ICM Collaboration Notes

Patric Fulop & Alex Agachi

The University of Edinburgh

November 8, 2017

1 Introduction

Identify potential, explain problem, literature review, explain briefly what you're predicting in light of lit review. Then conclude by using different benchmarks. Obviously explain your model at some point.

2 Data statistics

There are two main datasets, one with biological information and one with clinical data. We give a brief description of the merged dataset before preprocessing:

- **Key 1:** Patient ID - this is not unique across rows
- **Key 2:** Surgery date and clinical surgery date. These are sometimes off by one day so we took only surgery dates as being relevant.

There are a total of **7825** entries and **6688** unique patients. Each patient has **30** relevant attributes. For convenience, the attribute names have been renamed more intuitively, and in English :).

Some of the attributes have missing values.

1. Diagnostic dates are there only for one fifth of the patients, **1162**.
2. Date of birth (DoB) is missing for **1002** patients.
3. Date of death is missing for **4908** entries, should we assume these are survivors?
4. Gene data is very sparse, i.e. **Ch** markers.
5. Gender data has **332** entries missing.

2.1 Dealing with Missing data

Examples.[LR14]
Clearly some patients underwent some tests, while others did not. This is a problem we can deal with in a very robust manner as long as we can assume that the data is missing at random (i.e. the mechanism by which it is missing can be described as random, and does not contain relevant information in itself. For example whether a test was conducted or not for a patient does not say something highly relevant about that patient?s condition/survival expectation in itself.) Even if the data is not missing at random, similar techniques would be applied by default of statistics having invented better ones to date, but it would help to understand better the missing data reasons/mechanisms for our variables, to make sure we describe it properly.[GLAAA15]

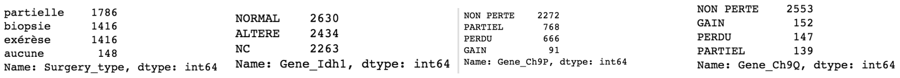
Attribute	Present	Missing	Encoding	Type
Age at surgery	to see	to see	Age	Numerical discrete
Gender	7493	332	Gender	Binary
Histo Grade	7825	0	Tumor Grade	Categorical (4)
Histo Type	7825	0	Tumor Type	Categorical
KPS	?	?	?	?
Outcome	4766	3059	Surgery Type	Categorical (3)
Radiotherapy	2722	5103	Rx Date	Time → Ultimately Binary
Chemotherapy	2950	4875	Chemo Date	Time → Ultimately Binary
IDH Mutation 1	7327	498	Gene IDH1	Categorical (3)
IDH Mutation 2	7078	747	Gene IDH2	Categorical (3)
Htert C228T	4336	3489	Gene C228T	Categorical (3)
Htert C250T	4333	3492	Gene C250T	Categorical (3)

Table 1: Present and missing variables and their encoding

3 Encoding clarifications and target variables

As previously discussed, in the first phase we are interested in a smaller subset of attributes. Table 1 above indicates some of the variables of interest. Please let us know if we got the right ones and whether we should add more from the dataset. For some of them, some things remain unclear.

1. We aim to add age at surgery as one variable, taking into account surgery date and date of birth.
2. We do not have any attribute for KPS (performance status score) as far as we know.
3. The outcome is encoded in the surgery type variable 1a. It is either a type of surgical removal or biopsy. For this variable, does missing data tell us that there was no surgery or that we do not know the outcome? We can see **aucune** as a type so I would assume that we do not know the outcome.
4. For radiotherapy and chemotherapy, should we assume that if the patient does not have a date, he did not undergo that treatment?
5. For IDH mutations 1b, IDH1 and IDH2 seem to predominate there. Are these the two main ones we are interested in? You mentioned IDH wild type which, I assume is the case for non-mutated IDH gene, so I would just say this is the **NORMAL** value of IDH1/IDH2 Gene. Is this correct? Furthermore, what does the value **NC** stand for?
6. In terms of genetic tests, is there any equivalence between the following coding schemes for various genes, i.e. can we treat **NORMAL** or **ALTERE** as carrying the same meaning across these schemes? i.e. 1b, 1c and 1d



```
partielle     1786
biopsie       1416
exérèse       1416
aucune         148
Name: Surgery_type, dtype: int64
```

(a) Surgery Type

```
NORMAL     2630
ALTERE     2434
NC         2263
Name: Gene_Idh1, dtype: int64
```

(b) Gene Idh1

```
NON PERTE    2272
PARTIEL       768
PERDU         666
GAIN           91
Name: Gene_Ch9P, dtype: int64
```

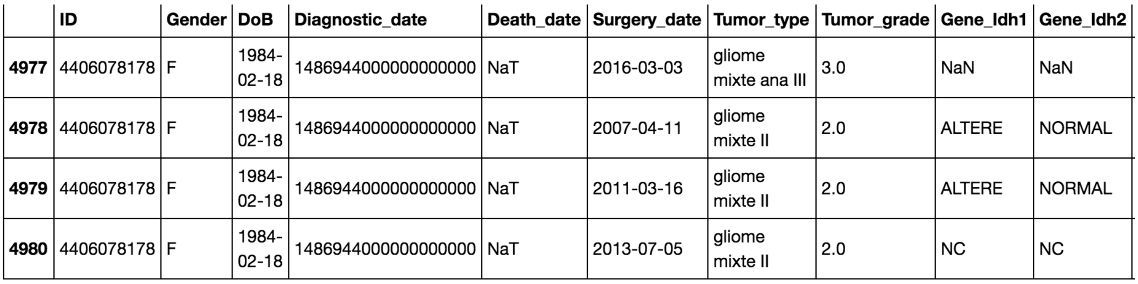
(c) Gene Ch9P

```
NON PERTE    2553
GAIN          152
PERDU         147
PARTIEL       139
Name: Gene_Ch9Q, dtype: int64
```

(d) Gene Ch9Q

Figure 1: Outcome and Gene Mutations

Multiple entries: On average, patients have more than 1 entry (see fig. 2) according to how many surgeries they went through. Is that correct and do you have any pointers as to how to treat patients with multiple surgeries? Should we aggregate them or treat them independently? Or perhaps find some other clever way of dealing with that.



	ID	Gender	DoB	Diagnostic_date	Death_date	Surgery_date	Tumor_type	Tumor_grade	Gene_Idh1	Gene_Idh2
4977	4406078178	F	1984-02-18	1486944000000000000	NaT	2016-03-03	gliome mixte ana III	3.0	NaN	NaN
4978	4406078178	F	1984-02-18	1486944000000000000	NaT	2007-04-11	gliome mixte II	2.0	ALTERE	NORMAL
4979	4406078178	F	1984-02-18	1486944000000000000	NaT	2011-03-16	gliome mixte II	2.0	ALTERE	NORMAL
4980	4406078178	F	1984-02-18	1486944000000000000	NaT	2013-07-05	gliome mixte II	2.0	NC	NC

Figure 2: Entries for patient 4406078178

Target Variable: We want to confirm our idea of selecting the target variable for our models. First of all, can we assume that all patients where there is no death date specified, are still alive? (as opposed to them not being alive anymore, but this record missing) Qualitatively we thought of incorporating the surgery date, cancer detection date and death date.

- The easiest target to model is a binary variable representing alive/dead.
- The time between diagnostic date and death date. We would scale this variable accordingly to account for no death.
- The time between the first/last surgery and death time. We could incorporate in this the number of surgeries a person had.

4 Literature Review

Find benchmarks and comparison situations for our pipeline and results, i.e. 'Cancer survivability'.

5 Miscellaneous

Add whatever crosses your mind

References

[GLAAA15] Pedro J. García-Laencina, Pedro Henriques Abreu, Miguel Henriques Abreu, and Noémia Afonoso. Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values. *Computers in Biology and Medicine*, 59:125–133, apr 2015.

[LR14] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2014.