Data in Brief 12 (2017) 667–675

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

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

ARTICLE INFO

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

ABSTRACT

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.

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).

http://dx.doi.org/10.1016/j.dib.2017.05.005

Data format	*Raw*
Experimental factors	*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.*
Experimental features	*Individual patient data were extracted from Kaplan-Meier figures and the number at risk reported below the figures for each of the 6 subgroups*
Data source location	*Kaplan-Meier figures published in Figs. 1B and 1C of Larkin et al. [2]*
Data accessibility	*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
Related research article	*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

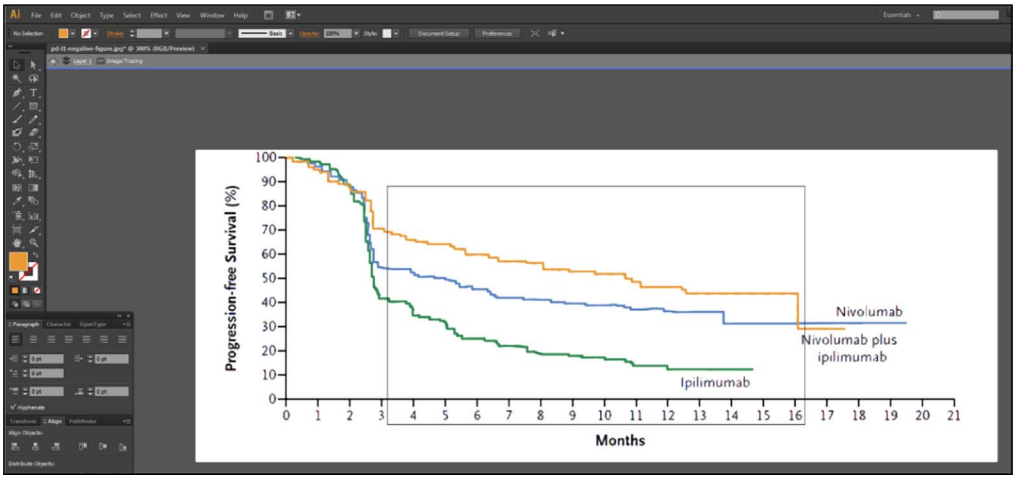

Fig. 1. Fig. 1C of Larkin et al. [2] imported into Adobe Illustrator.

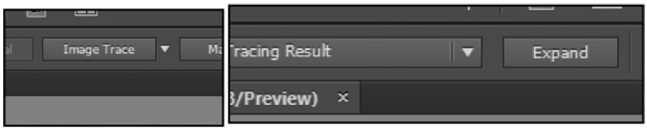

Fig. 2. Select the overall image and head to the top option to "Image Trace", selecting the arrow on the right and choosing "High Fidelity Photo". Next, select the button on the right of where Image Trace was, "Expand".

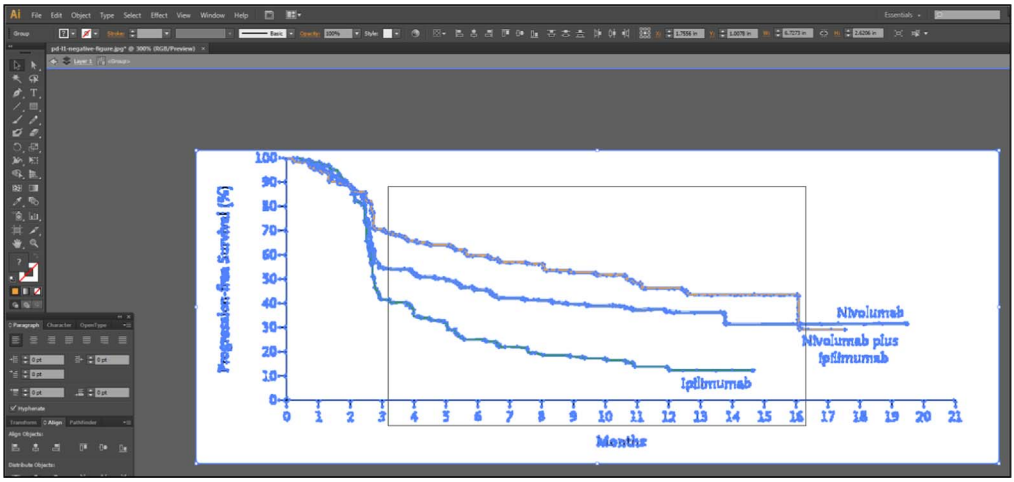

Fig. 3. The figure in Adobe Illustrator after expanding via Image Trace.

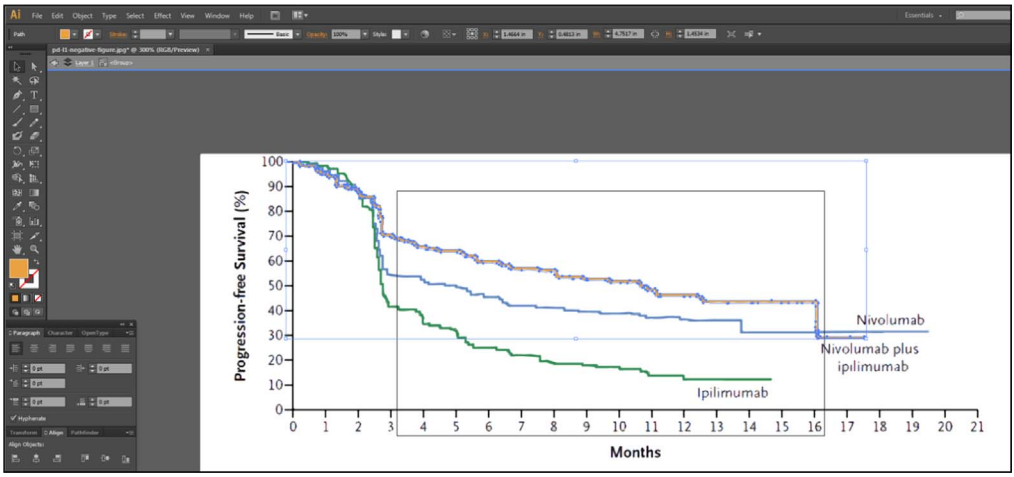

Fig. 4. It is now possible to select each line with just a click of the button. Because the trace was for a "High Fidelity Photo", Adobe Illustrator is able to understand that left clicking an orange line should highlight the entirety of the orange line and nothing else as displayed in this figure. Now,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

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.

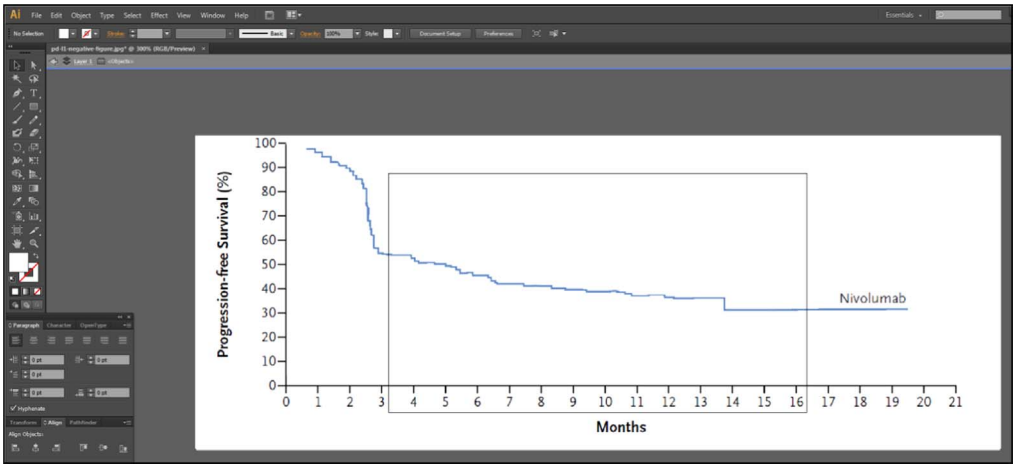

Fig. 5. The isolated Nivolumab line in Adobe Illustrator.

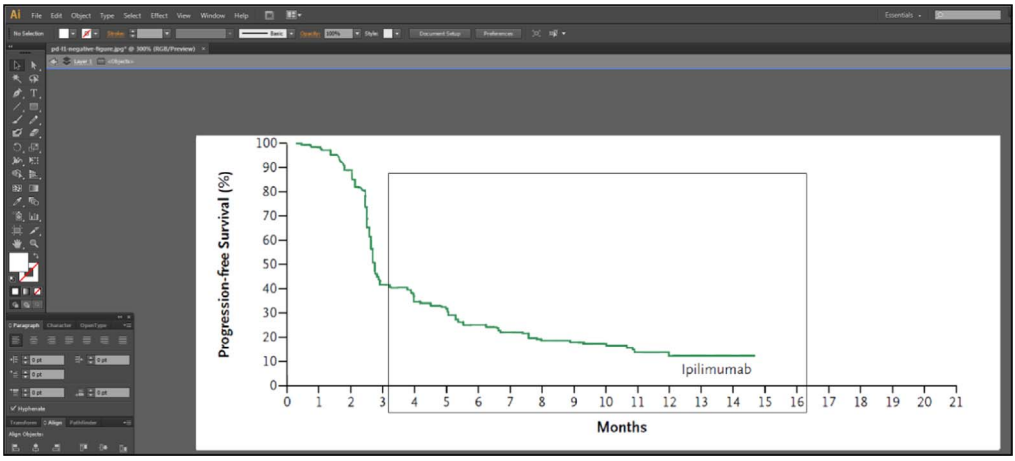

Fig. 6. The isolated Ipilimumab line in Adobe Illustrator.

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.

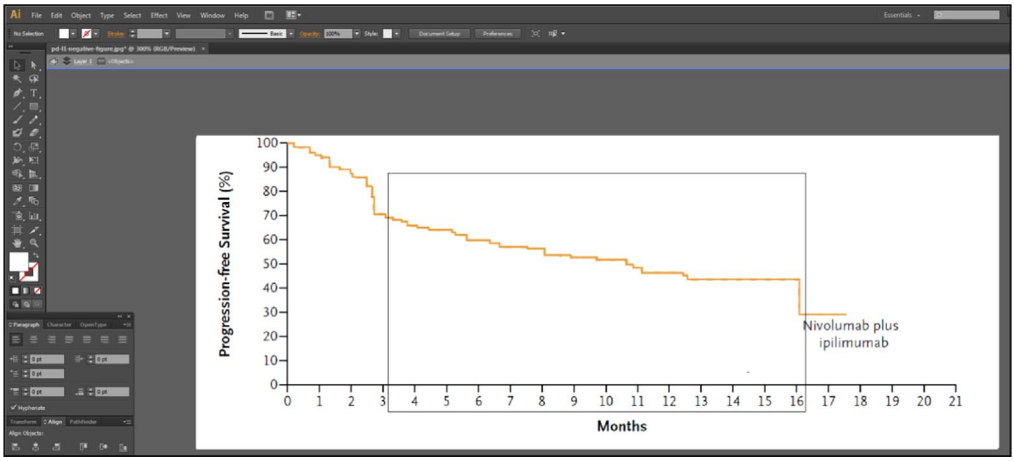

Fig. 7. The isolated Nivolumab plus Ipilimumab (combination therapy) line in Adobe Illustrator.

```
>
> source("program-1.R")
> source("program-2.R")
>
>
> ##############
> #
> # First, digitize each line in Figure 1B and 1C of Larkin et al (2015, PMID: 26027431)
> # to obtain the (x,y) coordinates for each line that correspond to times and
> # survival probabilities. This is described in Step 2 of the manuscript
> #
> # For each line, this will result in a matrix with 2 columns.
> # Column 1 is time and Column 2 is survival probability for that time.
> # This will be given for various time points.
> #
> # This will result in a total of 6 such data files, one file per line.
> # There are 3 lines in Figure 1B and 3 in Figure 1C or Larkin et al. Hence, 6 files.
> #
> # We have named 6 these files as follows:
> #
> # pdl1-negative-nivo.txt, pdl1-negative-ipi.txt, pdl1-negative-combo.txt,
> # pdl1-positive-nivo.txt, pdl1-positive-ipi.txt, pdl1-positive-combo.txt
> #
> # The R object digitized.file.names, given below, contains the names of these files.
> #
> ###############
>
>
> digitized.file.names = c("pdl1-negative-nivo.txt", "pdl1-negative-ipi.txt",
+                                    "pdl1-negative-combo.txt", "pdl1-positive-nivo.txt",
+                                    "pdl1-positive-ipi.txt", "pdl1-positive-combo.txt")
```

Fig. 8. First, read the two programs "program-1.R" and "program-2.R" using the "source" command in R. Here "program-1.R" contains the R function "preprocess.digitized.data" to perform the pre-processing step, and "program-2.R" contains the R function "Guyot.individual.data" that performs survival probability inversion steps described by Guyot et al. [6] to reconstruct patient-level data. These functions can be downloaded from https://www.mskcc.org/sites/default/files/node/137932/documents/2017-04-20-14-31-36/dataexample.zip. Next, create an R object "digitized.file.names", which is a character vector of the names of the text files containing the (x,y) data for the 6 lines. We have named the files as "pdl1-negative-nivo.txt", "pdl1-negative-ipi.txt" etc.

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 DigitizeIt software package [4] in your computer and open this jpeg file. To digitize the

```
>
> ################################
> #
> # Now, look below Figures 1B and 1C of Larkin et al and write down the
> # "number at risk" data given for various time points for each line.
> #
> # The list "numbers.below.figure", shown below, contains these data.
> #
> ################################
>
> numbers.below.figure = list(
+                           pdl1.negative.nivo = c(208, 192, 178, 108, 105, 98, 88, 80,
+                                                  76, 74, 63, 50, 31, 24, 9, 5, 4, 2, 1, 1) ,
+                           pdl1.negative.ipi = c(202, 183, 166, 82, 72, 59, 44, 39, 35,
+                                                 31, 26, 22, 12, 8, 3, 1),
+                           pdl1.negative.combo = c(210, 195, 181, 142, 134, 123, 112, 106,
+                                                   105, 96, 88, 79, 42, 36, 13, 9, 6, 2, 1),
+                           pdl1.positive.nivo =  c(80, 76, 71, 57, 56, 54, 51, 49, 49, 43, 38,
+                                                   32, 16, 13, 5, 4, 2),
+                           pdl1.positive.ipi = c(75, 69, 66, 40, 33, 24, 22, 21, 21, 17, 16, 15,
+                                                 9, 6, 3, 2, 2),
+                           pdl1.positive.combo =  c(68, 63, 61, 53, 52, 47, 44, 42, 42, 39,
+                                                    34, 24, 16, 12, 3, 1, 1)
+                           )
>
```

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].

```
>
> ####################
> #
> # Now, specify how far along on the x-axis of each line we want to go to extract data.
> # For example, in one line we may want to go up to time 15 units,
> # in another line up to time 18 units etc.
> #
> # As above, organize these times (integer values) for each line in the same order
> # as the sheets in the excel file.
> #
> ####################
>
> time = list( time.pdl1.neg.nivo = 0:18,
+              time.pdl1.neg.ipi = 0:15,
+              time.pdl1.neg.combo = 0:18,
+              time.pdl1.pos.nivo = 0:17,
+              time.pdl1.pos.ipi = 0:17,
+              time.pdl1.pos.combo = 0:17
+            )
>
>
>
> ################
> #
> # arm indicator
> #
> # 1 = pdl1-neg-nivo
> # 2 = pdl1-neg-ipi
> # 3 = pdl1-neg-combo
> # 4 = pdl1-pos-nivo
> # 5 = pdl1-pos-ipi
> # 6 = pdl1-pos-combo
> #
> ################
>
```

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 "pdl1.neg.nivo"), the second file corresponds to negative programmed death 1 ligand expression receiving ipilimumab (denoted "pdl1-neg-ipi") etc.

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

```
>
>
> ##############
> #
> # The R functions preprocess.digitized.data (Program 1)
> # and Guyot.individual.data (Program 2) are given below.
> # Read them into R first. Then execute the commands below to get "individual.data".
> #
> ##############
>
> individual.data = NULL
> for(ifile in 1:length(digitized.file.names)){
+     digitized.line = read.table(digitized.file.names[ifile], header=T)
+     processed.line.data = preprocess.digitized.data(digitized.line,
+                                                     numbers.below.figure[[ifile]],
+                                                     time[[ifile]])
+     individual.line.data = Guyot.individual.data(processed.line.data$condensed.data.set,
+                                                  processed.line.data$nrisk.data,
+                                                  input.arm.id=ifile)
+     individual.data = rbind(individual.data, individual.line.data)
+ }
>
>
>
```

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.

```
>
> individual.data[1:20,]
       time event tmt.arm.number
 [1,] 0.678     1              1
 [2,] 0.678     1              1
 [3,] 0.678     1              1
 [4,] 0.678     1              1
 [5,] 0.678     1              1
 [6,] 0.905     1              1
 [7,] 0.910     1              1
 [8,] 0.939     1              1
 [9,] 1.140     1              1
[10,] 1.140     1              1
[11,] 1.140     1              1
[12,] 1.390     1              1
[13,] 1.410     1              1
[14,] 1.410     1              1
[15,] 1.410     1              1
[16,] 1.610     1              1
[17,] 1.660     1              1
[18,] 1.680     1              1
[19,] 1.740     1              1
[20,] 1.910     1              1
>
>
```

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.

```
>
>
> treatment.type = c(
+                    rep("nivolumab", length(which(individual.data[,"tmt.arm.number"] == 1))),
+                    rep("ipilimumab", length(which(individual.data[,"tmt.arm.number"] == 2))),
+                    rep("combination", length(which(individual.data[,"tmt.arm.number"] == 3))),
+                    rep("nivolumab", length(which(individual.data[,"tmt.arm.number"] == 4))),
+                    rep("ipilimumab", length(which(individual.data[,"tmt.arm.number"] == 5))),
+                    rep("combination", length(which(individual.data[,"tmt.arm.number"] == 6))))
>
> pdl1.status = c(rep("negative", length(which(individual.data[,"tmt.arm.number"] < 4))),
+                 rep("positive", length(which(individual.data[,"tmt.arm.number"] >= 4))))
>
> individual.data = as.data.frame(individual.data)
> individual.data$treatment.type = treatment.type
> individual.data$pdl1.status = pdl1.status
>
>
```

Fig. 13. R commands to convert the treatment arm indicator numbers 1, 2, 3, 4, 5, 6 to treatment names ("nivolumab", "ipilimumab" and "combination") and programmed death 1 ligand status ("negative" and "positive"), and to append columns for treatment names and expression status to the patient-level data object "individual.data".

```
>
>
> individual.data[1:20,]
    time event tmt.arm.number treatment.type pdl1.status
1  0.678     1              1      nivolumab    negative
2  0.678     1              1      nivolumab    negative
3  0.678     1              1      nivolumab    negative
4  0.678     1              1      nivolumab    negative
5  0.678     1              1      nivolumab    negative
6  0.905     1              1      nivolumab    negative
7  0.910     1              1      nivolumab    negative
8  0.939     1              1      nivolumab    negative
9  1.140     1              1      nivolumab    negative
10 1.140     1              1      nivolumab    negative
11 1.140     1              1      nivolumab    negative
12 1.390     1              1      nivolumab    negative
13 1.410     1              1      nivolumab    negative
14 1.410     1              1      nivolumab    negative
15 1.410     1              1      nivolumab    negative
16 1.610     1              1      nivolumab    negative
17 1.660     1              1      nivolumab    negative
18 1.680     1              1      nivolumab    negative
19 1.740     1              1      nivolumab    negative
20 1.910     1              1      nivolumab    negative
>
>
```

Fig. 14. R output showing reconstructed patient-level data for the first 20 patients. The first 3 columns are the same as in Fig. 12. Columns 4 and 5 are the newly appended data on treatment and programmed death 1 ligand expression status using the commands shown in Fig. 13. The data for all 843 patients are given in 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. Sznol, 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.