

2017-12-01

# Reconstructing time-to-event data from published KaplanMeier curves

Wei, Yinghui

<http://hdl.handle.net/10026.1/11063>

---

10.1177/1536867X1801700402

Stata Journal

StataCorp

---

*All content in PEARL is protected by copyright law. Author manuscripts are made available in accordance with publisher policies. Please cite only the published version using the details provided on the item record or document. In the absence of an open licence (e.g. Creative Commons), permissions for further reuse of content should be sought from the publisher or author.*

# Reconstructing time-to-event data from published Kaplan-Meier curves

Yinghui Wei

Centre for Mathematical Sciences  
School of Computing and Mathematics  
Plymouth University, UK  
yinghui.wei@plymouth.ac.uk

Patrick Royston

MRC Clinical Trials Unit at UCL  
London, UK  
j.royston@ucl.ac.uk

**Abstract.** Hazard ratios can be approximated by data extracted from published Kaplan-Meier curves. Recently, this curve approach has been extended beyond hazard ratio approximation with the capability of constructing time-to-event data at the individual level. In this article, we introduce a Stata command `ipdfc` to implement the reconstruction method to convert Kaplan-Meier curves to time-to-event data. Examples are given to illustrate how to use the command.

**Keywords:** time-to-event data, Kaplan-Meier curves, hazard ratios.

## 1 Introduction

The hazard ratio is often recommended as an appropriate effect measure in the analysis of randomized controlled trials with time-to-event outcomes (Parmar et al. (1989), Deeks et al. (2008)) and has become the *de facto* standard approach to analysis. In meta-analysis of aggregated time-to-event data across trials, an essential step is to extract the (log) hazard ratio and its variance from published trial reports. Various extraction methods have been described (Parmar et al. (1989), Williamson et al. (2002) and Tierney et al. (2007)), including direct and indirect estimates of hazard ratios based on 95% confidence intervals, *P*-values for the logrank test or the Mantel-Haenszel test, and regression coefficients in the Cox proportional hazards model. An approximation to hazard ratios can also be derived by a “curve approach”, as described by Parmar et al. (1989) and Tierney et al. (2007). The curve approach makes use the extracted ordinate (*y*-) and abscissa (*x*-) values from the Kaplan-Meier curve, to calculate hazard ratios for each time interval for which the number of patients at risk was reported. The overall hazard ratio during the follow-up phase is then derived by a weighted sum of the individual estimates of hazard ratios across time intervals, with the weights inversely proportional to the variance of each estimate (Parmar et al. (1989)).

The curve approach has been extended (Guyot et al. (2012)) beyond the estimation of hazard ratios to the reconstruction of time-to-event data at the individual level. Availability of reconstructed individual-level data allows one to fit alternative models if desired in secondary analyses. Because non-proportional hazards are increasingly reported in trials, alternative measures, such as restricted mean survival time, that do not require the proportional hazards assumption, may have a more intuitive interpretation under non-proportional hazards (Royston and Parmar (2011)). As the proportional hazards assumption may not be satisfied for all trials in a meta-analysis, alternative effect measures to hazard ratios may be more appropriate in such settings (Wei et al. (2015)). However, by definition, newly developed effect measures are not reported in earlier trial publications. The use of these measures therefore relies either on collaborative sharing of individual-level data or on methods that enable reconstruction of such data from trial reports.

The reconstruction algorithm was written as an R function (Guyot et al. (2012)). In this article, we present an implementation of this algorithm with improvements by introducing a Stata command `ipdfc`, which has the following features:

- Makes use of the curve approach to reconstructing individual-level time-to-event data based on the published Kaplan-Meier curves;
- Utilizes the number of patients at risk, as reported in the trial publication;
- Can identify which extracted time points correspond to the lower and upper end-point of each time interval in the risk table;
- Can use survival/failure probabilities/percentages as data input;
- Incorporates correction of monotonicity violators in the extracted data for survival/failure probabilities/percentages.

In the following section, we briefly overview the methods underpinning the `ipdfc` command introduced in this article. This is followed by detailed descriptions of syntax, options and saved results. We then demonstrate its application to two examples from trial publications, and assess the approximation accuracy by comparing summary statistics between the reconstructed data and the original publications. We close with a discussion.

## 2 Methods

### 2.1 Extracting data from published Kaplan-Meier curves

The reconstruction of time-to-event observations is based on data extracted from published Kaplan-Meier curves. In such curves, the x-values usually represent the follow-up time since randomization; the y-values may represent the survival probabilities at the corresponding time points. As an alternative to survival probabilities, the y-values may be survival percentages, failure probabilities or percentages, as specified in the trial publication. All these measures can be transformed arithmetically into survival probabilities. In addition to data from curves, the number of patients randomized into each arm of a trial should be extracted from publications.

The DigitizeIt (<http://www.digitizeit.de/>) software application is a suitable tool for extracting data from a graphical image. Data extraction using this software is far more rapid, detailed, accurate and reliable than by applying pencil and ruler methods manually to a reduced image of the graph. If a curve is displayed as a clearly defined, unbroken line, DigitizeIt can automatically read off the x- and y-values at a large number of time points. This helps to ensure the good quality of data required as input in the reconstruction of time-to-event observations. However, if the curve is presented as a broken (e.g. dashed) line, the operator must extract data semi-manually by clicking on individual points on the curve using a mouse. Since each click returns only one data point, many clicks must be made to obtain sufficient data when there are many jumps in the curve. In contrast, within a specific time interval where there are few events or where the survival curves are flat, little information is available and correspondingly few clicks are required.

In addition, it is important to extract the number of patients at risk for each arm at a regular time intervals during the follow-up. This information, usually known as the risk table, is often presented beneath the published Kaplan-Meier curves. The accuracy of the approximated time-to-event data can be improved by incorporating information provided in the risk table (Tierney et al. (2007)).

### 2.2 Adjusting monotonicity violators

Since a survival function is by definition monotone decreasing with time, the y-values extracted from a survival curve should also be monotone when ordered by the corresponding x-values. However, there may be violators among the extracted data such that the monotonicity constraint is not satisfied. This is due

to publication quality of the curves or to errors in controlling the mouse clicks (Guyot et al. (2012)). It should be noted that the reconstruction algorithm involves estimating survival functions. Monotonicity violators can lead to incorrect estimates for the number of events, and subsequently incorrect estimates of the survival function, which prevents the reconstruction from working. It is therefore crucial to correct the values for violators to ensure monotonicity. Since violators are often multiple, a systematic method is required.

In the `ipdfc` command, we incorporate alternative methods for the correction of violators. The first method, isotonic regression (Barlow et al. (1972)), may help to detect violators and correct their values by using a Pool-Adjacent-Violators (PAV) algorithm. Adjacent violators occur where a pair of adjacent times and corresponding survival probabilities is inappropriately ordered, e.g. time = (1.0,1.1), survival = (0.91, 0.92). Briefly, the PAV algorithm replaces the adjacent violators with their mean so that the data satisfy the monotonicity constraint. The technique has been recently coded in a Stata command `irax` (van Putten and Royston (2014)), which can be called in our command. We also consider an alternative. We replace the value of a violator by the value of its adjacent violator, such that the corrected data satisfy the monotonicity. We expect the use of either method will lead to similar results because the absolute difference between the values of adjacent violators is often too small to have a material influence on the resulting data.

### 2.3 Algorithm to reconstruct survival data

We now briefly describe the algorithm underpinning the command `ipdfc`. We start with introducing notations. Let  $S_k$  denote survival probabilities at time  $t_k$ , where ( $k = 1, 2, \dots, K$ ) and  $K$  is the total number of data points extracted. The survival probabilities  $S_k$  and the corresponding time  $t_k$  may be extracted from respective y- and x-coordinates of a Kaplan-Meier curve. Let  $nrisk_i$  denote the number of patients at risk at time  $trisk_i$ , where  $i = 1, \dots, T$ , with  $T$  as the number of intervals where the number of patients at risk is reported. The number of extracted data points  $K$  is often much greater than  $T$ , the total number of intervals at the risk table. If the risk table is not reported, we have  $T = 1$ .

The four quantities  $S_k, t_k, nrisk_i$  and  $trisk_i$  are the required input in the algorithm. As mentioned above, the number of patients at risk, if available, should be included in the algorithm. Otherwise if  $T = 1$ , the number of patients randomized to each arm should be included to the algorithm. The total number of events  $D$  can also be used in the reconstruction.

In the algorithm, we will estimate the following quantities: the number of censoring  $\hat{c}_k$ , the number of events  $\hat{d}_k$ , the censoring time  $\widehat{ctime}_k$  and event time  $\widehat{dtime}_k$ . To estimate these quantities, we implement the algorithm described in Guyot et al. (2012) by adding three new components for improvements. First, we calculate  $lower_i$  and  $upper_i$  by using the input data  $t_k, trisk_i$ . Here  $lower_i$  and  $upper_i$  are respectively the indices for the first and last time points extracted from the time interval  $[trisk_i, trisk_{i+1}]$ . For each of such time interval,  $lower_i$  is equal to  $\min\{k : t_k \geq trisk_i\}$ , and  $upper_i$  is equal to  $\max\{k : t_k \leq trisk_{i+1}\}$ . Thus,  $lower_i$  and  $upper_i$  are not required as data input as in the R code of Guyot et al. (2012). Second, we adjust the values of monotonicity violators by using isotonic regression or its alternative as just described. Third, we extend the algorithm to the situation when the number at risk is reported at the last time interval, at which we allow the calculation of the number of censoring following the same methods as those for the other intervals. The full algorithm is given in the appendix of this article.

### 3 The ipdfc command

#### 3.1 Syntax

```
ipdfc, surv(varname) tstar(varname) trisk(varname) nrisk(varname)
generate(varname1 varname2) saving(filename[, replace]) [ probability
failure isotonic totevents(#) ]
```

This syntax converts data extracted from a Kaplan-Meier curve to time-to-event data. It should be noted that the syntax does not handle more than one sample at once. When dealing with a trial having more than one arm, the syntax converts data extracted from one curve at a time to time-to-event data for the respective arm. This should be done for all arms one by one, and further data management is needed to amalgamate the data from all arms of a trial, if the data are from a trial, as seen in the examples given later in this article.

#### 3.2 Required input

`surv(varname)` specifies the data extracted from the ordinate (y-axis) of a published Kaplan-Meier curve. It can be survival probabilities, survival percentages, failure probabilities or failure percentages. By default, `varname` is assumed to contain survival percentages.

`tstart(varname)` specifies the time since randomization as extracted from the abscissa (x-axis) of a published Kaplan-Meier curve. The time could be in any units (e.g. days, months, years) as specified in the publication.

`trisk(varname)` specifies the times at which the numbers of patients at risk in `trisk()`. Set `varname = 0` if only the total number of patients in the sample is known.

`nrisk(varname)` supplies the number of patients at risk for each time in `trisk()`. Both `trisk()` and `nrisk()` may often be found in a risk table displayed beneath published Kaplan-Meier curves. If no risk table is available, specify `nrisk()` as the number of patients in the sample and `trisk(varname)` as 0.

`generate(varname1 varname2)` generates the time-to-event outputs extracted from the input information. `varname1 varname2` specify two new variables, time to event and an event indicator (1 = event, 0 = censored). For example, `generate(time event)` would create `time` as the time to event and `event` as the event indicator.

`saving(filename[, replace])` saves the reconstructed survival data to a file named `filename.dta`.

### 3.3 Options

`probability` signifies that `varname` in `surv(varname)` contains probabilities rather than percentages.

`failure` signifies that `varname` in `surv(varname)` contains failure information rather than the default survival information.

`isotonic` invokes use of isotonic regression to adjust values which violate the time-related monotonicity in `surv(varname)`. By default, an alternative, simpler method is used to correct the values of violators, by replacing the value of a violator by the value of its adjacent violator.

`totevents(#)` is the total number of events and is used to adjust the number of observations censored in the final interval of the risk table.

## 4 Illustrative Examples

### 4.1 Example 1: Head and neck cancer trial

Our first example is a two-arm randomized controlled trial published in Bonner et al. (2006). A total of 424 participants with locoregionally advanced head and neck cancer were randomized to receive either radiotherapy plus cetuximab or radiotherapy alone. The primary outcome was the duration of locoregional control. Both Kaplan-Meier curves and the hazard ratio were reported. This example was first used in Guyot et al. (2012) to illustrate the application of the reconstruction method. Here we use `ipdfc` to reconstruct the survival data, to illustrate how it performs compared with Guyot et al. (2012) and with the results in the original publication. We run the steps for each arm separately to obtain arm-specific data based on the associated Kaplan-Meier curve from the trial report.

We read in a text file for the control arm by calling `import delimited`.

```
. import delimited using "Head_and_neck_arm0.txt", clear
```

The text file contains four variables: `ts` and `s` as the data extracted from the x-axis and y-axis of a curve; and `trisk` and `nrisk` from the risk table.

We regenerate data for the control group by calling `ipdfc`.

```
. ipdfc, surv(s) tstart(ts) trisk(trisk) nrisk(nrisk) isotonic ///  
> generate(t_ipd event_ipd) saving(temp0)
```

Since the extracted y-values are survival percentages in this example, we need not call either `probability` or `failure` to convert `s`. However, we call the option `isotonic` to evoke isotonic regression to correct monotonicity violators. The regenerated survival data are stored in the file `temp0.dta`.

We run the following steps for the treatment group.

```
. import delimited using "Head_and_neck_arm1.txt", clear  
  
. ipdfc, surv(s) tstart(ts) trisk(trisk) nrisk(nrisk) isotonic ///  
> generate(t_ipd event_ipd) saving(temp1)
```

The regenerated survival data are stored in the file `temp1.dta`.

The data simulated from both arms are then combined and specified with an arm indicator.

```
. use temp0, clear  
. gen byte arm = 0  
. append using temp1
```

```
. replace arm = 1 if missing(arm)
```

In the amalgamated data, there are three variables `t_ipd`, `event_ipd` and `arm`, which are time to an event, event indicator and arm indicator, respectively. We label the arm indicator as “Radiotherapy” and “Radiotherapy plus cetuximab”, as specified in the trial publication.

```
. label define ARM 0 "Radiotherapy" 1 "Radiotherapy plus cetuximab"
. label values arm ARM
```

We set `time` as the time to failure.

```
. stset t_ipd, failure(event_ipd)
```

By calling the `sts graph`, we reconstruct the survival curves.

```
. sts graph, by(arm) title("") xlabel(0(10)70) ylabel(0(0.2)1) ///
> risktable(0(10)50, order(2 "Radiotherapy" 1 "Radiotherapy plus" bottom(msize(5)))) ///
> xtitle("Months") l2title("Locoregional Control") ///
> scheme(s2color) graphregion(fcolor(white)) ///
> plot1opts(lpatter(solid) lcolor(gs12)) ///
> plot2opts(lpatter(solid) lcolor(black)) ///
> text(-0.38 -9.4 "Cetuximab") ///
> legend(off) ///
> text (0.52 53 "Radiotherapy plus Cetuximab") text(0.20 60 "Radiotherapy")
```

The survival analysis is carried out by calling `stcox arm`.

```
. stcox arm
      failure _d: event_ipd
      analysis time _t: t_ipd

Cox regression -- Breslow method for ties
No. of subjects =          424                Number of obs =          424
No. of failures =           241
Time at risk   =  8412.821526
Log likelihood = -1320.1899                LR chi2(1) =           6.31
                                                Prob > chi2 =          0.0120
```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
arm	.7208993	.0947487	-2.49	0.013	.5571859 .9327152

The reconstructed Kaplan-Meier curves (see Figure 1) look similar to the published curves (Bonner et al. (2006)). There is only a small discrepancy in the number of patients at risk in the radiotherapy plus Cetuximab arm. For this arm, based on the reconstructed data the number of patients at risk are 211, 130, 94, 67, 35 and 10, which are similar though not identical to 211, 143, 101, 66, 35 and 9 in the original publication. The discrepancy in the risk table

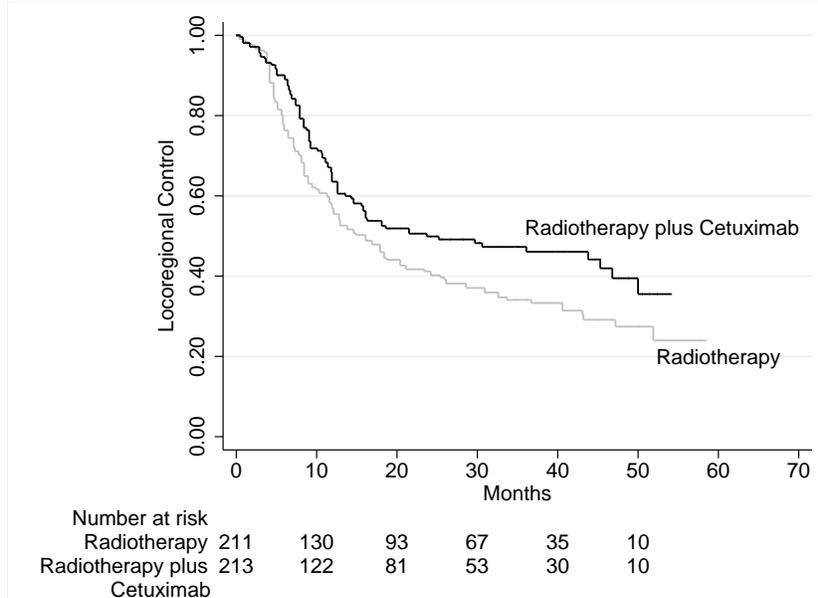


Figure 1: Reconstructed Kaplan-Meier curves for locoregional control among patients with head and neck cancer (Bonner et al. (2006)). Patients are randomized to receive radiotherapy plus cetuximab or radiotherapy alone.

between the approximation (see Figure 1) and the original publication is very small for the radiotherapy arm.

In Table 1, we report percentages of patients surviving 1, 2 and 3 years, median duration of locoregional control and hazard ratio estimates. The estimates of percentages of surviving and median time to event are close to those in the original publication. The hazard ratio (0.72, 95%CI: 0.56-0.93) estimated by our command is close to 0.73(95%: 0.57-0.94) estimated by Guyot et al. (2012). Since we digitize the data independently of Guyot et al. (2012), we do not expect to obtain identical data nor identical results. However, both approximated hazard ratios are similar to published hazard ratio (0.68, 95 %CI: 0.52-0.89), but not identical.

## 4.2 Example 2: ICON7 trial

Our second example is ICON7, a two-arm randomized controlled trial in advanced ovarian cancer (Perren et al. (2011)). A total of 1528 women were randomized to receive either standard chemotherapy plus bevacizumab or standard

Table 1: Example 1. Comparison of summary measures estimated from publication and their corresponding reconstructed data

	Original publication	Guyot (2012)	ipdfc
<b>Radiotherapy arm</b>			
Percent surviving 1 year	55	56.1 (49.6,63.3)	56.9 (49.9, 63.2)
Percent surviving 2 year	41	41.1 (34.7,48.6)	40.9 (34.2, 47.5)
Percent surviving 3 year	34	34.7 (28.4,42.5)	33.5 (27.1, 40.1)
Median duration	14.9	14.9 (11.9,23.0)	16.1 (11.9, 20.4)
<b>Radiotherapy plus cetuximab arm</b>			
Survival rate at 1 year	63	64.0 (57.8,70.9)	65.4 (58.2, 71.6)
Survival rate at 2 years	50	50.4 (43.9,57.8)	51.0 (43.3, 58.6)
Survival rate at 3 years	47	46.7 (40.1,54.4)	49.6 (40.4, 55.7)
Median duration	24.4	24.3 (15.7,45.7)	23.7 (15.6, 46.8)
<b>Hazard ratio with 95%CI</b>			
	0.68 (0.52,0.89)	0.73(0.57,0.94)	0.72 (0.56,0.93)

chemotherapy alone. From the analysis based on data with 30 months follow-up, Perren et al. (2011) concluded that bevacizumab improved progression-free survival (PFS) in this population, with HR 0.81 (95%CI, 0.70 to 0.94;  $P = 0.004$  from a logrank test). Perren et al. (2011) found significant non-proportional hazards ( $P < 0.001$ ) of the treatment effect. Kaplan-Meier curves and the associated risk table for PFS were reported in their Figure 2a, based on which we reconstructed the survival data using `ipdfc`. Also, we make use of the total number of events `tot` because they were available.

```

. local tot0 = 464
. local tot1 = 470
. import delimited using "ICON7_data_arm0.txt", clear
. ipdfc, surv(s) tstart(ts) trisk(trisk) nrisk(nrisk) probability isotonic ///
> tot(`tot0`) generate(t_ipd event_ipd) saving(temp0)

. import delimited using "ICON7_data_arm1.txt", clear
. ipdfc, surv(s) tstart(ts) trisk(trisk) nrisk(nrisk) probability isotonic ///
> tot(`tot1`) generate(t_ipd event_ipd) saving(temp1)

```

In this example, the extracted y-values are survival probabilities. According in the above codes we call the option `probability` to specify that `surv(s)` represents survival probabilities other than survival percentages.

The reconstructed Kaplan-Meier curves in Figure 2 look similar to those in the original publication (Perren et al. (2011)). The number of patients at risk is also well approximated, with most numbers identical to those in the

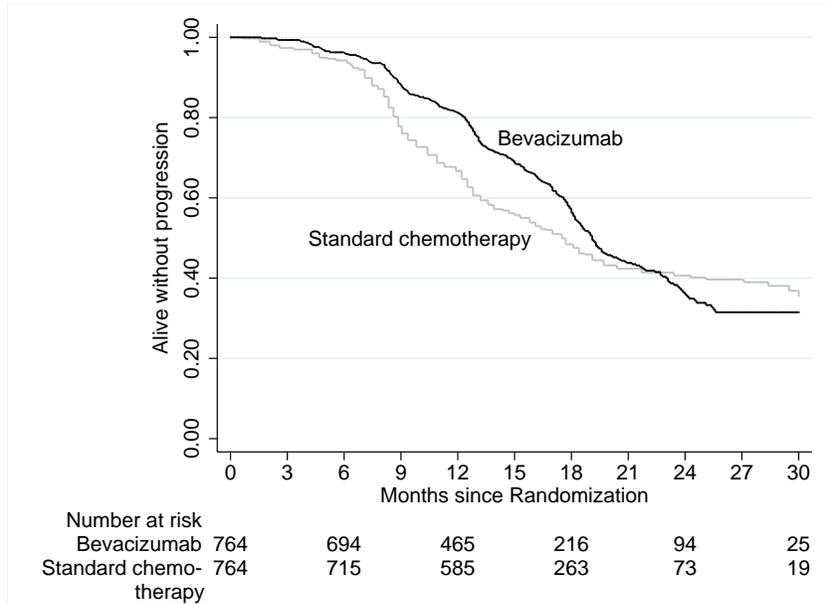


Figure 2: Reconstructed Kaplan-Meier curves for progression-free survival according to treatment group in ICON7 (Perren et al. (2011)). Patients are randomized to receive standard chemotherapy plus Bavacizumab or standard chemotherapy alone.

original publication. The little discrepancies lie in 6 months and 12 months. The numbers of patients at risk are 694 at 6 months and 465 at 12 months based on the approximated data, which compared similarly though not identical to 693 at 6 months and 464 at 12 months as reported in the original publication. The estimated hazard ratios, median survival time as well as the  $P$ -values from the log-rank test are also similar to those in the original publication (see table 2).

### 4.3 Example 3: EUROPA trial

Our third example EUROPA is a two-arm randomized placebo-controlled trial evaluating the efficacy of perindopril in reduction of cardiovascular events among patients with stable coronary artery disease (Fox et al. (2003)). In this trial, 12218 patients were randomly assigned perindopril 8 mg once daily ( $n=6110$ ), or placebo ( $n=6108$ ). Kaplan-Meier curves and the associated risk table were presented in Figure 2 of their trial report. In Fox et al. (2003) the Cox proportional hazards model was used, but the hazard ratio estimate was not reported. It was

Table 2: Example 2. Comparison of summary measures estimated from publication and their corresponding reconstructed data

	Original publication	Reconstructed data
<b>Log-rank test</b>	P=0.004	P=0.009
<b>Non-PH test</b>	P<0.001	P<0.001
<b>Hazard ratio</b>	0.81 (95%: 0.70 - 0.94)	0.83 (95%: 0.72 - 0.96)
<b>Median survival time</b>		
Chemotherapy arm	17.3	17.5 (95%CI: 16.1 - 18.7)
Bevacizumab arm	19.0	19.1 (95%CI: 18.3 - 19.9)

reported in Fox et al. (2003) that perindopril treatment was associated with a significant reduction in the composite events (cardiovascular mortality, non-fatal myocardial infarction, and resuscitated cardiac arrest), with  $P$ -value = 0.0003 from a log-rank test and absolute risk reduction 1.9%.

We extracted the failure percentages and the associated time points respectively from y-axis and x-axis of the Kaplan-Meier curves in their Figure 2. In the following codes, we evoke the option `failure` in the command `ipdfc` to specify that the input data are failure percentages instead of the default, survival percentages.

```
. import delimited using "europa_data_arm0.txt", clear
. ipdfc, surv(s) tstart(ts) trisk(trisk) nrisk(nrisk) failure isotonic ///
>      generate(t_ipd event_ipd) saving(temp0)

. import delimited using "europa_data_arm1.txt", clear
. ipdfc, surv(s) tstart(ts) trisk(trisk) nrisk(nrisk) failure isotonic ///
>      generate(t_ipd event_ipd) saving(temp1)
```

The Kaplan-Meier curves from the reconstructed data are presented in Figure 3. It is clear that the reconstructed curves appear nearly identical to the original. The reconstructed curves correctly reflect that the benefit of perindopril treatment began to appear at 1 year and gradually increased throughout the follow-up of the trial. The numbers of patients at risk are also very similar to the reported values, with only a small discrepancy in the placebo arm in 2 years of follow-up (5781 in the original report vs 5783 as the reconstructed data).

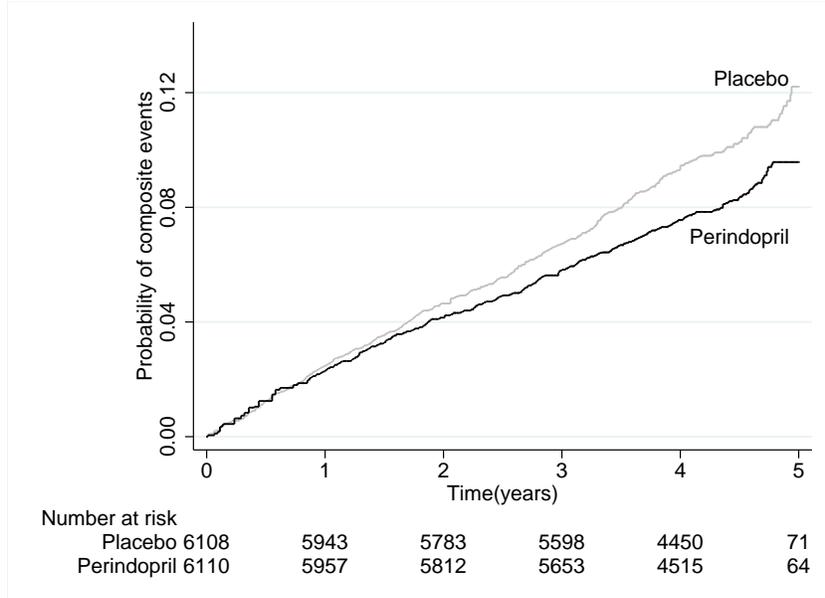


Figure 3: Reconstructed Kaplan-Meier curves for time to first occurrence of event. Patients are randomly assigned to perindopril treatment or placebo, EUROPA trial (Fox et al. (2003)).

Table 3: Example 3. Comparison of summary measures estimated from publication and their corresponding reconstructed data

	Publication	ipdfc with isotonic	ipdfc without isotonic
Log-rank test	P=0.0003	P=0.0006	P=0.0006
Absolute Risk Reduction (95%CI)	1.9%	1.82% (0.80% - 2.84%)	1.80% (0.80% - 2.82%)
Hazard ratio (95%CI)	NA	0.81 (0.72 - 0.91)	0.81 (0.72 - 0.92)

Using the reconstructed data, we obtain  $P$ -value = 0.0006 from a log-rank test (Table 3). Similar to the original report, this result also suggests that perindopril treatment was associated with a significant reduction in the composite events. We estimate the absolute risk reduction as 1.82%, looking similarly to 1.9% in the original publication. We are able to obtain the 95% confidence interval (0.80% to 2.84%) for this based on the reconstructed data. Using Cox proportional

hazards model, we obtain the hazard ratio estimate 0.81 (95%CI : 0.72 – 0.91). Table 3 shows that the correction of monotonicity violators by isotonic regression and by the default method lead to very similar results.

The availability of Kaplan-Meier curves has enabled us to reconstruct the time-to-event data and calculate the hazard ratio, which was not reported for this trial. This would be particularly helpful if this trial was included in a meta-analysis where the hazard ratio is used as an effect measure.

## 5 Discussion

In this article, we provide a Stata command `ipdfc` to implement the algorithm of reconstructing time-to-event data based on the information extracted from published Kaplan-Meier curves. Our command has greater flexibility, incorporating several additional features. It requires fewer inputs, automatically corrects data inconsistency which violates monotonicity, and allows one to utilize the number of patients at risk at the final interval, if reported.

Example 1 shows that the estimates of summary statistics (Table 1) based on `ipdfc` are similar to those by Guyot et al. (2012). Some estimates are better approximations than others. The approximations to median times to event are very close to those in the original publication (Perren et al. (2011)). The approximated hazard ratio is close to but not identical to that reported in the original publication. This small discrepancy is possibly due to the numbers and positioning of events not being entirely accurately estimated by the algorithm.

Although non-proportional hazards are evident in ICON7, the reconstructed Kaplan-Meiers curves and hazard ratio estimate are in reasonable agreement with those from the trial publication (see table 2). This suggests that non-proportional hazards may not much affect the approximation accuracy. However, further empirical evaluation of `ipdfc` in a larger number of trials, with or without obvious non-proportional hazards, is desirable. This is a topic for further research.

Where hazard ratios are not reported but Kaplan-Meier curves are available, the command `ipdfc` is particularly helpful as it enables the reconstruction of time-to-event data and hence allows for re-analysis of the data. For the EUROPA trial, we are able to obtain the estimate of hazard ratio as well as the 95% confidence interval for the absolute risk reduction, both of which were not reported in the trial publication. It is shown that the recovered Kaplan-Meier curves and the associated risk table are both very similar to the originals. This

is perhaps due to the large sample size in this trial and the accuracy of the reconstructed data increases accordingly.

We conclude that `ipdfc` appears to perform quite well in regenerating survival data, sufficient to produce reasonable approximations to summary statistics in time-to-event analysis.

## 6 Acknowledgements

PR was supported by the UK Medical Research Council (MRC) grant to the MRC Clinical Trials Unit Hub for Trials Methodology Research [Grant number MSA7355QP21].

## 7 References

- Barlow, R.E., Bartholomew, D.J., Bremner, J.M. (1972). Statistical inference under order restrictions; the theory and application of isotonic regression. *New York: Wiley*. ISBN: 0-471-04970-0.
- Fox, K.M., EUROpean trial On reduction of cardiac events with Perindopril in stable coronary Artery diseases investigators (2003) Efficacy of perindopril in reduction of cardiovascular events among patients with stable coronary artery disease: randomised, double-blind, placebo-controlled, multicentre trial (the EUROPA study). *Lancet*, **362**,782-788.
- Bonner, J.A., Harari, P.M., Giralt, J. Azarnia, N., Shin, D.M, Cohen, R.B., Jones, C.U., Sur, R., Raben, D., Jassem, J., Ove, R., Kies, M.S., Baselga, J., Youssoufian, H., Amellal, N., Rowinsky, E.K., Ang, K.K. (2006). Radiotherapy plus Cetuximab for Squamous-Cell Carcinoma of the Head and Nect. *The New England Journal of Medicine*, **354**:567-78.
- Deeks, J.J., Higgins, J.P.T., Altman, D.G. (2008). Analysing data and undertaking meta-analyses. Chapter 9 in *Higgins, J.P.T. and Green, S. (eds) Cochrane Handbook for Systematic Reviews of Interventions* Wiley-Blackwell. ISBN: 978-0-470-69951-5(H/B).
- Guyot, P., Ades, A.E., Ouwers, M. and Welton, N. (2012). Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Medical Research Methodology*, **12**, 9.

- Parmar, M.K.B., Torri, V. and Stewart, L.A. (1989). Extracting summary statistics to perform meta-analysis of the published literature for survival endpoints. *Statistics in Medicine*, **17**, 2815-2834.
- Parner, E., Anderson, P.(2010). Regression analysis of censored data using pseudo-observations. *Stata Journal*, **10**(3), 408-422.
- Perren, T.J., Swart, M., Pfisterer, J. et al (2011) A Phase 3 Trial of Bevacizumab in Ovarian Cancer. *The New England Journal of Medicine*, **365**, 2484-96.
- van Putten, W., Royston, P. Isotonic regression analysis, *in preparation*.
- Royston, P., Parmar, M.K. (2002). 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*, **21**(15), 2175-2197.
- Royston, P., Parmar, M.K. (2011). The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in Medicine* **30**(19), 2409-2441.
- Tierney, J.F., Stewart, L.A, Ghersi D., Burdett, S. and Sydes, M.R. (2007). Practical methods for incorporating summary time-to-event data into meta-analysis. *Trials* **8**, 16.
- Williamson, P.R., Turdus-Smith, C., Hutton, J.L. and Marson, A.G. (2002). Aggregate data meta-analysis with time-to-event outcomes. *Statistics in Medicine*, **21**, 3337-3351.
- Wei, Y., Royston, P., Tierney, J. and Parmar, M.K. (2015). Meta-analysis of time-to-event outcomes using restricted mean survival time: application to individual participant data. *Statistics in Medicine*, **34**(21), 2881-2898.

## About the Authors

Yinghui Wei is a statistician with research interests including statistical methods for medicine and health as well as statistical computing and algorithms. Her current work centres on survival analysis, meta-analysis, hierarchical models and infectious diseases epidemiology.

Patrick Royston is a medical statistician with 40 years of experience, with a strong interest in biostatistical methods and in statistical computing and algorithms. He works largely in methodological issues in the design and analysis of

clinical trials and observational studies. He is currently focusing on alternative outcome measures and tests of treatment effects in trials with a timeo-to-event outcome, on parametric modelling of survival data, and on novel clinical trial designs.

## Appendix

---

**Algorithm** Reconstructing survival data (adapted from Guyot et al(2012))

---

**Require:** The data extracted from published survival curves.

$S_k$ : survival percentages as extracted from y-axis,  $k = 1, 2, \dots, K$ . Here  $K$  is the total number of extracted data points

$t_k$ : time from randomization as extracted from x-axis.

$nrisk_i$ : number of patients at risk at time  $trisk_i$ ,  $i = 1, \dots, T$ . Here  $T$  is the number of intervals where the number of patients at risk is reported.

$trisk_i$ : time reported at the risk table.

**Ensure:**  $S_{k+1} \leq S_k$  for all  $k$  to meet the monotonicity constraint.

Set  $lower_i = \min\{k : t_k \geq trisk_i\}$ , and  $upper_i = \max\{k : t_k \leq trisk_{i+1}\}$ .

**if**  $i < T - 1$  and  $T > 1$  **then**

**Step 1** Calculate  $\widehat{nc}_i$ , the number of censored at time  $[trisk_i, trisk_{i+1}]$  by

$$\widehat{nc}_i = S_{lower_{i+1}} / S_{lower_i} \times nrisk_i - nrisk_{i+1}.$$

**Step 2** Distribute  $\widehat{nc}_i$  censored time evenly within  $[trisk_i, trisk_{i+1}]$ . The censored time is then

$$\widehat{ctime}_c = t_{lower_i} + c \times (t_{lower_{i+1}} - t_{lower_i}) / (\widehat{nc}_i + 1),$$

where  $c = 1, \dots, \widehat{nc}_i$ . We can then calculate the number of censored events  $\widehat{nc}_k$  in extracted intervals  $[t_k, t_{k+1}]$ , which is within  $[trisk_i, trisk_{i+1}]$ .

**Step 3** Calculate the number of events at  $t_k$  as

$$\widehat{nd}_k = \widehat{n}_k \times \left(1 - S_k / \widehat{S}_{last(k)}^{KM}\right).$$

The  $\widehat{n}_k$  is the estimated number at risk at time  $t_k$ ,  $\widehat{S}_{last(k)}^{KM}$  is the estimated survival probability at time  $t_{last(k)}$  with

$$last(k) = \begin{cases} 1, & \text{if } k = 1 \\ k', & \text{otherwise.} \end{cases}$$

Note that  $t_{k'} \leq t_k$ , and  $k'$  is such that the latest event occurs at  $t_{k'}$  and there are no events in  $(t_{k'}, t_k)$ . The estimated number of patients at risk at time  $t_{k+1}$  is then  $\widehat{n}_{k+1} = \widehat{n}_k - \widehat{nd}_k - \widehat{nc}_k$ , where  $k \in [lower_i, upper_i]$ . Thus  $\widehat{nrisk}_{i+1} = \widehat{n}_{upper_i+1}$ .

---

---

**Step 4** Set  $\Delta_t = \widehat{nrisk}_{i+1} - nrisk_{i+1}$ .

**if**  $\Delta_t \neq 0$  **then**

Adjust the estimated number of censored in time interval  $[trisk_i, trisk_{i+1}]$  by setting

$$\widehat{nc}_i = \widehat{nc}_i + (\widehat{nrisk}_{i+1} - nrisk_{i+1}).$$

We then repeat Step 1-4 until  $\widehat{nrisk}_{i+1} = nrisk_{i+1}$ .

**end if**  $\Delta_t = 0$

**Step 5** Repeat Step 1-4 until  $i + 1 = T$ .

**end if**  $i + 1 == T$  or  $T == 1$

**if**  $i == T$  or  $(i == 1$  and  $T == 1)$  **then**

**Step 6** Approximate  $\widehat{nc}_T$  within interval  $[trisk_{T-1}, trisk_T]$  by setting

$$\widehat{nc}_T = \min \left( \frac{t_{upperT} - t_{lowerT}}{t_{upperT-1} - t_{lowerT-1}} \times \sum_{i=1}^{T-1} \widehat{nc}_i; nrisk_T \right).$$

We then run Step 2-3 for the last interval  $[trisk_{T-1}, trisk_T]$ .

**end if**

**if** the total number of events  $D$  is not given **then**

Stop the algorithm.

**end if**

**if** the total number of events  $D$  is given **then**

**Step 7** Compute  $\sum_{k=1}^{upperT-1} \widehat{nd}_k$ .

**if**  $\sum_{k=1}^{upperT-1} \widehat{nd}_k \geq D$  **then**

Stop the algorithm.

**end if**

**if**  $\sum_{k=1}^{upperT-1} \widehat{nd}_k < D$  **then**

**Step 8** Adjust the number of censored  $\widehat{nc}_T$  by setting

$$\widehat{nc}_T = \widehat{nc}_T + \left( \sum_{k=1}^{upperT} \widehat{nd}_k - D \right).$$

Repeat Step 2-3, and Step 7-8 for the last interval.

**end if**

**end if**

---