The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*), *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

Subscriptions are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-782-8272, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

Subscription rates listed below include both a printed and an electronic copy unless otherwise mentioned.

U.S. and Canada		Elsewhere	
Printed & electronic		**Printed & electronic**	
1-year subscription	$124	1-year subscription	$154
2-year subscription	$224	2-year subscription	$284
3-year subscription	$310	3-year subscription	$400
1-year student subscription	$ 89	1-year student subscription	$119
1-year institutional subscription	$375	1-year institutional subscription	$405
2-year institutional subscription	$679	2-year institutional subscription	$739
3-year institutional subscription	$935	3-year institutional subscription	$1,025
Electronic only		**Electronic only**	
1-year subscription	$ 89	1-year subscription	$ 89
2-year subscription	$162	2-year subscription	$162
3-year subscription	$229	3-year subscription	$229
1-year student subscription	$ 62	1-year student subscription	$ 62

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.

The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press. Stata, Stata, Stata Press, Mata, Mata, and NetCourse are registered trademarks of StataCorp LLC.

The Stata Journal (2017)
17, Number 4, pp. 786–802

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

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

Patrick Royston
MRC Clinical Trials Unit
University College London
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 command, `ipdfc`, to implement the reconstruction method to convert Kaplan–Meier curves to time-to-event data. We give examples to illustrate how to use the command.

Keywords: st0498, ipdfc, 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, Torri, and Stewart 1989; Deeks, Higgins, and Altman 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, Torri, and Stewart 1989; Williamson et al. 2002; Tierney et al. 2007), including direct and indirect estimates of hazard ratios based on 95% confidence intervals (CIs), p-values for the log-rank 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, Torri, and Stewart (1989) and Tierney et al. (2007). The curve approach uses 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, Torri, and Stewart 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. The availability of reconstructed individual-level data allows one to fit alternative models in secondary analyses if desired. Because nonproportional 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 nonproportional hazards (Royston and Parmar 2011). Because 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 command, `ipdfc`, that has the following features:

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

In the following section, we briefly overview the methods underpinning the `ipdfc` command introduced in this article. We then give detailed descriptions of syntax and options. We then demonstrate its application in 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, survival percentages, failure probabilities, or failure percentages at the corresponding time points, as specified in the trial publication. 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 manually applying pencil and ruler methods 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 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 (for example, dashed) line, the operator must extract data semi-manually by clicking on individual points on the curve using a mouse. Because 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 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

Because 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 errors in controlling the mouse clicks (Guyot et al. 2012). 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. Because violators are often multiple, a systematic method is required.

With 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 algorithm. Adjacent violators occur where a pair of adjacent times and corresponding survival probabilities is inappropriately ordered, for example, time $= (1.0, 1.1)$, survival $= (0.91, 0.92)$. Briefly, the pool-adjacent-violators algorithm replaces the adjacent violators with their mean so that the data satisfy the monotonicity constraint. The technique has been recently coded in a command called `irax` (van Putten and Royston 2017), which can be called in our command. We also consider an alternative. We replace the value of a violator with the value of its adjacent violator such that the corrected data satisfy the monotonicity. We expect that using 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 `ipdfc` command. We start with introducing notations. Let S_k denote survival probabilities at time t_k, where $k = 1, 2, \ldots, K$ 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 the 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 $i = 1, \ldots, T$, 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 in 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, $\widehat{c}_k$; the number of events, $\widehat{d}_k$; the censoring time, $\widehat{ctime}_k$; and the 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 these time intervals, 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 like 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 where 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*`) tstart(`*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. The syntax does not handle more than one sample at a time. 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 individually, and further data management is needed to amalgamate the data from all arms of a trial, if the data are from a trial. We will illustrate this in the examples given later in this article.

3.2 Options

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

`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 (for example, days, months, or years), as specified in the publication. `tstart()` is required.

`trisk(`*varname*`)` specifies the times corresponding to the numbers of patients at risk in `nrisk()`. Set `trisk()` as `0` only if the total number of patients in the sample is known. `trisk()` is required.

`nrisk(`*varname*`)` supplies the number of patients at risk for each time in `trisk()`. Both `trisk()` and `nrisk()` are often 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 specify `trisk()` as `0`. `nrisk()` is required.

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

`saving(`*filename*[`, replace`]`)` saves the reconstructed survival data to *filename*`.dta`. `replace` allows the file to be replaced if it already exists. `saving()` is required.

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

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

`isotonic` uses isotonic regression to adjust values that violate the time-related monotonicity in `surv()`. By default, an alternative, simpler method is used to correct the values of violators by replacing the value of a violator with 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 and 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"
(4 vars, 102 obs)
```

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)
```

Because the extracted y values are survival percentages in this example, we need not use either `probability` or `failure` to convert `s`. However, we use 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
(4 vars, 87 obs)
. 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)
(213 real changes made)
```

In the amalgamated data, there are three variables—`t_ipd`, `event_ipd`, and `arm`—which are time to 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)
```
(*output omitted*)

By calling `sts graph`, we reconstruct the survival curves (see figure 1).

```
. sts graph, by(arm) title("") xlabel(0(10)70) ylabel(0(0.2)1)
> risktable(0(10)50, order(2 "Radiotherapy" 1 "Radiotherapy plus"))
> xtitle("Months") l2title("Locoregional control")
> scheme(sj) graphregion(fcolor(white))
> plot1opts(lpattern(solid) lcolor(gs12))
> plot2opts(lpattern(solid) lcolor(black))
> text(-0.38 -9.4 "cetuximab")
> legend(off)
> text (0.52 53 "Radiotherapy plus cetuximab")  text(0.20 60 "Radiotherapy")
        failure _d:  event_ipd
  analysis time _t:  t_ipd
```

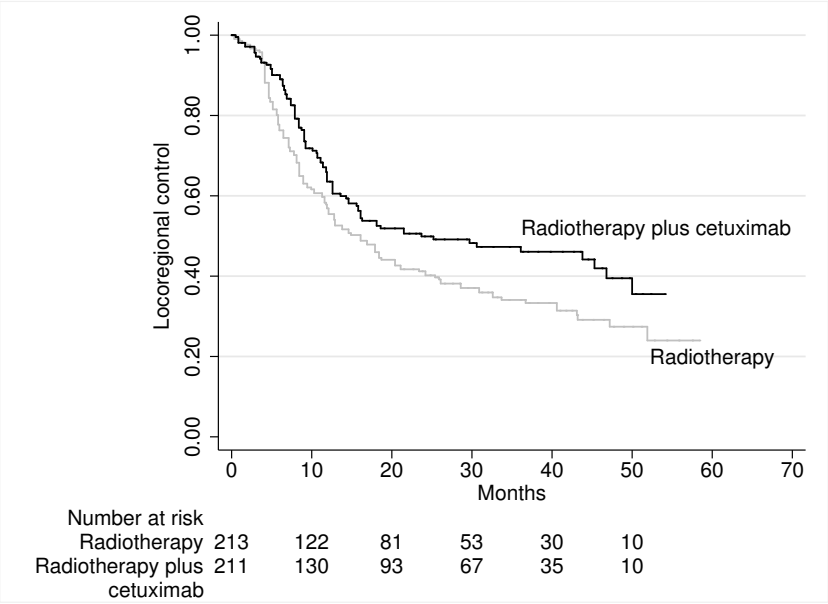

Figure 1. Reconstructed Kaplan–Meier curves for locoregional control among patients with head and neck cancer (Bonner et al. 2006). Patients are randomized to receive radiotherapy plus cetuximab or radiotherapy alone.

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

```
. stcox arm
        failure _d:  event_ipd
  analysis time _t:  t_ipd
Iteration 0:   log likelihood = -1323.3427
Iteration 1:   log likelihood = -1320.1905
Iteration 2:   log likelihood = -1320.1899
Refining estimates:
Iteration 0:   log likelihood = -1320.1899
Cox regression -- Breslow method for ties
No. of subjects =          424                 Number of obs    =         424
No. of failures =          241
Time at risk    =  8412.821523
                                               LR chi2(1)       =        6.31
Log likelihood  =   -1320.1899                 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 numbers of patients

at risk in the radiotherapy plus cetuximab arm. For this arm, based on the reconstructed data, the numbers of patients at risk are 211, 130, 93, 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 between the approximation (figure 1) and the original publication is very small for the radiotherapy arm.

In table 1, we report percentages of patients surviving one, two, and three years; median duration of locoregional control; and hazard-ratio estimates. The estimates of percentage 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 the hazard ratio (0.73, 95% CI: $[0.57, 0.94]$) estimated by Guyot et al. (2012). Because we digitize the data independently of Guyot et al. (2012), we do not expect to obtain identical data nor identical results. Though not identical, both approximated hazard ratios are similar to the published hazard ratio (0.68, 95% CI: $[0.52, 0.89]$).

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

	Original publication	Guyot et al. (2012)	`ipdfc`
Radiotherapy arm		Percent [95% CI]	Percent [95% CI]
Percent surviving one year	55	56.1 [49.6, 63.3]	56.9 [49.9, 63.2]
Percent surviving two years	41	41.1 [34.7, 48.6]	40.9 [34.2, 47.5]
Percent surviving three years	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]
Radiotherapy plus cetuximab arm		Percent [95% CI]	Percent [95% CI]
Survival rate at one year	63	64.0 [57.8, 70.9]	65.4 [58.2, 71.6]
Survival rate at two years	50	50.4 [43.9, 57.8]	51.0 [43.3, 58.6]
Survival rate at three 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]
Hazard ratio with 95% CI			
	0.68 [0.52, 0.89]	0.73 [0.57, 0.94]	0.72 [0.56, 0.93]

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 1,528 women were randomized to receive either standard chemotherapy plus bevacizumab or standard chemotherapy alone. From the analysis based on data with 30 months follow-up, Perren et al. (2011) concluded that bevacizumab improved progression-free survival in this population, with hazard ratio 0.81 (95% CI: $[0.70, 0.94]$; $P = 0.004$ from a log-rank test). Perren et al. (2011) found significant nonproportional hazards ($P < 0.001$) of the treatment effect. Kaplan–Meier curves and the associated risk table for progression-free survival were reported in their

figure 2a, on which we base our reconstruction of the survival data using `ipdfc`. Also, we use the total number of events, `tot`, because it is available.

```
. local tot0 = 464
. local tot1 = 470
. import delimited using "icon7_data_arm0.txt", clear
(4 vars, 86 obs)
. 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
(4 vars, 473 obs)
. 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 to the above codes, we use the `probability` option to specify that `surv(s)` represents survival probabilities rather than survival percentages.

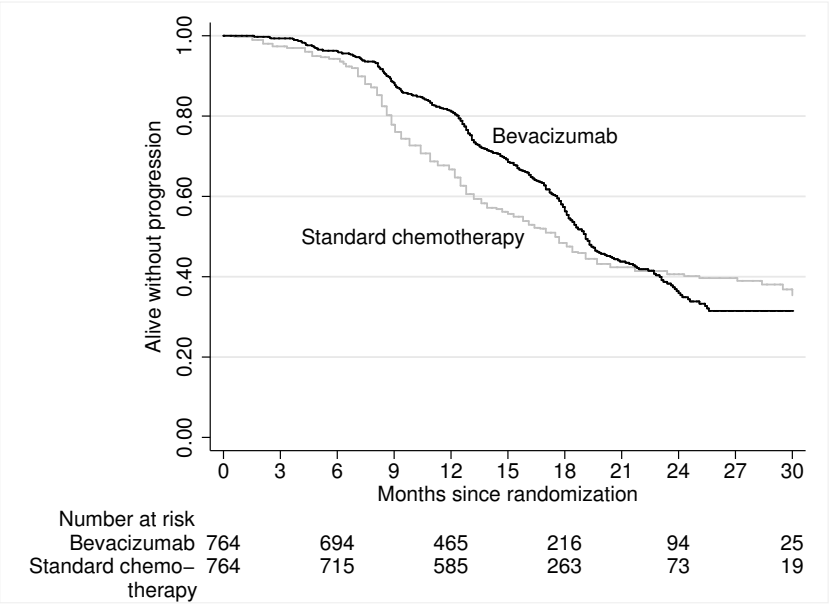

Figure 2. Reconstructed Kaplan–Meier curves for progression-free survival according to treatment group in ICON7 (Perren et al. 2011). Patients are randomized to receive standard chemotherapy plus bevacizumab or standard chemotherapy alone.

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 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 identically to the original publication numbers of 693 at 6 months and 464 at 12 months. The estimated hazard ratios, median survival time, and p-values from the log-rank test are also similar to those in the original publication. See table 2 for a comparison of summary measures.

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

	Original publication	Reconstructed data
Log-rank test	$P = 0.004$	$P = 0.009$
Nonproportional hazard test	$P < 0.001$	$P < 0.001$
Hazard ratio	0.81 (95% CI: $[0.70, 0.94]$)	0.83 (95% CI: $[0.72, 0.96]$)
Median survival time		
Chemotherapy arm	17.3	17.5 (95% CI: $[16.1, 18.7]$)
Bevacizumab arm	19.0	19.1 (95% CI: $[18.3, 19.9]$)

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 2003). In this trial, 12,218 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 the trial report. In Fox (2003), the Cox proportional hazards model was used, but the hazard-ratio estimate was not reported. It was reported in Fox (2003) that perindopril treatment was associated with a significant reduction in the composite events (cardiovascular mortality, nonfatal myocardial infarction, and resuscitated cardiac arrest), with p-value = 0.0003 from a log-rank test and absolute risk reduction of 1.9%.

We extracted the failure percentages and the associated time points, respectively, from the y axis and the x axis of the Kaplan–Meier curves in Fox's (2003) figure 2. In the following codes, we use the option `failure` to specify that the input data are failure percentages instead of the default survival percentages.

```
. import delimited using "europa_data_arm0.txt", clear
(4 vars, 743 obs)
. ipdfc, surv(s) failure isotonic tstart(ts) trisk(trisk) nrisk(nrisk)
> generate(t_ipd event_ipd) saving(temp0)
. import delimited using "europa_data_arm1.txt", clear
(4 vars, 650 obs)
. ipdfc, surv(s) failure isotonic tstart(ts) trisk(trisk) nrisk(nrisk)
> generate(t_ipd event_ipd) saving(temp1)
```

The Kaplan–Meier curves from the reconstructed data are presented in figure 3. The reconstructed curves appear nearly identical to the original. The reconstructed curves correctly reflect that the benefit of perindopril treatment began to appear at one 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 two years of follow-up (5,781 in the original report versus 5,783 in the reconstructed data).

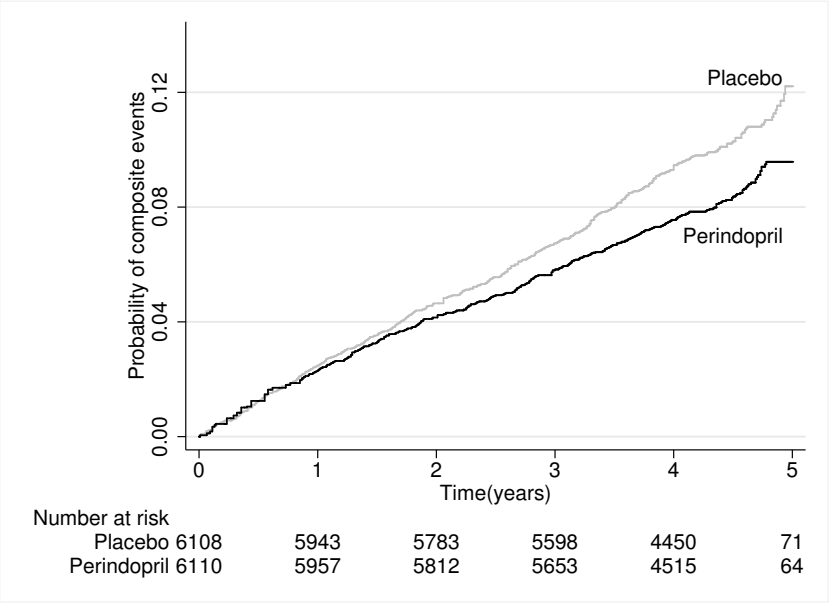

Figure 3. Reconstructed Kaplan–Meier curves for time to first occurrence of event. Patients are randomly assigned to perindopril treatment or placebo in the EUROPA trial (Fox 2003).

Using the reconstructed data, we obtain p-value = 0.0006 from a log-rank test (see 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%, similar to the 1.9% in the original publication. We are able to obtain the 95% CI $[0.80, 2.84]$ for this based on the reconstructed data. Using the 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.

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)	not applicable	0.81 $[0.72, 0.91]$	0.81 $[0.72, 0.92]$

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 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 that violates monotonicity, and allows one to use 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 also close, 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 nonproportional hazards are evident in ICON7, the reconstructed Kaplan–Meier curves and hazard-ratio estimate are in reasonable agreement with those from

the trial publication (see table 2). This suggests that nonproportional hazards may not much affect the approximation accuracy. However, further empirical evaluation of `ipdfc` in a larger number of trials, with or without obvious nonproportional hazards, is desirable; this is a topic for further research.

Where hazard ratios are not reported but Kaplan–Meier curves are available, `ipdfc` is particularly helpful because it enables the reconstruction of time-to-event data and hence allows for reanalysis of the data. For the EUROPA trial, we are able to obtain the estimate of the hazard ratio and obtain the 95% CI 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 Acknowledgments

Patrick Royston 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., D. J. Bartholomew, J. M. Bremner, and H. D. Brunk. 1972. *Statistical Inference Under Order Restrictions: Theory and Application of Isotonic Regression*. New York: Wiley.

Bonner, J. A., P. M. Harari, J. Giralt, N. Azarnia, D. M. Shin, R. B. Cohen, C. U. Jones, R. Sur, D. Raben, J. Jassem, R. Ove, M. S. Kies, J. Baselga, H. Youssoufian, N. Amellal, E. K. Rowinsky, and K. K. Ang. 2006. Radiotherapy plus cetuximab for squamous-cell carcinoma of the head and neck. *New England Journal of Medicine* 354: 567–578.

Deeks, J. J., J. P. T. Higgins, and D. G. Altman. 2008. Analysing data and undertaking meta-analyses. In *Cochrane Handbook for Systematic Reviews of Interventions*, ed. J. P. T. Higgins and S. Green, 243–296. Chichester, UK: Wiley.

Fox, K. M. 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.

Guyot, P., A. E. Ades, M. J. N. M. Ouwens, and N. J. Welton. 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., V. Torri, and L. Stewart. 1989. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. *Statistics in Medicine* 17: 2815–2834.

Perren, T. J., A. M. Swart, J. Pfisterer, J. A. Ledermann, E. Pujade-Lauraine, G. Kristensen, M. S. Carey, P. Beale, A. Cervantes, C. Kurzeder, A. du Bois, J. Sehouli, R. Kimmig, A. Stähle, F. Collinson, S. Essapen, C. Gourley, A. Lortholary, F. Selle, M. R. Mirza, A. Leminen, M. Plante, D. Stark, W. Qian, M. K. B. Parmar, and A. M. Oza. 2011. A phase 3 trial of bevacizumab in ovarian cancer. *New England Journal of Medicine* 365: 2484–2496.

van Putten, W., and P. Royston. 2017. irax: Stata module to perform isotonic regression analysis. Statistical Software Components S458406, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s458406.html.

Royston, P., and M. K. B. Parmar. 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: 2409–2421.

Tierney, J. F., L. A. Stewart, D. Ghersi, S. Burdett, and M. R. Sydes. 2007. Practical methods for incorporating summary time-to-event data into meta-analysis. *Trials* 8: 16.

Wei, Y., P. Royston, J. F. Tierney, and M. K. B. Parmar. 2015. Meta-analysis of time-to-event outcomes from randomized trials using restricted mean survival time: Application to individual participant data. *Statistics in Medicine* 34: 2881–2898.

Williamson, P. R., C. T. Smith, J. L. Hutton, and A. G. Marson. 2002. Aggregate data meta-analysis with time-to-event outcomes. *Statistics in Medicine* 21: 3337–3351.

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 centers 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 time-to-event outcome, on parametric modeling of survival data, and on novel clinical trial designs.

Appendix

Algorithm 1 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, $k = 1, \ldots, K$, where 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 $trisk_i, i = 1, \ldots, T$, where 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 $\text{lower}_i = \min\{k : t_k \geq trisk_i\}$ and $\text{upper}_i = \max\{k : t_k \leq trisk_{i+1}\}$.

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

$$\widehat{nc}_i = S_{\text{lower}_{i+1}}/S_{\text{lower}_i} \times nrisk_i - nrisk_{i+1}$$

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

$$\widehat{ctime}_c = t_{\text{lower}_i} + c \times (t_{\text{lower}_{i+1}} - t_{\text{lower}_i})/(\widehat{nc}_i + 1)$$

where $c = 1, \ldots, \widehat{nc}_i$. 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

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

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

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

Note that $t_{k'} \leq t_k$, 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 $k \in [\text{lower}_i, \text{upper}_i]$. Thus, $\widehat{nrisk}_{i+1} = \widehat{n}_{\text{upper}_i+1}$.

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

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

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

end if

Step 5. Repeat steps 1–4 until $i + 1 = T$.

end if

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

$$\widehat{nc}_T = \min\left(\frac{t_{\text{upper}_T} - t_{\text{lower}_T}}{t_{\text{upper}_{T-1}} - t_{\text{lower}_1}} \times \sum_{i=1}^{T-1} \widehat{nc}_i;\ nrisk_T\right)$$

We then run steps 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 $\sum_{k=1}^{\text{upper}_{T-1}} \widehat{nd}_k$.

if $\sum_{k=1}^{\text{upper}_{T-1}} \widehat{nd}_k \geq D$ **then**

Stop the algorithm.

end if

if $\sum_{k=1}^{\text{upper}_{T-1}} \widehat{nd}_k < D$ **then**

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

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

Repeat steps 2–3 and steps 7–8 for the last interval.

end if

end if