EUROPEAN JOURNAL OF CANCER 45 (2009) 3298–3302

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejconline.com

Correcting population-based survival for DCOs – Why a simple method works and when to avoid it

Paul Silcocks[a,b,*], *Catherine S. Thomson*[c,d]

[a]*Trent Cancer Registry, 5 Old Fulwood Road, Sheffield S10 3TG, England, United Kingdom*
[b]*Trent Research and Development Support Unit, University of Nottingham, NIHR Research Design Service for the East Midlands, 14th Floor, Tower Building, University Park, Nottingham NG7 2RD, England, United Kingdom*
[c]*Statistical Information Team, Cancer Research UK, 61 Lincoln's Inn Fields, London WC2A 3PX, United Kingdom*

ARTICLE INFO

Article history:
Received 24 December 2008
Received in revised form 9 June 2009
Accepted 12 June 2009
Available online 3 August 2009

Keywords:
Death certificates
England
Epidemiology
Cancer registries
Survival analysis

ABSTRACT

A high proportion of cancer registrations solely based on a death certificate (DCOs) indicates poor data quality and biases cancer survival estimates. Intensive trace-back of registrations initiated after death (DCIs) can reduce the proportion of DCOs to an acceptable level and also improve data quality in other areas (such as increasing the information on disease extent, morphology and treatment) but is expensive in staff time. Our approach – based on a proportional hazards model for DCOs relative to all other cases – can be used to predict what the likely effect of the trace-back will be on survival and to justify the extra work involved. It can also be used to correct results from other sources (including historical data) especially when these sources contain high percentages of DCOs. Of course, the ability to make this correction is no excuse for omitting trace-back of DCI cases when resources permit. With our model the true survival tends ultimately to $(1-p)*S$ where p is the proportion of DCOs and S is the observed survival, which is a simple correction noted by others. The worse the assumed survival of DCOs is relative to all other cases, the earlier is the time for the maximum difference between observed and true survival. Correction to the later part of survival curves is easy and an example is shown using EUROCARE data. This paper shows why the simple method works and suggests that researchers should always think about adjusting their survival estimates with regard to the percentage of DCOs. This paper also shows when the simple correction can (on 5-year survival estimates) and cannot (on 1-year survival estimates, generally) be used to adjust survival figures when comparisons are made across regions or countries with differing percentages of DCOs. We also present examples of some hazard ratios found in practice.

1. Introduction

Cancer Registries across the United Kingdom (UK) are meant to record the details of all patients diagnosed with cancer in their respective areas. Most cancer registrations are made by the cancer Registry while the patient is still alive (Registered In Life, or RIL, cases) but some registrations are only initiated after the patient has died – these are termed Death Certificate Initiated registrations (DCIs). Theoretically, most of these can later be followed-back through GP or hospital

* *Corresponding author:* Address: Trent Cancer Registry, 5 Old Fulwood Road, Sheffield S10 3TG, England, United Kingdom. Tel.: +44 114 226 3560x63563; fax: +44 114 226 3561.
E-mail address: paul.silcocks@nhs.net (P. Silcocks).
[d] Previously at Trent Cancer Registry.
0959-8049/$ - see front matter
doi:10.1016/j.ejca.2009.06.013

records to obtain the correct original date of diagnosis (followed-back DCIs), and other information about the cancer patients, their tumour and their treatment. True death certificate only registrations (DCOs) occur either when hospital or GP records contain no supporting evidence for cancer, or when no GP or hospital records can be linked to the case at all. However, in many Cancer Registries, the 'reported' percentage of DCOs includes both 'true' DCOs and DCIs for which attempts to follow-back have not (yet) been made, or have not yet been completed.

Since the only evidence that the patient had cancer for a reported DCO case is on the death certificate, the date of diagnosis has to be taken to be the same as the date of death and survival time is therefore zero. These cases are excluded from survival analyses, hence overestimating the true survival, assuming that the unknown survival would be worse on average than cases with a known diagnosis date.

Previously, a DCO percentage of 15% has been called 'large'[1] and worldwide DCO rates of 25% or over are common for a range of sites.[2] Even within the territory covered by a single Registry some centres may perform less well. For instance as on 1st February 2005, in Trent the reported percentage of DCOs for 2003 ranged from 2.6% to 27.7% across the old Health Authority areas, although subsequent follow-back efforts did reduce the range for 2003 to 1.0–7.5% as on 3rd March 2007.

The target for DCO percentage for UK Registries is now 2%, a level actually achieved or bettered by some European and UK Registries, although not all. The impact of DCO registrations has recently been explored by Robinson and colleagues,[3] including the effect of 'lost' cases that never come to the attention of the Registry (a topic we do not address). Their paper adjusted for DCOs by assuming a single survival time for such cases whereas our method makes the possibly more realistic assumption of a single hazard ratio, thus allowing DCOs to have a distribution of survival times.

It is, however, helpful to be able to make a rapid assessment of the bias in survival estimates caused by DCOs, and whether there needs to be a correction applied when making comparisons across Registries with very different % DCO values. In the EUROCARE I report, Berrino and colleagues[4] examined the bias induced by DCOs by comparing survival estimates before and after an intensive trace-back exercise. They observed that the percentage reduction in estimated survival resulting from the inclusion of DCO cases was generally of the same order as the proportion of such cases in the series under study; and as such that this could be used as a quick estimate of the likely true survival. The present paper demonstrates why this should be so, indicates when this simple approximation is likely to be reasonable to apply, and allows appropriate correction to be made when comparisons are being made.

2. Methods

The basic idea is to regard the true survival of all cases as being a mixture of the survival of observed cases (i.e. the pool of RIL and followed-back DCI cases generally used by Registries to report survival estimates) and the necessarily unobservable survival of the DCOs. Secondly, we assume that this unobservable survival of DCOs can be described relative to observed cases by a proportional hazards model with hazard ratio R. With these assumptions the true survival at time t after diagnosis is then given by:

$$S_{\mathrm{TRUE}}(t) = (1-p)S_{\mathrm{OBSERVED}}(t) + pS^{R}_{\mathrm{OBSERVED}} \quad (1)$$

where p is the 'reported' proportion of DCOs (i.e. 'true' DCOs plus DCIs for which no follow-back has been attempted/completed).

If the hazard ratio R can be estimated then the observed survival can be corrected using this formula to give an estimate of the true survival. The problem of course is how to estimate R. We describe how this can be done in Appendix A. We refer to Eq. (1) as the 'exact' correction, for reasons that will become apparent.

It can be shown that there is an upper bound for the difference between the true survival and the observed survival, provided we have an estimate for the hazard ratio. Also, in the long term (i.e. as follow-up increases), and especially for small p, or R much greater than 1, Eq. (1) simplifies to:

$$S_{\mathrm{TRUE}}(t) \approx (1-p)S_{\mathrm{OBSERVED}}(t) \quad (2)$$

We refer to Eq. (2) as the 'approximate' correction, and is most useful when only a point estimate of the observed survival is available.

If we make an additional reasonable assumption that the true survival of DCOs is worse on average than that of cases in general (so that R is greater than 1) then the greater the value of R is, the earlier in the survival curve the maximum bias will be, and in fact for all hazard ratios above 2 the maximum bias occurs progressively earlier than the median observed survival time. Fig. 1 illustrates how the value of the observed survival at the point of maximum difference (or bias) increases with the hazard ratio[5] for DCOs (corresponding to smaller values of t). Furthermore it turns out that even if DCOs have only a marginally worse prognosis than observed cases (i.e. with R being only slightly greater than 1) once observed survival has fallen below 36.7% the true and observed survival curves must converge, Fig. 2. Finally, these

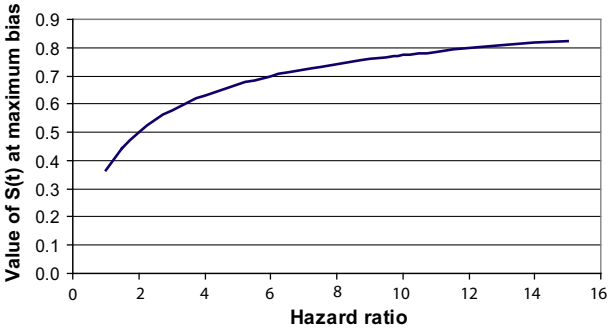

Fig. 1 – This shows that as the true prognosis for DCOs worsens (measured by increasing hazard ratio), the maximum bias occurs at higher values of observed survival $S(t)$ and hence earlier survival times. For values of R above 2, the maximum bias always occurs earlier than the median survival time.

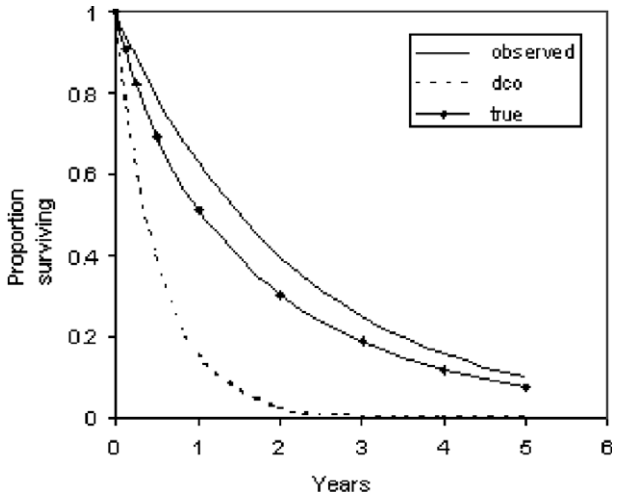

Fig. 2 – For a DCO: Observed hazard ratio (R) of 4, and 25% DCOs the maximum absolute difference (observed minus true) in the figure above is 11 percentage points at about $7\frac{1}{2}$ months, but assuming proportional hazards, for any hazard ratio the maximum possible difference cannot be more than 25 percentage points.

approaches assume that the percentage of DCOs, the incidence of and survival from cancer are in a steady state, and that the survival of DCOs is not better than that of RIL cases.

3. Application

We need to know roughly how big the hazard ratio R is likely to be in practice. This ultimately depends on the existence of intensive trace-back exercises to give estimates of true survival. This was possible at Trent Cancer Registry following implementation of new procedures and intensive effort to improve trace-back. Hazard ratios were estimated from full survival data for individual colorectal, breast and lung cancer patients diagnosed in 2000 in the old Trent Region. Further details are given in Appendix B. If a trace-back exercise only reports point estimates of true and observed survival, as in Berrino and colleagues,[4] it is still possible to rearrange Eq. (1) to obtain an estimate of the hazard ratio, see Appendix A Eq. (A7).

4. Results

Table 1 shows the results of the approximate correction formula of Eq. (2) applied to the point estimate data presented in the EUROCARE I study (which gives 5-year survival estimates before and after an intensive active follow-back). In addition, Table 1 displays the hazard ratios for the followed-back DCIs relative to all observed cases, estimated using Eq. (A7) given given in Appendix A, as applied to the point estimates in the EUROCARE I data; the corresponding values for the proportion of survivors at which the difference between observed and true survival is at a maximum; and also the hazard ratios obtained by Cox regression using Trent Cancer Registry data on year 2000 cases (i.e. after the intensive trace-back).

The estimates of the hazard ratios for three most common tumour sites were very similar despite being obtained for different countries and by different methods. These hazard ratios were about 4 for lung cancer (with poor prognosis), about 6.5 for colorectal cancer (with intermediate prognosis) and about 13.5 for breast (with good prognosis).

5. Discussion

We have shown that a simple proportional hazards model for the survival of DCO cases relative to all RIL cases allows the time at which the biasing effect of DCO registrations on survival is maximised to be estimated. Also, the observation of Berrino and colleagues[4] is simply a consequence of the model as follow-up increases with time, as in Eq. (2). This happens to coincide with the naïve estimate which would result from assuming (incorrectly) that DCOs are a random sample of all cases, but the present analysis shows how this naïve correction only applies in the long term, after the cases destined to become DCOs have died out.

We suggest as a rule of thumb that Eq. (2) should be applied after the time of median survival, on the assumption that a value of R less than 2 is *a priori* unlikely. The approximate correction results in point estimates of 5-year survival very close to those obtained by the intensive follow-back of EUROCARE I, and this is attributable to the fact that the correction is applied after a 'reasonable' lapse of time from diagnosis, which need not be very long. For $R = 6$, as seen for colorectal cancer in Table 1, using Eq. (A6) of Appendix A

Table 1 – Comparison of active follow-back and correction formula.

	Lung	Breast	Colorectal
Proportion of DCOs (=p)	**0.255**	**0.118**	**0.22**
% 5-year Survival excluding DCO (=S)	**6.4**	**60.1**	**33.3**
Adjusted % survival figure in EUROCARE I after follow-back	**4.8**	**54.1**	**26.6**
Result of approximate correction = $(1 - p) * S$ applied to EUROCARE I (%)	4.77	53.01	25.97
Hazard ratio R derived from point estimate of 5-year survival in EUROCARE I, using Eq. (A7) – Appendix A	3.42	13.07	5.99
% Survival when bias is maximal (using R)	60.2	80.8	69.9
Hazard ratio (R) estimate from TCR[a] full survival data (year 2000 cases)	4.36	14.25	6.50

Items in bold obtained from Table 1.2 of EUROCARE I.[4]
a TCR = Trent Cancer Registry.

shows that the approximate correction will be within 2.5% of the 'exact' value once the observed survival has fallen to 59%, even if the proportion of DCOs is as high as 25%. For a value of R as low as 2, for which the maximum bias occurs at 50% survival, the approximate correction will be within 10% of the true value provided that there are no more than 15% of DCOs.

Empirically, from this limited exercise, it also seems that R is inversely related to observed survival, hence the time taken to reach an appropriate time to apply the approximation is much the same for many scenarios, so that 5-year survival would be a reasonable starting point for many tumours except for ones with particularly good survival. Hence, the naive approximate correction should not be used for correcting short-term survival estimates, e.g. 1-year survival, which is currently suggested as a proxy for measuring the effect of late presentation across the countries included in the EUROCARE I, or similar later studies. Given a value for R, however, the 'exact' correction, based on Eq. (1) can be applied at any time point.

While DCOs are clearly not a random sample of all cases, we explored whether it is reasonable to assume that they are a random sample of DCIs. While it is unlikely to be true, the difficulty is that by their very nature we know little about DCOs. Appendix B describes work we performed with traced-back DCI cases to see if there was a relation between the time from death to registration and survival time (so regarding DCOs as traced-back DCIs with an 'infinite' duration from diagnosis to registration). No relation was found and so in the absence of any other evidence to the contrary, we believe that the assumption is justifiable on the parsimony principle of Occam's razor.[6]

The existence of DCOs may also be taken as evidence that possible sources of information have not been covered; conceivably long-term survivors may therefore have been missed out. However, if they exist such cases would only affect the later part of the survival curve and would have a small effect unless both the proportion of DCOs and the proportion of long-term survivors were high. In practice it is unlikely that there would be a high proportion of long-term survivors amongst DCOs. This is partly because of selection bias (in that these patients have died) and partly because DCOs are assumed effectively to be a random sample of traced-back DCIs – a group that has poorer survival than cases registered in life and for whom the proportion of survival time missed out seems relatively modest.[7]

If no 'plug-in' value for the relevant hazard ratio R based on published data is available, or an informed guess is not sufficiently accurate, R must be estimated by means of an intensive bespoke trace-back exercise. We have shown that estimates of the hazard ratios for three most common tumour sites with poor (lung), intermediate (colorectal) and good (breast) prognosis obtained in different countries and by different methods agree remarkably well. The value of a trace-back exercise lies not only in producing valid survival estimates by reducing the reported percentage of DCOs, but also by helping to estimate the hazard ratio for DCOs relative to observed cases. Once estimates of the hazard ratios are obtained they may be applied cautiously in a 'plug-in' fashion to the corresponding tumour sites in other years or at geographic locations where trace-back is poor. We suggest that the estimates obtained here for the three common sites could be used elsewhere as needed, either for other studies for these three cancer sites, or for other cancer sites with similar prognoses to these three cancer sites.

Even with intensive trace-back there may still be some residual DCOs and the approach outlined here can still be applied to correct the bias due to these cases, although the effect will then be marginal.

In practice, of course the best way to avoid DCOs is through improved routine trace-back. However, our approach can be used to predict what the likely effect on survival of the trace-back will have, and justify the extra work involved. Of course, reducing the proportion of DCO cases will improve data quality in other areas besides overall survival (e.g. increasing the information on disease extent, morphology and treatment) so the existence of this simple correction should not be perceived as an excuse for not performing trace-back of DCI cases. Alternatively, our method can be used to correct survival results from other sources, including historical data, especially when these sources contain high percentages of DCOs. As this correction is applied to the observed survival, it can be used for adjusting both observed and relative survival results. It may also be of interest to compare the results of our approach to correcting for DCOs with those of Robinson and colleagues.[3]

In short, when comparing survival across regions or countries, researchers should always consider adjusting for the % DCOs. We have shown why the simple 'approximate' correction of Berrino and colleagues[4] works, and suggest as a rule of thumb that it is probably acceptable in most cases for use on 5-year estimates of survival although not earlier than the time of median survival, nor for correcting short-term survival estimates, e.g. 1-year survival. On the other hand, we also suggest that a more generally applicable 'exact' alternative applicable at any time point provided a plausible value of R is available. We have reported estimates of these hazard ratios for the three most common cancer sites, with differing prognoses which can be applied by others elsewhere.

Conflict of interest statement

None declared.

Acknowledgements

Funding – no external funding.

Appendix A. Theory

We assumed that the true survival is a mixture of the survival of cases who will become DCOs and cases whose survival time will be known. We also assumed that the hazard for the DCO group is proportional to that of the observed group, with hazard ratio R and that the proportion of DCO cases is p.

The true survival at time t after diagnosis can then be written as:

$$S_{\mathrm{TRUE}}(t) = (1-p)S_{\mathrm{OBS}}(t) + pS_{\mathrm{DCO}}(t) \quad \text{(A1)}$$

Which, by the proportional hazards assumption:

$$= (1-p)S_{OBS}(t) + pS_{OBS}^{R}(t) \quad \text{(A2)}$$

If R can be estimated then the observed survival can be corrected using this formula to give an estimate of the true survival.

A.1. *Consequences of this model*

1) The difference between the observed and true survival is:

$$S_{OBS}(t) - S_{TRUE}(t) = S_{OBS}(t) - [(1-p)S_{OBS}(t) + pS_{OBS}^{R}(t)] \quad \text{(A3)}$$

In the context of comparing survival distributions for equivalence, Wellek[5] showed that under a proportional hazards model the maximum difference between two survival curves is related to the hazard ratio.

If we write the difference between observed and true survival as Δ, and differentiate with respect to $S_{OBS}(t)$

$$\frac{\partial \Delta}{\partial S_{OBS}} = p - pRS_{OBS}^{R-1} \quad \text{(A4)}$$

Which on equating to zero and rearranging gives $S_{OBS}(t_{max}) = \left(\frac{1}{R}\right)^{\frac{1}{R-1}}$ where t_{max} is the time at which this maximum difference occurs, and the maximum difference itself is given by:

$$\Delta_{max} = p\left[(1/R)^{1/(R-1)} - (1/R)^{R/(R-1)}\right] \quad \text{(A5)}$$

Use of the maximum difference is preferable to an arbitrary choice of the difference at 5 or 10 years follow-up, which may be less than the actual maximum.

2) As $t \to \infty$, that is, for long term survival it is also clear that $S_{TRUE}(t) \to (1-p)S_{OBS}(t)$ more rapidly if p is small or if R is large. This is because as R is greater than 1, $S_{OBS}^{R}(t)$ tends to zero faster than $S_{OBS}(t)$. Of course if R is close to 1 then there is little bias anyway.

3) Also it can be shown (by L'Hospital's rule and application of limit theorems) that for cases in which DCOs are in fact very much sicker than observed cases, ie, as $R \to \infty$, $\Delta_{max} \to p = \Delta^{*}_{max}$, the *maximum possible value* of Δ_{max} over all values of R, for a proportional hazard model. It provides a worst-case bound, if we are unwilling to make assumptions about R and have no data. This is because the greater the value of R, the closer this maximum difference occurs to t = 0, so that in the limit all the DCO deaths occur immediately, and the true survival curve is $(1-p)S_{OBS}(t)$ for all t.

4) If the approximate corrected survival of Eq. (2) in the main paper is expressed as a fraction of the "exact" corrected survival of Eq. (1) then its relative accuracy at the maximum discrepancy is:

$$\%\ \text{accuracy} = \frac{1}{1 + \frac{p}{(1-p)}\frac{1}{R}} \times 100 \quad \text{(A6)}$$

These approaches assume that the percentage of DCOs, incidence and survival from the cancer are in a steady state, and that survival of DCOs is not better than that of RIL cases.

5) If a traceback exercise only reports point estimates of true and observed survival, as in Berrino et al.,[4] it is possible to rearrange Eq. (1) from the main paper to obtain the following estimate of the hazard ratio:

$$R = \frac{\log_e\left[\frac{\widehat{S_{TRUE}(t)} - (1-p)S_{OBSERVED}(t)}{pS_{OBSERVED}(t)}\right]}{\log_e[S_{OBSERVED}(t)]} + 1 \quad \text{(A7)}$$

This may be used to compare trace-back exercises reported in different ways.

Appendix B. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ejca.2009.06.013.

REFERENCES

1. Jensen OM, Parkin DM, MacLennan R, Muir CS, Skeet RG, editors. *Cancer registration principles and methods*. Lyon: IARC Scientific Publications No. 95; 1991.
2. Parkin DM, Whelan SL, Ferlay J, Teppo L, Thomas DB, editors. *Cancer incidence in five continents*, vol. VIII. Lyon: IARC Scientific Publications No. 155; 2002.
3. Robinson D, Sankila R, Hakulinen T, Møller H. Interpreting international comparisons of cancer survival: the effects of incomplete registration and the presence of death certificate only cases on survival estimates. *Eur J Cancer* 2007;**43**:909–13.
4. Berrino F, Esteve J, Coleman MP. Basic issues in estimating and comparing the survival of cancer patients. In: Berrino F, Sant M, Verdecchia A, Capocaccia R, Hakulinen T, Esteve J, editors. *Survival of cancer patients in Europe. The EUROCARE study*. Lyon: IARC Scientific Publications No. 132; 1995. p. 1–14.
5. Wellek S. A log-rank test for equivalence of two survivor functions. *Biometrics* 1993;**49**:877–81.
6. Last JM, Spasoff RA, Harris S. *A dictionary of epidemiology*. 4th ed. New York: Oxford University Press; 2000.
7. Silcocks P. Survival of death certificate initiated registrations: selection bias, incomplete trace-back or higher mortality? *Br J Cancer* 2006;**95**:1576–8.