A reconstructed melanoma data set for evaluating differential treatment benefit according to biomarker subgroups

Jaya M. Satagopan*, Alexia Iasonos, Joseph G. Kanik

Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA

A R T I C L E   I N F O

Article history:
Received 9 March 2017
Received in revised form 23 April 2017
Accepted 3 May 2017
Available online 5 May 2017

A B S T R A C T

The data presented in this article are related to the research article entitled “Measuring differential treatment benefit across marker specific subgroups: the choice of outcome scale” (Satagopan and Iasonos, 2015) [1]. These data were digitally reconstructed from figures published in Larkin et al. (2015) [2]. This article describes the steps to digitally reconstruct patient-level data on time-to-event outcome and treatment and biomarker groups using published Kaplan-Meier survival curves. The reconstructed data set and the corresponding computer programs are made publicly available to enable further statistical methodology research.

© 2017 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Specifications Table

| Subject area       | Biostatistics          |
|--------------------|------------------------|
| More specific subject area | Clinical Biostatistics |
| Type of data       | Text file              |
| How data was acquired | Digital extraction techniques and statistical methods using Adobe Illustrator [3], DigitizeIt software package[4] and the R programming language [5] |

DOI of original article: http://dx.doi.org/10.1016/j.cct.2017.02.007
* Corresponding author.
E-mail address: satagopj@mskcc.org (J.M. Satagopan).
A total of 843 melanoma patients with positive or negative programmed death 1 ligand expression were randomized to receive nivolumab monotherapy, ipilimumab monotherapy or combination therapy. The study has 6 subgroups defined by 3 treatments and two levels of programmed death 1 ligand expression.

Individual patient data were extracted from Kaplan-Meier figures and the number at risk reported below the figures for each of the 6 subgroups.

The reconstructed data and R functions are available at https://www.mskcc.org/sites/default/files/node/137932/documents/2017-04-20-14-31-36/dataexample.zip

J. M. Satagopan, A. Iasonos, Measuring differential treatment benefit across marker specific subgroups: the choice of outcome scale, Contemp Clin Trials. [1]

Value of the data

- The data set presents reconstructed information on progression free survival in metastatic melanoma patients and could be used by other researchers.
- This reconstructed data set allows other researchers to develop statistical methodologies for evaluating differential treatment benefit according to biomarker level.
- This reconstructed data set allows other researchers to extend the statistical analyses and compare the results to other similar studies.

1. Data

We present reconstructed data based on Fig. 1B and C of Larkin et al. [2]. The reconstructed data set includes information on time to disease progression, progression status, treatment, and the status of programmed death 1 ligand expression for 843 metastatic melanoma patients: 620 with negative expression (210 randomized to the combination therapy arm, 202 to ipilimumab monotherapy and 208 to nivolumab monotherapy) and 223 with positive expression (68 randomized to the combination therapy arm, 75 to ipilimumab monotherapy and 80 to nivolumab monotherapy). The
reconstructed data are only approximate data to facilitate statistical methodology research, and do not represent actual patient-level data. These reconstructed data are new and original in the sense that the reconstructed time to progression free survival and progression status data has not been published elsewhere.

Fig. 2. Select the overall image and head to the top option to “Image Trace”, selecting the arrow on the right and choosing “High Fidelity Photo”. Next, select the button on the right of where Image Trace was, “Expand”.

Fig. 3. The figure in Adobe Illustrator after expanding via Image Trace.

Fig. 4. It is now possible to select each line with just a click of the button. Because the trace was for a “High Fidelity Photo”, Adobe Illustrator is able to understand that left clicking an orange line should highlight the entirety of the orange line and nothing else as displayed in this figure. Now, each line can be removed to obtain separate files for each line of data.
2. Experimental design, materials and methods

We used the following steps to reconstruct data from Figs. 1B and 1C of Larkin et al. [2].

**Step 1: Isolating individual lines from Kaplan-Meier figures**

Fig. 1C of Larkin et al. [2] contains 3 lines representing the Kaplan-Meier estimates of survival probabilities for patients with negative programmed death 1 ligand expression randomized to nivolumab monotherapy, ipilimumab monotherapy and combination therapy. Isolate these 3 lines using Adobe Illustrator [3], as described in Figs. 1–7. Use similar methods to isolate the 3 lines from Fig. 1B of Larkin et al. [2] that correspond to patients with positive programmed death 1 ligand expression. Save the isolated lines as separate jpeg files.
**Step 2: Digital extraction of time and survival probabilities**

Consider a jpeg file containing a single line – for example, the jpeg file corresponding to Fig. 7. Launch the Digitizelt software package [4] in your computer and open this jpeg file. To digitize the
line, select the desired minimum and maximum points on the horizontal (i.e., x) and vertical (i.e., y) axes, click the “Line” icon and left click the mouse on any part of the line. This will digitize the line and show the times (x-axis) and survival probability estimates (y-axis) in the output frame, which can

Fig. 9. Create an R object “numbers.below.figure” as a list containing 6 elements. Each element is a vector containing the numbers at risk given below Figs. 1B and C of Larkin et al. [2].

```r
> numbers.below.figure = list(  
>   pdll.negative.nivo = c(208, 192, 178, 108, 105, 98, 88, 80,  
>    76, 74, 63, 50, 31, 24, 9, 5, 4, 2, 1, 1),  
>   pdll.negative.ipi = c(202, 183, 166, 82, 72, 59, 44, 39, 35,  
>    31, 26, 22, 12, 8, 3, 1),  
>   pdll.negative.combo = c(210, 195, 181, 142, 138, 123, 112, 106,  
>    105, 86, 88, 79, 62, 36, 13, 9, 6, 2, 1),  
>   pdll.positive.nivo = c(80, 76, 71, 57, 56, 54, 31, 45, 49, 43, 38,  
>    32, 16, 13, 5, 4, 2,  
>   pdll.positive.ipi = c(78, 69, 66, 40, 35, 24, 22, 21, 21, 17, 16, 15,  
>    9, 6, 3, 2, 2),  
>   pdll.positive.combo = c(68, 63, 61, 53, 52, 47, 44, 42, 42, 39,  
>    34, 24, 16, 12, 3, 1, 1)  
> )
```

Fig. 10. Create an R object “time” as a list containing 6 vectors. Each vector is a set of integers giving the time points along the x-axis of Figs. 1B and C of Larkin et al. [2]. The commented items referred to as “arm indicator” denote the treatment/biomarker arm. This is a simple book-keeping strategy for the user to note that the first file to be digitized corresponds to data from patients with negative programmed death 1 ligand expression receiving nivolumab (denoted “pdll.neg.nivo”), the second file corresponds to negative programmed death 1 ligand expression receiving ipilimumab (denoted “pdll-neg-tp” etc.)
Fig. 11. The R object “individual.data” will contain the patient-level digitized data. This object is assembled by running the functions preprocess.digitized.data (in program-1.R) and Guyot.individual.data (in program-2.R) using the (x,y) data sets corresponding to each of the 6 digitized lines. The “for” loop runs these functions for each (x,y) data set.

Fig. 12. R output showing the first 20 rows of the digitized patient level data. These are the first 20 rows of the object “individual.data”. Column 1 gives the progression free survival time, Column 2 is the event status (1 = disease progression, 0 = no progression). Column 3 is treatment arm number indicating the treatment/biomarker arm, which takes values 1, 2, 3, 4, 5 or 6 (see Fig. 10). These first 20 patients have treatment arm number as 1 in Column 3 since these are patients with negative programmed death 1 ligand expression receiving nivolumab treatment. The data for all the 843 patients can be downloaded from https://www.mskcc.org/sites/default/files/node/137932/documents/2017-04-20-14-31-36/dataexample.zip.
be saved as a text file. The demo video in the DigitizeIt software page [4] gives a detailed description of this step. Apply this step to each jpeg file to obtain 6 text files.

**Step 3: Reconstructing patient-level data**

To obtain patient-level data, first pre-process the (x,y) values corresponding to each line obtained in Step 2 using Program 1. Next, use these parameters as the input for Program 2, which is an R function written by Guyot et al. [6], to obtain the reconstructed patient-level data. These steps are shown in Figs. 8–14.
Funding sources

This work was supported by research grants R01 CA137420, R01 CA197402 and P30 CA008748 from the National Cancer Institute, USA, and grant UL1RR024996 from the Clinical and Translational Science Center at Weill Cornell Medical College, New York, USA. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Acknowledgements

The authors thank an anonymous reviewer for insightful comments that improved the presentation.

Transparency document. Supporting information

Transparency data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.dib.2017.05.005.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.dib.2017.05.005.

References

[1] J.M. Satagopan, A. Iasonos, Measuring differential treatment benefit across marker specific subgroups: the choice of outcome scale, Contemp. Clin. Trials (2017), http://dx.doi.org/10.1016/j.cct.2017.02.007.
[2] J. Larkin, V. Chiarion-Sileni, R. Gonzalez, J.J. Grob, C.L. Cowey, C.D. Lao, D. Schadendorf, R. Dummer, M. Smylie, P. Rutkowski, P.F. Ferrucci, A. Hill, J. Wagstaff, M.S. Carlino, J.B. Haanen, M. Maio, I. Marquez-Rodas, G.A. McArthur, P.A. Ascierto, G.V. Long, M.K. Callahan, M.A. Posstow, K. Grossman, M. Szol, B. Dreno, L. Bastholt, A. Yang, L.M. Rollin, C. Horak, F.S. Hodi, J. D. Wolchok, Combined nivolumab and ipilimumab or monotherapy in untreated melanoma, N. Engl. J. Med. 373 (2015) 23–34.
[3] M. Golding, Adobe Illustrator CS5: for Web and Interactive Design, Lynda.com, California, 2010.
[4] DigitizeIt, Digitizer software – digitize a scanned graph or chart into (x,y)-data. (http://www.digitizeit.de/), 2008 (accessed 20.04.17).
[5] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2008.
[6] P. Guyot, A.E. Ades, M.J. Ouwens, N.J. Welton, Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves, BMC Med. Res. Methodol. 12 (2012) 9.