Type of manuscript: original research article

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

Authors: Milena Falcaro[1*] and James R. Carpenter [2,3]

Affiliation:

[1]University College London, UK

[2]London School of Hygiene and Tropical Medicine, UK

[3]MRC Clinical Trials Unit at UCL, London, UK

** Corresponding author:*

Milena Falcaro
Department of Primary Care and Population Health
UCL Medical School
Rowland Hill Street
London NW3 2PF, UK
E-mail: m.falcaro@ucl.ac.uk

Suggested running head: multiple imputation for missing stage in net survival

Sources of financial support: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. MF is supported by the National Institute for Health Research (NIHR) School for Public Health Research. JRC is supported by the Medical Research Council.

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

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.

Key words: cancer, informative censoring, multiple imputation, net survival, Pohar-Perme estimator, uncongeniality.

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 [1,2,3]. 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 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 50387 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.

The background general population mortality rates for our relative survival analysis were retrieved from life tables for England stratified by age, sex, calendar year, region and deprivation.

	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)	(63,77.1)	(64.1,78.2)	(62.0,77.0)	(61.9,77.3)

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

2.2.2 Missing data generation and analysis setting

From our fully observed data set of 50387 cancer patients we created 1000 independent random samples, each of 5000 patients. We then introduced missing tumour stage values. We considered two scenarios depending on whether the overall rate of incomplete stage information was set to around 30% (scenario A) or 50% (scenario B). In both scenarios the missingness was assumed to depend only on observed quantities, i.e. to be missing at random (MAR). Specifically, for each of the 1000 independent resamples we proceeded as follows.

Step 1: we randomly sampled 5000 patients from the population.

Step 2: using a missing at random mechanism dependent on survival time, event indicator, age at diagnosis (linear and quadratic effect) and deprivation, we induced missing data on stage under two scenarios, A and B, giving around 30% and 50% of missing stage values respectively. See the Appendix for more details about the simulated missing data mechanisms.

Step 3: we conducted a complete records analysis (CRA) and obtained stage-specific net survival estimates at time t (t = 1, 2, …, 5 years) post diagnosis using the Pohar-Perme estimator.

Step 4: we carried out multiple imputation (MI). In more detail, we used a multinomial logistic imputation model for stage and included the Nelson-Aalen estimate of the cumulative hazard, the event indicator, dummy variables for deprivation and a restricted cubic spline function for age at diagnosis (knots placed at the 0^{th}, 33^{rd}, 67^{th} and 100^{th} centiles of the distribution). Other life-table variables (e.g. indicators for region) were initially included in the model but, since they were found to be non-significant, we dropped them from the final model. For both scenarios A and B we generated 100 imputed data sets.

Step 5: in each of the imputed data sets we estimated stage-specific net survival at $t = 1,...,5$ years after diagnosis using the Pohar-Perme method.

Step 6: we pooled the imputation-specific results. The MI point estimates and their 95% confidence intervals were derived by applying the Rubin's rules after a suitable transformation to improve normality and by then back transforming to the original scale.

A complementary log-log transformation is usually recommended for predicted survival probabilities [19] as it maps the interval (0, 1) to $(-\infty, +\infty)$. However, in the relative survival context the application of this transformation is sometimes problematic as numerical instability may arise if the probability is very close to 0 or 1. An additional complication is that estimates of net survival above 1 may also occasionally occur. In our study we faced these problems when we tried to obtain MI estimates for stage 1 as a few of the estimates to be transformed fell just below or above 1. We therefore decided to use a log transformation for stage 1 and a complementary log-log transformation for the other stage categories. For comparison, we also calculated the MI point and interval estimates using Rubin's rules without a prior transformation to normality.

Steps 1 to 6 were then repeated 1000 times. The results across these replications were compared to the reference values (here treated as the "true" values) obtained using the Pohar-Perme estimator on the fully observed population data. Under both scenarios we evaluated the performance of CRA and our MI method in terms of bias, coverage rate and average length of the 95% confidence intervals [20]. The bias is defined as the difference between the average of the estimates across the repeated samples and the reference value, while the coverage is the proportion of times the 95% confidence interval includes the reference value.

Coverage rates are considered acceptable if they are approximately not more than 2 standard errors away from the nominal coverage probability [20]. In our case the coverage rates should therefore fall approximately between 93.6% and 96.4% (i.e. $95\% \pm 2 * \sqrt{0.95*0.05/1000}$).

As a rule of thumb, it is recommended that the number of imputations *m* should be at least equal to the percentage of incomplete records in the dataset. For example, if there are 20% of records with missing values then we should set *m* to at least 20 [11]. However, White et al. [11] and Royston and White [21] pointed out that in simulations studies where the interest lies in comparing statistical methods larger values of *m* are needed. In our study we therefore set *m*=100 but for real data analyses we recommend using the rule of thumb.

All the analyses were performed using Stata 13 [22]. In particular, the Pohar-Perme estimates were obtained using the *stns* command [23].

3. Results

Table 2 displays the findings of our resampling study. The application of the Rubin's rules with and without a prior normalizing transformation yielded similar results so, for simplicity, hereafter we only report those without the transformation. CRA led to severe bias and very poor coverage under both scenarios, clearly showing that this type of analysis should not be used when estimating stage-specific net survival from data with a non-trivial proportion of missing stage. We note that our MI strategy performed much better than CRA in terms of bias and coverage. Under scenario A, i.e. with around 30% of missingness, MI succeeded in recovering the missing stage information from the incomplete records, all empirical coverage rates being satisfactory and the largest relative bias being 6.4%. Some of the coverage rates

(especially those for stage 1) were higher than 96.4%, suggesting that the MI strategy may sometimes be too conservative.

Increasing the proportion of missing values to 50% (scenario B) somehow worsened the performance of MI for stage 4, for which we now observed coverage rates as low as 84.9%. However, despite this under-coverage, MI still greatly outperformed CRA. Indeed, under scenario B, the coverage rates from CRA did not go above 8.1%.

	Full data	Scenario A (30% missing values)								Scenario B (50% missing values)							
	(reference	Complete records analysis				Multiple imputation				Complete records analysis				Multiple imputation			
	value)	*bias*	*rbias*	*avL*	*coverage*	*bias*	*rbias*	*avL*	*coverage*	*bias*	*rbias*	*avL*	*coverage*	*bias*	*rbias*	*avL*	*coverage*
stage 1																	
$S_1(1)$	94.50	3.07	3.2%	4.09	17.6%	0.03	0.0%	5.66	98.8%	4.52	4.8%	3.89	2.0%	-0.04	0.0%	6.60	99.5%
$S_1(2)$	93.63	4.34	4.6%	5.32	13.6%	-0.64	-0.7%	7.28	97.5%	6.59	7.0%	5.17	0.4%	-1.27	-1.4%	8.73	97.4%
$S_1(3)$	92.26	5.48	5.9%	6.61	11.0%	-0.84	-0.9%	8.59	97.3%	8.50	9.2%	6.57	0.3%	-1.77	-1.9%	10.32	95.0%
$S_1(4)$	90.71	6.38	7.0%	8.00	13.1%	-0.77	-0.9%	9.81	97.3%	10.29	11.3%	8.11	0.3%	-1.66	-1.8%	11.72	95.7%
$S_1(5)$	89.13	7.21	8.1%	9.59	18.8%	-0.46	-0.5%	11.15	96.7%	11.86	13.3%	9.87	0.6%	-1.25	-1.4%	13.18	97.1%
stage 2																	
$S_2(1)$	90.03	4.75	5.3%	3.37	0.1%	0.14	0.2%	4.41	97.8%	7.13	7.9%	3.34	0.0%	0.10	0.1%	5.18	98.8%
$S_2(2)$	86.29	6.52	7.6%	4.41	0.0%	-0.41	-0.5%	5.47	96.4%	10.13	11.7%	4.47	0.0%	-0.85	-1.0%	6.50	95.7%
$S_2(3)$	82.73	7.78	9.4%	5.34	0.0%	-0.63	-0.8%	6.24	96.0%	12.49	15.1%	5.54	0.0%	-1.25	-1.5%	7.36	93.3%
$S_2(4)$	79.62	8.66	10.9%	6.24	0.0%	-0.61	-0.8%	6.92	96.0%	14.25	17.9%	6.60	0.0%	-1.26	-1.6%	8.07	94.2%
$S_2(5)$	76.97	9.24	12.0%	7.25	0.0%	-0.57	-0.7%	7.67	95.6%	15.54	20.2%	7.77	0.0%	-1.20	-1.6%	8.82	94.9%
stage 3																	
$S_3(1)$	83.00	5.70	6.9%	4.36	0.0%	-0.69	-0.8%	5.25	95.7%	8.97	10.8%	4.58	0.0%	-1.09	-1.3%	6.16	94.3%
$S_3(2)$	71.38	8.06	11.3%	5.69	0.0%	-0.10	-0.1%	6.21	96.9%	13.45	18.8%	6.18	0.0%	-0.14	-0.2%	7.23	97.3%
$S_3(3)$	62.23	9.00	14.5%	6.51	0.0%	0.26	0.4%	6.65	96.6%	15.74	25.3%	7.27	0.0%	0.64	1.0%	7.69	95.3%
$S_3(4)$	55.91	9.32	16.7%	7.10	0.0%	0.40	0.7%	6.98	96.2%	16.84	30.1%	8.08	0.0%	0.99	1.8%	8.01	94.5%
$S_3(5)$	51.67	9.37	18.1%	7.70	0.3%	0.32	0.6%	7.37	96.6%	17.35	33.6%	8.87	0.0%	0.93	1.8%	8.41	94.9%
stage 4																	
$S_4(1)$	37.29	6.30	16.9%	7.56	7.8%	0.38	1.0%	6.44	97.9%	10.89	29.2%	9.85	0.6%	0.88	2.4%	7.26	95.5%
$S_4(2)$	18.39	5.05	27.5%	6.53	11.5%	0.81	4.4%	5.32	94.1%	9.55	51.9%	8.93	0.8%	1.62	8.8%	6.15	86.6%
$S_4(3)$	10.66	3.84	36.1%	5.51	18.7%	0.68	6.4%	4.33	93.6%	7.65	71.8%	7.80	1.3%	1.42	13.3%	5.06	84.9%
$S_4(4)$	7.50	3.10	41.3%	4.95	28.9%	0.43	5.7%	3.77	96.3%	6.33	84.4%	7.13	3.5%	0.92	12.3%	4.40	90.6%
$S_4(5)$	5.90	2.62	44.3%	4.72	39.2%	0.25	4.2%	3.51	96.8%	5.47	92.6%	6.84	8.1%	0.55	9.3%	4.06	95.0%

Table 2. Results for CRA and MI under scenarios A (30% missingness) and B (50% missingness). $S_r(t)$ denotes net survival (%) for stage=r and time=t years after diagnosis (r=1,…,4; t=1,…,5). The reference values correspond to the net survival estimates obtained using the fully observed population data. The performance of CRA and MI are evaluated in terms of bias, percentage relative bias (rbias), average length (avL) and coverage rate of the 95% CIs. The bias is defined as the difference between the average of the estimates across the repeated samples and the reference value, while the coverage is the proportion of times the 95% confidence interval includes the reference value.

4. Discussion

To the best of our knowledge, this is the first study to discuss how missing values should be handled in the context of the non-parametric Pohar-Perme estimator. Our results are very encouraging.

We focused on tumour stage at diagnosis for two reasons. Firstly, it is an important determinant of treatment and prognosis and so a key predictor of cancer survival. Information on stage is of vital importance for assessing the impact of early detection programs and for a better understanding of trends over time or differences in cancer survival across countries [6,7]. Secondly, stage is often missing for a non-trivial fraction of patients and the missingness is typically not completely at random.

In this setting, the ad-hoc and yet relatively popular approach of creating an extra category for the missing values has been shown to lead to severe bias [8,24]. Further, as our findings show, the other popular approach of estimating net survival using complete records results in biased estimates and poor confidence interval coverage. Complete records analysis should therefore be avoided for the estimation of stage-specific net survival when the percentage of missing stage values is non-trivial.

Another option is needed for researchers. While multiple imputation is a natural and practical approach, its performance in this setting needs to be evaluated before it can be recommended for routine use. This is because the parametric imputation model is uncongenial [12] with the non-parametric Pohar-Perme estimator. Despite this, because of the practicality of MI, we were motivated to explore this approach by Schafer's statement "Experience suggests that Bayesian MI does interact well with a variety of semi- and nonparametric estimation procedures" [14].

We specified the imputation model in a similar manner to that proposed by Falcaro *et al.* [9] for the estimation of stage-specific net survival via a flexible proportional hazards model

[25,26]. In line with theory [10] we used a multinomial logistic imputation model for stage and included (i) predictors of both the values of the incomplete variable and whether it was missing, (ii) the variables affecting the inverse probability of censoring weights and (iii) the outcome. For the latter we followed the work of White and Royston [27] and incorporated the event indicator and the Nelson-Aalen cumulative hazard estimate. This can easily be derived in standard statistical software. In Stata, for example, before carrying out the imputation we can stset the data and use the “sts gen H=na” command to generate a new variable H containing the Nelson-Aalen estimate for each patient.

Encouragingly, our MI strategy reduced the bias to a practically negligible level, and vastly improved confidence interval coverage. Overall, MI confidence interval coverage was close to, or slightly above, 95%. This is in line with the slightly conservative behaviour typically found with uncongenial imputation (see chapter 2 in [10]). Only for two estimates of net survival with tumour stage 4, under the 50%-missingness scenario, did the MI confidence interval coverage drop below 87%. This is because the missing data mechanism was such that, while 50% of values were missing overall, stage 4 was the category with the highest proportion of missingness (around 64%). The large number of missing values meant that the impact of the approximation implicit in the White and Royston approach of including the Nelson-Aalen cumulative hazard estimate and the event indicator in the imputation model became detectable in a small increase in bias and corresponding reduction in confidence interval coverage.

Our results show MI gives reliable inferences with this high proportion of missing data and therefore provide confidence that MI will give reliable inferences when a lower proportion of data are missing. They further suggest that our MI approach can be expected to perform reasonably well also in the non-parametric cause-specific survival setting.

Some readers may feel that the proportion of missing stage data we chose was too high, given the improvement in capturing stage data in recent years. However, many researchers are

still interested in comparisons with earlier data. Moreover, many cancer registries in developing countries have far from the level of stage completeness observed in the UK, US or Scandinavian countries. With lower proportions of missing data, is MI worth bothering with? Since (i) we can't know for sure the extent of bias in any specific case and (ii) the time and effort involved in MI are small in relation to gathering and cleaning the data, we would argue that MI should be used as a matter of routine if the percentage of missing values is not negligible.

It is however important to stress that the specification of an imputation model needs to be carefully tailored for each real data set under investigation to address the particular missing data mechanism at hand.

As we have commented elsewhere [8,9], it is usually implausible that stage is missing completely at random because its missingness is typically strongly associated with survival. The appropriateness of the MAR assumption is important to consider. In our setting it assumes that the distribution of stage, given survival and other variables, is the same whether or not it is observed (this interpretation of MAR is set out in [10]). This makes it a natural starting point for the analysis. Alongside this, in applications when a large proportion of data are missing and inferences are critical sensitivity analysis to plausible departures from MAR should be considered; some possible MI approaches are sketched in chapter 10 of [10] and will be developed further in future work.

We chose a re-sampling study (as opposed to simulating the population data) because, as pointed out by Marshall *et al.* [19] and Lee and Carlin [28], it offers the advantage of working with data that reflect the characteristics and variability of a realistic population. Resampling studies share many similarities with simulation studies where the data generating mechanism is an explicitly specified probability model. The main difference between these two techniques is that in the resampling framework the analyst draws the repeated samples from a real data set rather than generating the data from a theoretical probability distribution. Drawing samples from

an existing population avoids making the inevitable simplifying assumptions about how the population variables should be distributed and interrelated. When, as here, our existing population is large (n=50387) and our random samples are about 10% of this (n=5000), we get valid inferences from our samples for this population when no data are missing. Then, the missing data mechanism is under our control and we know that it is responsible for the inferential biases. We can then directly assess how successful multiple imputation is in correcting these. Inevitably, as with all simulation studies, our results could be further strengthened by experience with alternative populations, cancer sites and missing data mechanisms. Nevertheless, we believe that our study, being based on a real complex data set, gives valuable insights on the performance of multiple imputation when combined with the Pohar-Perme method. For colorectal cancer (and cancers with similar survival profiles) it provides robust evidence that the widespread use of MI would give substantial, scientifically important, improvements to the inferences made using currently popular methods for missing data. Further work is planned to generalize these findings to other cancer sites.

In passing, note that the complete population used for our resampling study was obtained by extracting the registry records of patients with no missing values on our key set of variables. We did this so to have full control on the missing data mechanism. This however means that the net survival estimates reported in this paper are not representative of the UK population or even the cancer registries from which the resampling population was extracted; but such representativeness was not the aim of our work.

In conclusion, this study demonstrates that MI offers a substantial, practically important, improvement on complete records analysis when faced with missing data in the non-parametric estimation of stage-specific colorectal cancer net survival.

Acknowledgments

The authors are very grateful to Prof. Michel Coleman and Dr Bernard Rachet (London School of Hygiene and Tropical Medicine) for providing the real cancer registry data considered in this paper, and to the referees whose comments have led to a greatly improved manuscript.

Bibliography

1. Pohar Perme M, Stare J, Estève J. On estimation in relative survival. Biometrics. 2012; 68: p. 113-120.
2. Estève J, Benhamou E, Croasdale M, Raymond L. Relative survival and the estimation of net survival: elements for further discussion. Statistics in Medicine. 1990; 9: p. 529-538.
3. Dickman P, Sloggett A, Hills M, Hakulinen T. Regression models for relative survival. Statistics in Medicine. 2004;(23): p. 51-64.
4. Danieli C, Remontet L, Bossard N, Roche L, Belot A. Estimating net survival: the importance of allowing for informative censoring. Statistics in Medicine. 2012; 31: p. 775-786.
5. Roche L, Danieli C, Belot A, Grosclaude P, Bouvier A, Velten M, et al. Cancer net survival on registry data: use of the new unbiased Pohar-Perme estimator and magnitude of the bias with the classical methods. International Journal of Cancer. 2013; 132: p. 2359-2369.
6. Maringe C, Walters S, Rachet B, et al.. Stage at diagnosis and colorectal cancer survival in six high-income countries: a population-based study of patients diagnosed during 2000-2007. Acta Oncologica. 2013; 52(5): p. 919-932.
7. Walters S, Maringe C, Butler J, et al.. Breast cancer survival and stage at diagnosis in Australia, Canada, Denmark, Norway, Sweden and the UK, 2000-2007: a population-based study. British Journal of Cancer. 2013; 108(5): p. 1195-1208.
8. Nur U, Shack L, Rachet B, Carpenter J, Coleman M. Modelling relative survival in the presence of incomplete data: a tutorial. International Journal of Epidemiology. 2010; 39: p. 118-128.
9. Falcaro M, Nur U, Rachet B, Carpenter J. Estimating excess hazard ratios and net survival when covariate data are missing: strategies for multiple imputation. Epidemiology. 2015; 26: p. 421-428.
10. Carpenter J, Kenward M. Multiple imputation and its application Chichester: John Wiley & Sons; 2013.
11. White I, Royston P, Wood A. Multiple imputation using chained equations: issues and guidance for practice. Statistics in Medicine. 2011; 30: p. 377-399.
12. Meng XL. Multiple-imputation inferences with uncongenial sources of input. Statistical Science. 1994; 9: p. 538-573.
13. Schafer J, Graham J. Missing data: our view of the state of the art. Psychological Methods. 2002; 7(2): p. 147-177.
14. Schafer J. Multiple imputation in multivariate problems when the imputation and analysis models differ. Statistica Neerlandica. 2003; 57: p. 19-35.
15. Sarfati D, Blakely T, Pearce N. Measuring cancer survival in populations: relative survival vs cancer-specific survival. International Journal of Epidemiology. 2010; 39: p. 598-610.

16. Ederer F, Axtell L, Cutler S. The relative survival rate: a statistical methodology. National Cancer Institute Monograph. 1961; 6: p. 101-121.
17. Robins J. Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. Proceedings of the Bio-pharmaceutical Section. 1993;: p. 24-33.
18. Rebolj Kodre A, Pohar Perme M. Informative censoring in relative survival. Statistics in Medicine. 2013; 32: p. 4791-4802.
19. Marshall A, Altman D, Royston P, Holder R. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. BMC Medical Research Methodology. 2010; 10: p. 7.
20. Burton A, Altman D, Royston P, Holder R. The design of simulation studies in medical statistics. Statistics in Medicine. 2006; 25: p. 4279-4292.
21. Royston P, White IR. Multiple Imputation by Chained Equations (MICE): implementation in Stata. Journal of Statistical Software. 2011; 45(4): p. 1-20.
22. StataCorp. Stata statistical software: release 13 College Station, TX: StataCorp LP; 2013.
23. Clerc-Urmès I, Grzebyk M, Hédelin G. Net survival estimation with stns. The Stata Journal. 2014; 14: p. 87-102.
24. Greenland S, Finkle W. A critical look at methods for handling missing covariates in epidemiologic regression analyses. American Journal of Epidemiology. 1995; 142(12): p. 1255-1264.
25. Royston P, Parmar M. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Statistics in Medicine. 2002; 21: p. 2175-2197.
26. Nelson C, Lambert P, Squire I, Jones D. Flexible parametric models for relative survival with application in coronary heart disease. Statistics in Medicine. 2007; 26: p. 5486-5498.
27. White I, Royston P. Imputing missing covariate values for the Cox model. Statistics in Medicine. 2009; 28: p. 1982-1998.
28. Lee K, Carlin J. Recovery of information from multiple imputation: a simulation study. Emerging Themes in Epidemiology. 2012; 9: p. 3.

Appendix

Let R denote the missing data indicator and Z be a vector of covariates. The missing values for tumour stage were induced with a probability generated using the model

$$logit(P(R = 1 \mid Z)) = \alpha + \beta\,\mathrm{T} + \gamma\,\mathrm{D} + \delta_1\,\mathrm{age} + \delta_2\,\mathrm{agesq} + \sum_{i=2}^{5} \nu_i\, I\{\mathrm{dep} = i\}$$

where T represents the observed survival time in years, D is the event indicator, age refers to standardised age at diagnosis, agesq is age squared and $I\{\mathrm{dep} = i\}$ is the indicator function equal 1 when dep (deprivation quintile) $= i$ and 0 otherwise (i = 1,…,5). The parameters were chosen as follows.

(a) scenario A (30% missingness):

$\alpha = -0.212, \beta = -0.25, \gamma = -0.1, \delta_1 = 0.35, \delta_2 = 0.12, \nu_2 = \nu_3 = 0, \nu_4 = 0.2$ and $\nu_5 = 0.3$.

(b) scenario B (50% missingness):

$\alpha = 0.78,\ \beta = -0.25, \gamma = -0.1, \delta_1 = 0.35,\ \delta_2 = 0.12, \nu_2 = \nu_3 = 0, \nu_4 = 0.2$ and $\nu_5 = 0.3$.

The equation defining the missing data mechanism is a logistic model. Therefore, if for example we set $\nu_5 = 0.3$, this corresponds to an odds ratio of 1.35 (i.e. $e^{0.3}$) for patients living in the most deprived areas (dep=5) versus those living in the most affluent areas (dep=1), conditional on other variables being held constant.