A Method for Reconstructing Individual Patient Data From Kaplan-Meier Survival Curves That Incorporate Marked Censoring Times

Basia Rogula, Greta Lozano-Ortega, and Karissa M. Johnston

Abstract

Introduction. Access to individual patient data (IPD) can be advantageous when conducting cost-effectiveness analyses or indirect treatment comparisons. While exact times of censoring are often marked on published Kaplan-Meier (KM) curves, an algorithm for reconstructing IPD from such curves that allows for their incorporation is presently unavailable. Methods. An algorithm capable of incorporating marked censoring times was developed to reconstruct IPD from KM curves, taking as additional inputs the total patient count and coordinates of the drops in survival. The reliability of the algorithm was evaluated via a simulation exercise, in which survival curves were simulated, digitized, and then reconstructed. To assess the reliability of the reconstructed curves, hazard ratios (HRs) and quantiles of survival were compared between the original and reconstructed curves, and the reconstructed curves were visually inspected. Results. No systematic differences were found in HRs and quantiles in the original versus reconstructed curves. Upon visual inspection, the reconstructed IPD provided a close fit to the digitized data from the published KM curves. Inherent to the algorithm, censoring times were incorporated into the reconstructed data exactly as specified. Conclusion. This new algorithm can reliably be used to reconstruct IPD from reported KM survival curves in the presence of extractable censoring times. Use of the algorithm will allow health researchers to reconstruct IPD more closely by incorporating censoring times exactly as marked, requiring as additional inputs the total patient count and coordinates of the drops in survival.

Keywords
algorithm, data recovery, individual patient data, Kaplan-Meier, survival analysis

Introduction

Access to individual patient data (IPD) can be advantageous and expand the range of analytic options when conducting analyses such as cost-effectiveness analyses or indirect treatment comparisons.1,2 Frequently, key evidence to inform these analyses comes from time-to-event data available in aggregate form only, from published Kaplan-Meier (KM) curves. CONSORT guidelines recommend that for time-to-event outcomes, hazard ratios (HRs) should be used as a comparative measure.3 However, a HR for the comparison of interest is not always available, so reconstruction of IPD from the reported KM curves can help fill in this data gap. In addition to calculating HRs, reconstruction of IPD is also required to be able to fit parametric curves, if it is of interest to extrapolate survival curves beyond the follow-up period covered by the KM analysis. The reconstruction of...
IPD is also particularly useful for meta-analyses, allowing for the combination of patients from multiple publications. These meta-analyses require the assumption that the populations are similar enough to validate their pooling. The quality of these secondary analyses performed on reconstructed IPD are affected by the accuracy of the reconstructed data relative to the original source data, and when original IPD are not available they may need to be reconstructed. Utilizing an algorithm that makes use of all available survival curve data leads to maximization of accuracy in the data reconstruction process. In the case of survival data, this can include the timing of events, the numbers at risk, a total event count, and time points at which individuals are censored.

None of the existing methods to reconstruct IPD from KM curves make use of marked censoring times, which are typically indicated on published KM curves with symbols such as crosses or ticks. An algorithm developed by Guyot and colleagues utilizes, where available, information on number of events, survival probabilities, and numbers at risk to reconstruct IPD. A modified version of Guyot’s algorithm was implemented with adaptions as a STATA command by Wei and Royston. A method was developed to reconstruct IPD from KM survival curves capable of incorporating censoring times exactly as they are marked, and avoid making assumptions about the distribution of censoring times. By improving the quality of the reconstructed IPD, this method allows researchers to enhance the quality of secondary analyses performed using time-to-event data.

Methods
A method was developed to reconstruct IPD from KM survival curves which incorporate censoring times exactly as they are marked, under the context of right-censored data. The method follows a greedy algorithm, that is, one that aims to find an optimal overall solution by making the locally optimal choice at each step. Specifically, the algorithm runs through each sequential drop in the KM curve, estimating the number of events occurring at each drop that results in the closest match to the desired survival value. Censoring points occurring during the interval directly before the drop are added to the reconstructed IPD before the number of events at the drop is estimated. Unique to this method, the location and number of censoring times are not estimated by the algorithm; they are entered as data inputs to the algorithm after being digitized by the analyst.

The data required are the following: 1) the total number at risk at time zero, n; 2) the coordinates of the bottoms of the J drops in survival, that is, two vectors: i) the extracted x-axis coordinates, \(x = (x_1, x_2, \ldots, x_J)\) and ii) the y-axis coordinates, \(y = (y_1, y_2, \ldots, y_J)\); and 3) a vector of the K censoring times, \(c = (c_1, c_2, \ldots, c_K)\). The vectors \(x, y\), and \(c\) can be extracted using digitization software such as DigitizeIt (www.digitizeit.de). Alternatively, in some cases the curves can be directly recovered from the electronic format in which they were published.

The x-coordinate of the last point in the curve represents the longest follow-up time. If the last point is at the end of a vertical line, its coordinates should be included in the coordinates of drops in survival. If the last point is at the end of a horizontal line, it represents censoring, and its time could be included in \(c\) or alternatively its coordinates can be included in \(x\) and \(y\). The algorithm could be run without the incorporation of marked censoring points; in this case, the coordinates of the last point should be included in \(x\) and \(y\).

A flowchart of the algorithm is shown in Figure 1. All individuals are first set to be censored at the longest follow-up time, initializing the reconstructed IPD dataset. This step is important as it ensures the denominator is set correctly when estimating numbers of events occurring at each drop in the subsequent steps. Each drop in
survival is then visited, starting at the first drop at $x_1$ and moving forwards until the last one at $x_J$, visiting each drop only once. At each drop, $x_j$, first, censoring events in $e$ occurring in the time interval directly before the drop (between $x_{j-1}$ and $x_j$, inclusive of $x_{j-1}$ but exclusive of $x_j$) are added to the reconstructed data. Then, the number of events to add at $x_j$ is estimated by finding the number of events which results in a survival probability at $x_j$ from a KM analysis fitted to the reconstructed IPD closest to the digitized value, $y_j$. This can be achieved using a number of methods. One strategy, illustrated in Figure 1, is to start at zero events, and add more events until the fitted survival probability drops below $y_j$, then finding whether that survival probability or the one prior is closer to the digitized value. In some cases, no events will be added. After the number of events to add at the drop

---

Figure 1 Algorithm flowchart.
is determined, the events are added to the reconstructed IPD before moving to the next drop. After all drops have been visited and all censoring points have been added (including any after the last drop), the reconstructed IPD are complete.

The method was implemented as function getIPD in package KMtoIPD using R 3.4.0. This R package can be installed from GitHub with devtools::install_github("rogula/KMtoIPD"). The function reads in 1) the number of individuals at time zero, 2) the coordinates of the drops in survival, and 3) censoring times, and returns the reconstructed IPD.

Simulation Validation

The reliability of the algorithm was evaluated via a simulation exercise, in which survival curves for two hypothetical treatments were simulated from three survival distributions (loglogistic, lognormal, and exponential; two curves with varying parameters from each distribution), for each of four sample sizes (20, 50, 100, 200). This resulted in a total of $2 \times 3 = 6$ data sets for each sample size. The code to generate the simulated data is included in the supplementary material (Appendix 3) and documents the model parameters used. The HRs between the two treatment curves simulated for each sample size and distribution combination were calculated. All resulting KM curves were digitized, and IPD were reconstructed with the R function using the digitized data. For the two treatment curves simulated for each sample size and distribution combination, the HR was calculated for the reconstructed IPD and compared with the HR calculated with the original data to assess the reliability of the reconstructed curves. Quantiles of survival (25%, 50% [median], 75%) were compared between the reconstructed and original curves. Reconstructed curves were visually inspected for their fit to the digitized points.

The algorithm developed by Guyot and colleagues was applied to the same curves; the numbers at risk at 10-unit time intervals were used as input in Guyot algorithm, and the reconstructed curves were compared between the two approaches for data reconstruction.

Results

Illustrative Example

As an example, IPD were reconstructed from a published figure showing KM survival curves by lymphocyte count ($\geq 1000/\mu L$, $< 1000/\mu L$) among patients with advanced,
The published curves are shown in Figure 2. The solid gray circles in Figure 3 show the points on the curve for the \( \frac{1000}{m} \) group that were digitized. The resulting digitized points and digitized times of censoring are shown in Appendix 1. The function call in R to reconstruct the IPD for the \( \frac{1000}{m} \) group would be
\[
\text{reconstr(IPD} = \text{getIPD(30, t, S, cens.t = cens_t)}.
\]
The reconstructed KM curves for both lymphocyte groups are shown in Figure 4. The reconstructed HR of \( \geq 1000/\mu L \) versus \( < 1000/\mu L \) and its 95% confidence interval (CI) limits, both reported to two decimal places, matched the reported HR exactly: 0.49 (0.26-0.91). The number of events in the reconstructed IPD was identical to the reported numbers in both groups (19 for \( \geq 1000/\mu L \) and 22 for \( < 1000/\mu L \)). The reconstructed medians + 95% CIs were identical to original values for both lymphocyte groups (22.2 [13.6, –] for \( \geq 1000/\mu L \) and 13.3 [6.1, 17.5] for \( < 1000/\mu L \)). The numbers in the reconstructed numbers at risk table were identical to the original table. As shown in Figure 5, the reconstructed curves fit very closely through the digitized points.

**Algorithm Validation**

KM curves fitted to the reconstructed IPD from the simulated data are shown in Figure 6. Upon visual inspection, the reconstructed IPD provided a close fit to the digitized data from the original KM curves. Inherent to the algorithm, censoring times were incorporated into the reconstructed data exactly as specified. No systematic differences were found in HRs comparing treatment 2 versus treatment 1 across scenarios (Figure 6 and Table 1). Quantiles were similar between the reconstructed and original curves (Table 1).

Results from running the data reconstruction method developed by Guyot and colleagues are shown in Appendix 2. For the smallest sample size of 20, the novel method presented here provided a visually closer fit to the original KM curves, most notably in the curve tails. The methods performed similarly for the larger sample sizes which had lower proportions of censoring events. When comparing reconstructed HRs, both methods generally provided a close match to the original HRs. However, there were a few scenarios for which the reconstructed HRs from this novel method were notably closer: lognormal, \( n = 50 \) (original: 1.65, Guyot: 1.61, novel method: 1.65), exponential, \( n = 20 \) (original: 3.08, Guyot: 3.15, novel method: 3.08), and exponential, \( n = 200 \) (original: 3.55, Guyot: 3.46, novel method: 3.55). The locations of censoring points were different in the data reconstructed using the Guyot algorithm versus this algorithm; this algorithm incorporated censoring points at their exact marked times.

**Discussion**

Needing to recreate IPD from published KM curves is a common occurrence in medical research. The
methodology proposed here allows researchers to reconstruct IPD from KM curves in the presence of marked censoring times, incorporating the censoring times exactly as they are marked. Based on the simulation performed, our method provides an advantage over existing methods, as incorporating censoring points in the recreation of IPD led to a closer fit to the KM curve.

One limitation is that censoring times, if being digitized from a survival curve, should be clearly marked and easily distinguishable for this algorithm to reliably work. Censoring times are not always marked or clearly distinguishable. The simulation example was run for sample sizes between 20 and 200. For larger sample sizes, it may be difficult to clearly distinguish individual censoring marks, particularly if there is a relatively high rate of censoring leading to crowding on the curve. For this reason, sample sizes larger than 200 were not generated for the simulation, and the proposed method may be more appropriate for use in relatively small sample sizes (depending on the rate of censoring and image quality). Even if censoring points appear clearly marked and distinguishable, there may be multiple censoring events that happen at the same time and overlap in the figure. Sometimes overlapping censoring points can become apparent once the data are reconstructed and the curve does not provide an adequate fit, and the manual editing of censoring points can improve the fit and/or result in better-matching numbers at risk tables. Note that this is

### Table 1  Comparison of the Original and Reconstructed Individual Patient-Level Survival Data

| Distribution  | n  | Treatment | Hazard ratio: Treatment 2 vs. Treatment 1 | Quantiles, 25%, 50%, 75% |
|---------------|----|-----------|----------------------------------------|---------------------------|
|               |    |           | Original                  | Reconstructed             |
| Loglogistic   | 20 | 1         | 1.79 (0.74, 4.29)         | 1.78 (0.74, 4.29)         |
|               |    | 2         | 2.00 (1.17, 3.41)         | 2.03 (1.19, 3.47)         |
|               | 50 | 1         | 1.68 (1.16, 2.43)         | 1.69 (1.17, 2.44)         |
|               |    | 2         | 1.76 (1.38, 2.24)         | 1.76 (1.38, 2.24)         |
|               | 100| 1         | 2.23 (0.92, 5.42)         | 2.25 (0.93, 5.46)         |
|               |    | 2         | 1.65 (1.02, 2.69)         | 1.65 (1.02, 2.68)         |
|               | 200| 1         | 1.93 (1.37, 2.73)         | 1.95 (1.38, 2.76)         |
|               |    | 2         | 2.56 (2.04, 3.22)         | 2.52 (2.01, 3.16)         |
| Lognormal     | 20 | 1         | 3.08 (1.18, 8.08)         | 3.08 (1.18, 8.05)         |
|               |    | 2         | 2.29 (1.33, 3.93)         | 2.30 (1.34, 3.95)         |
|               | 50 | 1         | 1.85 (1.28, 2.69)         | 1.84 (1.27, 2.67)         |
|               |    | 2         | 3.55 (2.73, 4.62)         | 3.55 (2.74, 4.61)         |
| Exponential   | 20 | 1         | 26.0, NR, NR              | 26.0, NR, NR              |
|               |    | 2         | 6.9, 18.5, 55.9           | 6.9, 18.4, 56.0           |
|               | 50 | 1         | 20.7, 57.0, NR            | 20.6, 57.0, NR            |
|               |    | 2         | 9.2, 28.1, 46.5           | 9.3, 28.1, 46.6           |
|               | 100| 1         | 21.0, 50.6, NR            | 21.1, 50.3, NR            |
|               |    | 2         | 11.1, 24.8, 53.3          | 11.1, 24.8, 53.2          |
|               | 200| 1         | 23.8, 61.9, NR            | 24.0, 63.8, NR            |
|               |    | 2         | 7.6, 16.2, 32.0           | 7.4, 16.1, 34.7           |

NR, not reached.

![Figure 5](image-url)  
**Figure 5** Example of checking the fit of reconstructed Kaplan-Meier survival curves through digitized points.
Figure 6 Kaplan-Meier survival curves reconstructed based on digitized survival data from various distributions and sample sizes. Black points mark digitized survival times and probabilities. Original and reconstructed hazard ratios are reported for comparison. Abbreviations: Orig., original; Recons., reconstructed.
not an issue for the last point in the curve where there are often many individuals censored at once; the algorithm initially sets everyone to censored at the longest follow-up time and those without a set event time or earlier set censoring time remain censored at the end of the curve. Also, some KM curve plotting functions do not mark censoring points recorded as occurring at the same time as an event. These censoring points would thus be missed in data reconstruction.

Drops in the KM survival curve are also sometimes difficult to see which can influence the quality of the reconstructed data. For example, overlapping curves can cause obstruction. If the locations of the drops can be estimated, then this algorithm could still be applied with the caveat that quality may be compromised. However, this is a limitation to all IPD reconstruction algorithms.

Another limitation is that the reliability of the reconstructed data relies on the data inputs and could be influenced by user error introduced in the digitizing process. For example, consistently marking points to digitize too low or too high could lead to bias in the reconstructed data. However, this is also a limitation of all IPD reconstruction algorithms. We suggest comparing as many results as possible in the original publication to those from the reconstructed data to assess the quality of the reconstructed data. HRs, numbers at risk tables, event counts, quantiles of survival, and visual fit of reconstructed curves through digitized coordinates could all be used to assess the quality of the reconstructed data.

Conclusions

None of the existing algorithms to reconstruct IPD from KM survival curves utilize censoring times often marked directly on the curves. In the presence of extractable censoring times, this novel method can reliably reconstruct IPD incorporating censoring times exactly as they are marked, requiring as additional inputs the total patient count and coordinates of the drops in survival. In this way, neither a number at risk table nor event count are required for the reconstruction, no assumption about censoring distribution is made, and censoring times are incorporated into the reconstructed IPD in the closest way possible. As this method relies on the quality of the extracted censoring times, it is more appropriate to use for KM curves where censoring times are easily distinguishable.

Authors’ Note

The abstract was presented at the 40th annual North American meeting for the SMDM (https://smdm.confex.com/smdm/2018/meetingapp.cgi/Paper/11679). This study was an independent research project conducted by Broadstreet HEOR.

ORCID iD

Basia Rogula https://orcid.org/0000-0002-2115-8275

Supplemental Material

Supplemental material for this article is available on the Medical Decision Making Policy & Practice website at https://journals.sagepub.com/home/mpp.

References

1. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ*. 2010;340:c221.
2. Simmonds MC, Higgins JP, Stewart LA, Tierney JF, Clarke MJ, Thompson SG. Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. *Clin Trials*. 2005;2(3):209–17.
3. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ*. 2010;340:c869.
4. Guyot P, Ades AE, Ouwens MJNM, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Med Res Methodol*. 2012;12(1):9.
5. Wei Y, Royston P. Reconstructing time-to-event data from published Kaplan-Meier curves. *Stat Asa*. 2017;17(4):786–802.
6. Liu N, Zhou Y, Lee JJ. IPD from KM: reconstruct individual patient data from published Kaplan-Meier survival curves. *BMC Med Res Methodol*. 2021;21(1):111.
7. Hoyle MW, Henley W. Improved curve fits to summary survival data: application to economic evaluation of health technologies. *BMC Med Res Methodol*. 2011;11(1):139.
8. Irvine AF, Waise S, Green EW, Stuart B. A non-linear optimisation method to extract summary statistics from Kaplan-Meier survival plots using the published P value. *BMC Med Res Methodol*. 2020;20(1):269.
9. Liu Z, Rich B, Hanley JA. Recovering the raw data behind a non-parametric survival curve. *Syst Rev*. 2014;3:151.
10. Takahashi M, Takahashi S, Araki N, et al. Efficacy of trabectedin in patients with advanced translocation-related sarcomas: pooled analysis of two phase II studies. *Oncologist*. 2017;22(8):979–88.