interested in evaluating covariate effects but merely seeks to estimate a summary measure (e.g. net survival or cumulative excess hazard) for all patients or for groups of patients. For instance, it can be used to estimate net survival by tumour stage at diagnosis, a measure which is of great importance for cancer surveillance and health planning and evaluation [6,7].

Although completeness of stage has considerably improved in recent years in many cancer registries, stage is often unavailable for a non-trivial number of patients. For example, in a recent series of papers by the International Cancer Benchmarking Partnership [6,7] the authors excluded from the analysis some of the cancer registries because of their high percentage of missing stage information. In particular, Maringe et al. [6] focused on colorectal cancer survival and excluded the registries that had less than 50% of patients with recorded stage data in the study period. Unfortunately, missingness on tumour stage is typically not completely at random. For example, older and more frail patients with relatively poor prognosis may be less likely to receive a thorough staging investigation [8]. Restricting the analysis to patients with complete records can lead to misleading results [8,9]. This situation is exacerbated when calculating net survival, where complete records analysis is only valid when data are missing completely at random. While multiple imputation has been successfully applied to parametric relative survival settings [8,9], to the best of our knowledge, no work has yet been published on the non-parametric estimation of stratum-specific (e.g. stage-specific) net survival when the stratification variable is not fully observed.

In this paper we report a resampling study from an extract of a population-based cancer registry data set. The aim is to evaluate the ability of multiple imputation [10,11], used in conjunction with the Pohar-Perme estimator of net survival, to reduce bias and improve confidence interval coverage when a key covariate (tumour stage) is missing at random.

Our proposed approach combines parametric imputation with a non-parametric estimator of net survival. This makes it an uncongenial imputation strategy [12]. Several authors [13,14] have argued that, unless the imputation model is grossly misspecified, uncongenial strategies like ours may perform better and be more robust than methods where missingness and estimation are handled in a "single step". However, it is important to evaluate the performance of our approach empirically; this is especially the case as it is unclear how to perform an efficient "single step" analysis for the non-parametric Pohar-Perme estimator.

The paper is structured as follows. We start by briefly introducing the Pohar-Perme estimator. Next, we describe the resampling design and the analysis setting. We then report our results and conclude with a discussion of our findings.

2. Methods

2.1. The Pohar-Perme estimator

In the relative survival setting the total hazard at time t, here denoted by $\lambda^*(t)$, is usually decomposed as

$$\lambda^*(t) = \lambda_E(t) + \lambda_P(t) \quad (1)$$

where $\lambda_E(t)$ is the excess or cancer-related hazard and $\lambda_P(t)$ represents the background or expected hazard. Two data sources are then used: $\lambda^*(t)$ is estimated from the cancer registry data, whereas $\lambda_P(t)$ is treated as a known quantity and is retrieved from the life tables of a comparable general population, usually matched to the cancer patients by at least age, sex, calendar time and geographical area [15]. The excess hazard is derived as the difference between the estimated total hazard and the expected hazard. By integrating over time we obtain the cumulative excess hazard $\Lambda_E(t)$ as

$$\Lambda_E(t) = \Lambda^*(t) - \Lambda_P(t),$$

where $\Lambda^*(t)$ is the total cumulative hazard and $\Lambda_P(t)$ is the expected cumulative hazard. Until recently, the decomposition (1) and the estimation of the excess hazard were commonly made by assuming independence between the cancer and non-cancer mortality processes. Pohar Perme et al. [1] argued that these two processes are very likely to be correlated, giving rise to an informative censoring that could grossly bias the results if ignored. To overcome this problem they proposed to adjust the continuous version of the Ederer II estimator [16] by using inverse probability of censoring weights [17], where the weights are the reciprocal of the individual-specific expected survival probabilities. Without going into much detail, the Ederer II estimator of $\Lambda_E(t)$ can be derived as the difference between the Nelson-Aalen estimator of $\Lambda^*(t)$ and the cumulative expected hazard of the patients still at risk at each failure. More details can be found in Pohar Perme et al. [1] and Rebolj Kodre and Pohar Perme [18].

2.2. Resampling study

2.2.1. The data

The population for our resampling study was extracted from four English cancer registries and consists of 50,387 male patients who were diagnosed with colorectal cancer between 1996 and 2006 with follow-up until the end of 2009 and for whom we had complete information on age at diagnosis, survival time, vital status, stage at diagnosis and deprivation quintile (based on the income domain of the Index of Multiple Deprivation). Table 1 summarises the data.

Table 1
Descriptive statistics of the complete cancer registry data set used for the resampling study.

	All patients	Patients with stage 1	stage 2	stage 3	stage 4
Overall	50387	13.9%	32.4%	30.0%	23.7%
Deaths	32267	8.6%	25.7%	30.2%	35.5%
Deprivation					
1 – least deprived	10599	15.0%	32.2%	31.2%	21.6%
2	10773	14.9%	32.4%	30.6%	22.1%
3	9914	14.0%	33.5%	29.2%	23.3%
4	9983	13.0%	33.0%	29.3%	24.7%
5 – most deprived	9118	12.4%	30.7%	29.3%	27.6%
Age at diagnosis					
Median	70.9	70.7	72.1	70.2	70.3
IQR	(62.8,77.5)	(63.0,77.1)	(64.1,78.2)	(62.0,77.0)	(61.9,77.3)