ICM Collaboration Notes

Patric Fulop & Alex Agachi

The University of Edinburgh

February 20, 2018

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.

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** - meaningless, not reliable
2. Date of birth (DoB) is missing for **1002** patients. Figure out first whether there is a relation between DOB and other missing variables. Take them out one by one starting with the IDH, see below.
3. Date of death is missing for **4908** entries, should we assume these are survivors? - no the data might be just missing because the patient didn't follow up.
4. Gene data is very sparse, i.e. **Ch** markers. No rule yet, we need to figure out how to deal with sparsity - not much to do, will imputation work? Should we even consider it?
5. Gender data has **332** entries missing. DELETE.
6. IDH1 and IDH2 is very important, if the study wasn't done the record is kind of pointless, so delete all those guys that don't have it.

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.

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

2. Merge KPS from Marc's email (performance status score)

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? Does **aucune** mean that no surgical act whatsoever was undertaken?

4. For radiotherapy and chemotherapy, should we assume that if the patient does not have a date, he did not undergo that treatment, or is this data missing instead? If missing CHX and RDX and gene data is sparse then the record is poor

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/mutated, so we concluded the **NORMAL** value of IDH1/IDH2 Gene stands for wild-type. Is this correct? In this context, what does the value **NC** stand for? It's a valid result and the result means non-conclusive - we don't know.

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/genes? i.e. 1b, 1c and 1d. No, just use it the same as in the table for different genes.

```
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. **Use the first date surgery date after you order it by dates. DO THIS FIRST**

	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, but we would much prefer if you would confirm.

- The time between surgery date and death date. We would scale this variable accordingly to account for no death. (**USE THIS ONE**)
- The easiest target to model is a binary variable representing alive/dead.
- 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

We conducted a more focused meta study in the past, of studies tailored to datasets and target variables such as ours. We broke down the study among the important features: datasets used, machine learning models used, validation mechanisms and feature engineering. Following this review, a possible first study on our dataset would include refining the following models:

- Clustering classifiers such as Gaussian Mixture models (GMM)
- Support Vector Machines (SVM)
- Decision Trees
- Neural Networks

We would then compare the results across all four criteria commonly used: accuracy, sensitivity, specificity, ROC curve and area under the curve (AUC). A similar development was tackled by Garcia-Laencina et. al. (2015) [GLAAA15], in which they focused on breast cancer.

5 Dealing with Missing data

Traditional techniques of dealing with missing data in medical studies, available case analysis, pairwise deletion, and single imputation are suboptimal and lead to either/both of loss of statistical power and increased bias in the results [BFHH14], [LDC+12]. As such, dealing with missing data in our study in a statistically robust manner has been a key area of focus in our data preparation to date.

Clearly some patients in this dataset 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. The missing data mechanism relates to why values are missing and the connection of those reasons with treatment outcomes [GLAAA15]. Can we assume that these are random in our study?

More concretely, for **C228T** we have **3489** missing observations in our dataset. Can we assume that these observations are missing in a way that is not reflective of the patients condition and outcome i.e. doctors simply did not order this test, or this test was not available at a specific time, all independent of the patient?s condition. Or if the data is not missing at random - for example the patients for whom we do not have the results to this test were significantly more ill-suffering from a more aggressive form of cancer than the ones for whom we do have the results? It would help to understand the mechanisms better, to make sure our methods are sound. We narrowed down here on a technique called multiple imputation (MI) by chained equations [GOG07], [Gra09]. This type of imputation method involves replacing missing values by suitable estimates and then applying standard complete-data methods to the filled-in data. The basics ideas of MI as proposed by Rubin [Rub96] involves the following three steps:

1. Imputation: Impute missing values using an appropriate model that incorporates appropriate random variation. During this first step, sets of plausible values for missing observations are created that reflect uncertainty about the non-response model. These sets of plausible values can then be used M times to 'complete' the missing values and create M 'completed' data sets.

2. Analysis: Perform the desired analysis on each of these M data sets using standard complete-data methods.
3. Combination: During this final step, the results are combined, which allows the uncertainty regarding the imputation to be taken into account.
4. Keep track of the uncertainty when we impute and do not bias the study too much.
5. Predictive mean matching - some regression

5.1 Ideas

- Key Aspect: Figure out exactly how many unique patients we have so we can work with that dataset first. This is a very tricky aspect, it can break down everything later on.
- Expectation maximization imputation. I guess we could use any density estimation technique for generating datapoints. Apparently KNN imputation did best for the [GLAAA15] paper. We should start with that approach perhaps. Their target variable is binary, we can have different models here.
- Given the dataset we have, perform PCA to see whether you can remove some attributes.
- Three missing data mechanisms a la Rubin [LDC+12]. I think in our case should be MAR (missing at random) as well. Find out what the percentage is for missing data for each attribute and if this is below a critical threshold in comparison with the sample size. Q: What is the right threshold? See *fraction of missing information* λ described by Rubin

5.2 Feature engineering

Feature engineering (changing variables for example in terms of scale, or creating interaction variables?) is a very large part of building more accurate models. We already thought of adding extra variables based on the current ones, i.e. the time from diagnostic date to surgery (to check the effects of patients delaying significantly recommended surgeries) and the age at diagnostic. Are there other obvious ones you suspect?

References

[BFHH14] Melanie L Bell, Mallorie Fiero, Nicholas J Horton, and Chiu-Hsieh Hsu. Handling missing data in RCTs; a review of the top medical journals. *BMC medical research methodology*, 14:118, nov 2014.

[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.

[GOG07] John W. Graham, Allison E. Olchowski, and Tamika D. Gilreath. How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory. *Prevention Science*, 8(3):206–213, aug 2007.

[Gra09] John W. Graham. Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology*, 60(1):549–576, 2009.

[LDC+12] Roderick J. Little, Ralph D'Agostino, Michael L. Cohen, Kay Dickersin, Scott S. Emerson, John T. Farrar, Constantine Frangakis, Joseph W. Hogan, Geert Molenberghs, Susan A. Murphy, James D. Neaton, Andrea Rotnitzky, Daniel Scharfstein, Weichung J. Shih, Jay P. Siegel, and Hal Stern. The Prevention and Treatment of Missing Data in Clinical Trials. *New England Journal of Medicine*, 367(14):1355–1360, oct 2012.

[Rub96] Donald B. Rubin. Multiple Imputation After 18+ Years. *Journal of the American Statistical Association*, 91(434):473, jun 1996.