Survival analysis of two versions of a database

2016-07-27 16:03:45 -0300

Abstract:

Diffuse intrinsic pontine tumors (DIPG) are the brain tumors that have the highest mortality in the age group of children and adolescents (0-18 years). Children with a median age of 7 years are the most affected, with no difference of gender. There is no effective curative treatment. Our research group actively studies this disease. We have been mounting cloud-based database of DIPG patients since 2009. Two different versions at different times, giving rise to two sets of data, like two different snapshots, each showing the aspect of the database at a particular time were deposited in the repository of this project. We wanted to compare these two snapshots through survival analysis. That is, we have done the same retrospective analysis in two different versions of the database. We demonstrated an important difference in the retrospective analysis results just because we used two different snapshots of the same dataset. Because all files can be publicly accessed, the scientific information is thus completely reproducible by third parties. This retrospective analysis is part of a clinical trial project that aims to be an open clinical research model.

Introduction:

Diffuse intrinsic pontine tumors (DIPG) are the brain tumors that have the highest mortality in the age group of children and adolescents (0-18 years). Children with a median age of 7 years are the most affected, with no difference of gender. [1] There is no effective curative treatment. Most of these gliomas arise in the ventral region of the pons and their growth pattern is infiltrative, affecting most or virtually all of the cross diameter of the pons. As a rule, they does not form a contrast-enhanced tumor mass and have no precise limits on magnetic resonance imaging. [2]

Our research group actively studies this disease. The information we have on patients with DIPG come from a database that we have been mounting since 2009. This database is currently fully digital and based on a cloud server. Several versions of it already existed more or less well documented over time. Two different versions at different times, giving rise to two sets of data were pub-

lished in the repository of this project. They are like two different snapshots, each showing the aspect of the database at a particular time. This post aims to compare these two snapshots through survival analysis. That is, we will do the same retrospective analysis in two different versions of the database.

The first version tries to recapitulate the exact moment of the data set when the VALKYRIE project was written in 2014. Unfortunately, this exact version of the database has not been saved, only the immediatelly next version. So, I used the version that had been stored in 2014, making a single modification (including patient 56). The data was de-identified as already described before [3] and the database was saved in a csv file [4] stored in the project repository. The file includes all patients diagnosed with DIPG between 2000-2013, regardless of the treatment.

The second version is the 2015 database (snapshots are always written to the end of the year) and includes patients diagnosed with brainstem tumors between 2000-2015, diffuse or focal, treated or not. Also, only one change was made (**topography** label modification for patient 64 from *focal pons* to *DIPG*). It is also stored on the same server in .csv format. [4]

Survival analysis used to design VALKYRIE [5] was a retrospective analysis in which we compared the outcome (overall survival) of patients with DIPG diagnosed between 2000-2013 treated with at least 40 Gy of RT and two different types of chemotherapy (chemo): treated according to a HIT GPOH protocol [6] _versus_ those treated with other chemotherapy protocols (or without chemo). It is this analysis to be repeated in the two versions of the database.

Methodology:

The databases are stored in a Github repository publicly accessible. They contain no elements that could identify patients, although individual patient data is depicted (they were de-identified [7]). The information was remotely captured using the package *RCurl* [8] and R [9] programming language. Using the package *knitr* [10], a script written in the R language was executed to perform survival analysis with the package *survival* [11]. The calculations performed were: survival curve by computing the Kaplan-Meier (KM) estimate [12], a log-rank or Mantel-Haenszel test for comparison between two survival curves obtained by the KM method [13], and calculation of the follow-up time (from the survivors, when there were any). Only patients that received at least 40 Gy of radiation therapy (RT) were included in both versions of the database. The comparisons were made between patients treated according to HIT protocols [6], and patients treated in accordance to COG protocols (or RT only treated) [14]. The script was written in a .Rmd file [15] stored on Github, which was, at one time, parsed and deployed in Markdown format [16] using the package *servr* [17], and 3.2.1 Jekyll program [18]. The .md file was saved in a local repository and then commited to GitHub Pages, becoming available on the internet. All files

can be publicly accessed. The current scientific information is thus completely reproducible by third parties.

Results and discussion:

The tabulated data (Tables 1 and 2) show only DIPG patients treated with RT from both versions of the database. The difference between the two versions is a larger number of cases in the second version (6 more cases) and a higher follow-up time (greater impact in the cases treated with HIT protocols because they were accrued later). Although the follow-up time tends to not affect the Kaplan-Meier estimate, in this case there was a significant difference between the versions. This greatly changed the results of the survival analysis, as discussed below.

Table 1: patients with DIPG, diagnosed 2000-2013, treated with RT (minimum 40 Gy)

1-B.Treated as per HIT protocols

	sex	age	os	status	ecog	chemo	tt.chemo	cycles1	rt	tt.rt	ttp	os2	vpa
43	1	6.9	15.8	0	1	10	5	NA	5400	32	5.0	330	1
44	1	5.7	5.2	1	1	10	34	NA	5400	30	NA	NA	1
45	1	8.3	12.9	0	1	10	22	NA	5400	28	12.1	26	1
48	0	7.3	11.4	0	3	10	235	NA	5400	226	NA	NA	1
50	0	3.5	6.7	0	2	10	100	NA	5400	19	6.7	2	1
51	1	3.7	5.6	0	2	10	68	NA	5400	22	5.5	2	0
52	0	6.0	4.0	0	4	10	NA	NA	5400	76	NA	NA	1
53	0	9.4	3.3	0	3	10	NA	NA	5400	52	NA	NA	1
54	1	5.7	1.4	0	3	10	19	NA	5400	NA	NA	NA	1
55	0	5.7	0.6	0	1	10	23	NA	5400	NA	NA	NA	1
56	1	4.0	25.0	0	0	10	71	NA	5400	36	NA	NA	0

Table 2: patients with DIPG, diagnosed 2000-2015, treated with RT (minimum 40 Gy)

2-B.Treated as per HIT protocols

	sex	age	os	status	ecog	lpps	chemo	tt.chemo	cycles	delta.chemo	rt	ttp	vpa
48	1	4.4	12.1	1	3	NA	10	1	10	113	5400	170	1
63	1	6.9	20.6	1	1	70	10	5	NA	147	5400	152	1
64	1	4.0	25.0	0	0	100	10	71	NA	NA	5400	NA	1
66	1	8.3	15.7	1	1	80	10	22	NA	342	5400	368	1
69	0	7.3	15.4	1	3	40	10	235	NA	182	5400	NA	1
71	0	3.5	7.2	1	2	0	10	100	NA	NA	5400	203	1
72	1	3.7	12.4	1	2	60	10	68	NA	NA	5400	168	0

	sex	age	os	status	ecog	lpps	chemo	tt.chemo	cycles	delta.chemo	rt	ttp	vpa
73	0	6.0	11.6	1	4	20	10	150	NA	NA	5400	NA	1
74	0	9.4	9.2	1	3	40	10	122	NA	NA	5400	262	1
76	1	5.7	16.5	1	3	40	10	19	NA	275	5400	304	1
77	0	5.7	8.8	1	1	80	10	23	NA	199	5400	199	1
79	0	5.6	15.2	1	1	80	10	26	NA	226	5400	236	0

Survival analysis:

Both versions of the database shows a larger group of patients who were treated with RT alone or chemotherapy regimens according to COG protocols. In the second version, this group has 5 more patients (20% increase). Summary of overall survival, in addition to survival in 12 months, illustrating that these variables have not really changed:

`First dataset version - RT alone/ RT+COG`

records	n.max	n.start	events	median	0.95LCL	0.95UCL
31	31	31	29	11.4	8.1	15.6

```
## Call: survfit(formula = Surv(os, status) ~ 1, data = dipg, subset = (rt >
##     4000 & chemo != 10))
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    12     13      17    0.447    0.09        0.302        0.664
```

`Second dataset version - RT alone/ RT+COG`

records	n.max	n.start	events	median	0.95LCL	0.95UCL
36	36	36	31	11.9	9.4	16.8

```
## Call: survfit(formula = Surv(os, status) ~ 1, data = brainstem, subset = (topo ==
##     "DIPG" & rt > 4000 & chemo != 10))
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    12     15      17    0.493  0.0872        0.349        0.698
```

Follow-up time has also remained essentially the same. Summary of follow-up time (classical method of calculating from the survivors):

`First dataset version - RT alone/ RT+COG`

- 29 observations deleted due to missingness

records	n.max	n.start	events	median	0.95LCL	0.95UCL
2	2	2	2	12.1	10.4	NA

`Second dataset version - RT alone/ RT+COG`

- 31 observations deleted due to missingness

records	n.max	n.start	events	median	0.95LCL	0.95UCL
5	5	5	5	9.7	6.8	NA

As can be seen in Tables 1 and 2, a considerable number of patients in both versions of the database, received valproic acid (VPA) during treatment. Thus, the comparison between the different types of treatment with chemotherapy was biased, since it was also part of the design hypothesis that treatment with VPA may modify survival. To try to reduce this bias, it was necessary to redo the calculations, taking into account only patients who were NOT treated with VPA. Summary of survival variable, in addition to survival in 12 months:

`First dataset version - RT alone/ RT+COG, no VPA`

records	n.max	n.start	events	median	0.95LCL	0.95UCL
7	7	7	7	8	7.3	NA

```
## Call: survfit(formula = Surv(os, status) ~ 1, data = dipg, subset = (rt >
##     4000 & chemo != 10 & vpa == 0))
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    12      1       6    0.143   0.132       0.0233        0.877
```

`Second dataset version - RT alone/ RT+COG, no VPA`

records	n.max	n.start	events	median	0.95LCL	0.95UCL
12	12	12	7	9	7.4	NA

```
## Call: survfit(formula = Surv(os, status) ~ 1, data = brainstem, subset = (topo ==
##     "DIPG" & rt > 4000 & chemo != 10 & vpa == 0))
##
```

```
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    12      2       6    0.404   0.155         0.19        0.859
```

Follow-up time:

`First dataset version - RT alone/ RT+COG, no VPA`

records	n.max	n.start	events	median	0.95LCL	0.95UCL
7	7	7	7	8	7.3	NA

Follow-up time from the first database version (patients treated with RT alone or *as per* COG), excluding those patients who received VPA is not different from its median overall survival since all included patients experienced the event (death) ie, there are no survivors in this subgroup.

`Second dataset version - RT alone/ RT+COG, no VPA`

- 7 observations deleted due to missingness

records	n.max	n.start	events	median	0.95LCL	0.95UCL
5	5	5	5	9.7	6.8	NA

In both versions of the database, the groups treated *as per* HIT chemotherapy protocols are smaller. However, there is an important difference regarding the follow-up time between the two versions. This reflects also in a substantial difference in the number of censored patients, which means, ultimately, a change in meaningful outcomes and, therefore, in the final results. Summary of overall survival, and survival in 12 months:

`First dataset version - RT+HIT`

records	n.max	n.start	events	median	0.95LCL	0.95UCL
11	11	11	1	NA	NA	NA

```
## Call: survfit(formula = Surv(os, status) ~ 1, data = dipg, subset = (rt >
##     4000 & chemo == 10))
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    12      3       1    0.857   0.132        0.633            1
```

`Second dataset version - RT+HIT`

records	n.max	n.start	events	median	0.95LCL	0.95UCL
12	12	12	11	13.8	11.6	NA

```
## Call: survfit(formula = Surv(os, status) ~ 1, data = brainstem, subset = (topo ==
##     "DIPG" & rt > 4000 & chemo == 10))
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    12      8       4    0.667   0.136        0.447        0.995
```

Difference in follow-up time (calculated by inverse KM function, since almost all patients were alive, in the first dataset version):

```
First dataset version - RT+HIT
```

records	n.max	n.start	events	median	0.95LCL	0.95UCL
11	11	11	10	6.7	4	NA

```
Second dataset version - RT+HIT
```

records	n.max	n.start	events	median	0.95LCL	0.95UCL
12	12	12	11	13.8	11.6	NA

At this point, it is possible to conclude that despite the fact that follow-up time usually do not interfere in the estimation of the Kaplan-Meier function [19], what really happened here was a great change in the meaningful outcomes. While in the first version of the database nearly all patients treated *as per* HIT were censored and almost no study subject experienced the event (death), the opposite occurred in the second version of the database. In this second version, with a follow-up time increase of just over 6 months most of the patients had experienced the event and there was hardly any censored time. This illustrates the importance of ensuring that the patient follow-up time is at least clinically significant, consistent with the natural progression of the disease studied [20].

As a final bonus, from the second database version it is possible to carry out a new analysis of retrospective survival, this time comparing patients treated with chemotherapy according to HIT (associated with VPA) with patients treated *as per* COG (or RT alone) and rigorously WITHOUT the use of VPA. The results are as follows:

```
First dataset version - RT/RT+COG versus RT+HIT
```

Table 15: Call: Surv(os, status) ~ ifelse(chemo == 10, 0, 1) Chisq = 3.931154 on 1 degrees of freedom, p = 0.047399

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
ifelse(chemo == 10, 0, 1)=0	11	1	5.004	3.204	3.931
ifelse(chemo == 10, 0, 1)=1	31	29	25	0.6413	3.931

`Second dataset version - RT/RT+COG versus RT+HIT`

Table 16: Call: Surv(os, status) ~ ifelse(chemo == 10, 0, 1) Chisq = 0.084035 on 1 degrees of freedom, p = 0.771903

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
ifelse(chemo == 10, 0, 1)=0	12	11	11.83	0.058	0.08403
ifelse(chemo == 10, 0, 1)=1	36	31	30.17	0.02274	0.08403

`Second dataset version - RT/RT+COG (no VPA) versus RT+HIT`

Table 17: Call: Surv(os, status) ~ g Chisq = 0.499920 on 1 degrees of freedom, p = 0.479535

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
g=1	12	7	5.648	0.3236	0.4999
g=2	12	11	12.35	0.148	0.4999

A graph comparing groups g1 and g2, each one with 12 patients:

```
## Call: survfit(formula = Surv(os, status) ~ g, data = brainstem, subset = (topo ==
##     "DIPG" & rt > 4000))
##
##    24 observations deleted due to missingness
##      n events median 0.95LCL 0.95UCL
## g=1 12      7    9.0     7.4      NA
## g=2 12     11   13.8    11.6      NA
```

It is quite evident that there seems to be no difference between the curves. This would indicate that the treatment proposed in our trial design does not correlate with survival enhancement, in particular the use of valproic acid seems

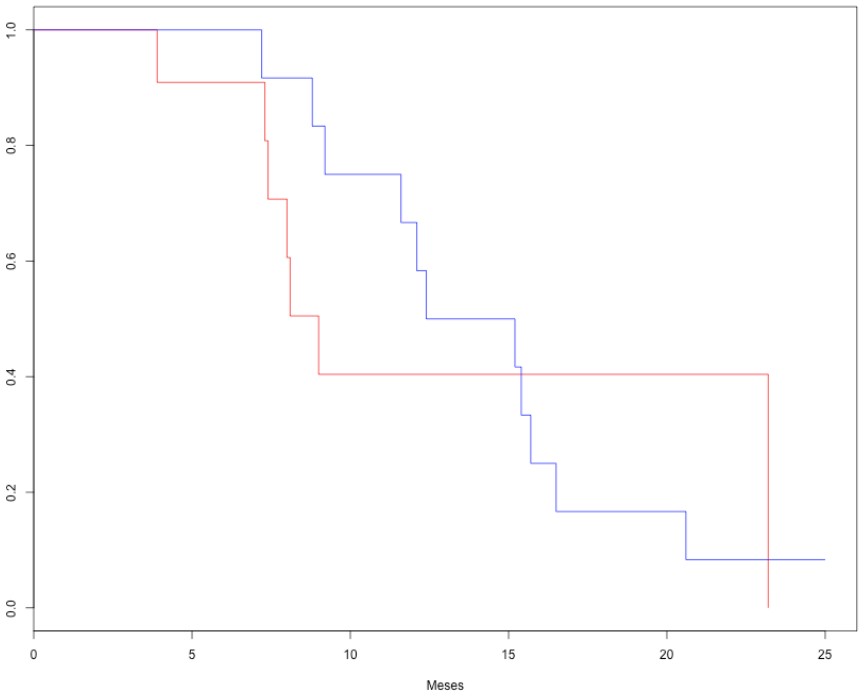

Figure 1: Patients with DIPG - overall survival

to have no effect. Guarding the limitations of retrospective analysis and the small number of patients in this comparison, it is a possible conclusion.

References:

1. Felix FHC. Diagnostic criteria for DIPG (2016). Figshare, doi:10.6084/m9.figshare.3489941.v1
2. Felix FHC. A little about the trial's rational (2016). Figshare, doi:10.6084/m9.figshare.3489917.v1
3. Felix FHC. Historical control group (2016). Figshare, doi:10.6084/m9.figshare.3489956.v2
4. Felix FHC. Banco de dados de pacientes portadores de tumor pontino intrínseco difuso (2016). Repositório do Github, disponível em: https://github.com/fhcflx/valkyrie/tree/master/data
5. Felix FHC. Estimate of the number of patients (2016). Figshare, doi:10.6084/m9.figshare.3489947.v1
6. Felix F and Fontenele J. Chemoradiotherapy with etoposide, cisplatin, and ifosfamide associated with valproic acid for patients with diffuse intrinsic pontine glioma [v1; not peer reviewed]. F1000Research 2015, 4:1301 (poster) [Portuguese] (doi:10.7490/f1000research.1111018.1)
7. Felix FHC. De-identification - anonymization (2016). Figshare, doi:10.6084/m9.figshare.3545471.v1
8. Duncan Temple Lang, the CRAN team (2016). RCurl: General Network (HTTP/FTP/...) Client Interface for R, https://CRAN.R-project.org/package=RCurl
9. R Core Team. R: A Language and Environment for Statistical Computing, 2016, https://www.R-project.org/
10. Yihui Xie (2016). knitr: A General-Purpose Package for Dynamic Report Generation in R, http://yihui.name/knitr/
11. Terry M Therneau (2015). A Package for Survival Analysis in S, http://CRAN.R-project.org/package=survival
12. Kablfleisch, J. D. and Prentice, R. L. (1980). The Statistical Analysis of Failure Time Data. New York:Wiley.
13. Harrington, D. P. and Fleming, T. R. (1982). A class of rank test procedures for censored survival data. Biometrika 69, 553-566.
14. Felix FH, de Araujo OL, da Trindade KM, Trompieri NM, Fontenele JB (2014). Retrospective evaluation of the outcomes of children with diffuse intrinsic pontine glioma treated with radiochemotherapy and valproic acid in a single center. J Neurooncol. 116(2):261-6. doi:10.1007/s11060-013-1280-6.
15. Felix FHC (2016). Análise de sobrevida em duas versões do mesmo banco de dados. Arquivo .Rmd armazenado em repositório do GitHub. https://github.com/fhcflx/valkyrie/blob/gh-pages/_source/2016-07-27-Análise-de-sobrevida-em-duas-versões-do-mesmo-banco-de-dados.Rmd
16. John Gruber (2004). Markdown, © 2002–2016 The Daring Fireball Company LLC.

17. Yihui Xie (2016). servr: A Simple HTTP Server to Serve Static Files or Dynamic Documents, https://CRAN.R-project.org/package=servr
18. GitHub (2016). Jekyll, a blog-aware, static site generator in Ruby. @ MIT License.
19. Shuster J. (1991). Median follow up in clinical trials. J. Clin. Oncol., 9, 191-192.
20. Altman DG, De Stavola BL, Love SB, Stepniewska KA (1995). Review of survival analyses published in cancer journals. Br J Cancer 72:511–518