15.095 Project report
Identifying most responsive patients to a fibrinogen injection: a prescriptive approach to treat traumatic hemorrhagic schocks

Jocelyn Beauchesne and Alexandre Saillard

December 2018

1 Problem statement and importance

Fibrinogen is a widely used molecule to treat patients suffering from a traumatic hemorrhagic shock (HS); indeed, it is the precursor of fibrin paramount for the constitution of a robust clot. As a matter of fact, French and European guidelines have recommended to target a fibrinogen concentration of 1.5-2g/L in patients' blood.

However, despite this strong physiological rationale there is a lack of scientific evidence of its positive impact on all-cause mortality. This project builds on a previous statistical study which concluded that 'fibrinogen administration within the first 6 hours of a traumatic hemorrhagic shock did not decrease all-cause day-one mortality'. The confidence interval for the computed average treatment effect was [-8.3%, 2.1%] and thus inconclusive.

For this reason, our goal was to identify a most responsive sub-population to fibrinogen injection in order to overcome this past inconclusive study. We leverage Optimal Prescriptive Trees [3] to prescribe fibrinogen to patients and thus distinguished them based on their prescription value. We further robustified this approach by training 10K prescriptive trees which in turn gave us more granularity.

Although still preliminary, our findings could be of a substantial impact for the medical community. It could help detect sub-optimal medical strategies without implementing costly, time-consuming and uncertain clinical trials.

However, preliminary results of this study should be tempered by the presence of several human bias within medical data when it is not acquired through randomization. Future work could help assess the robustness of our pipeline to these bias.

2 Data

2.1 Source

"Created in 2012, the French Traumabase® Group is a collaboration focusing on Major Trauma. The Group's objectives are to improve Major Trauma care, to inform public health decisions and to facilitate research." [4]

The complete dataset is made out of 14336 patients which were admitted in one of the participating trauma centres. Out of those 14336, 1027 were selected for the causal inference analysis based on medical criteria.

Furthermore, the covariates used in both the previous study and this one, have been selected by surveying a Delphi of 16 European experts. The question asked was which variables are important for a clinician to

decide on injecting fibrinogen, as well as which are the most predictive of early mortality. Variable were selected if at least 30% of the experts agreed on its relevance. In total, there were 22 selected variables including:

- Treatment W: whether the patient was administrated fibrinogen or not in the first 6 hours. $W \in \{0, 1\}$ and takes value 1 if treatment was administred, 0 otherwise.
- Outcome Y: all-cause day-one mortality. $Y \in \{0, 1\}$ and takes value 1 if the patient deceased, 0 otherwise.

Out of 1027 patients, 74% received a fibrinogen injection and 42% of patients died during the process.

2.2 Pre-processing

Data points examples can be seen in table 1. It is worth noting that there is a significant amount of missing values. Thus, a missing value imputation is necessary before any kind of modelling.

Row	**CPA**	...	**Fibrinogen**	**DC**	**Treatment**
1	Missing	...	1.5	1	1
2	1	...	0.9	0	0

Table 1: Data points examples

As our dataset is a mix of categorical and quantitative variables, we used the function `imputeFAMD` from package missMDA [5] in `R` [6] to impute our missing values. This was also the package used in the previous study, making our results even more comparable. However, as this imputation is paramount to the rest of the analysis, part of future work could be to compare this to other methods, such as MICE [8].

3 Methods

3.1 Average treatment effect

The average treatment effect (ATE) evaluates the effect of a treatment W (fibrinogen injection in our case) on an outcome Y (patient deceased or not) given some covariates X (medical measurements):

$$ATE = E[Y_i(1) - Y_i(0)]$$

However, in our case, counter-factual $Y_i(1)$ and $Y_i(0)$ are unknown. The idea behind causal inference is to construct and compute unbiased estimators to infer the value of the ATE. In the previous study and this one, we used the package `GRF` [7] in `R` [6] that is an implementation of [1]. While one might question the validity of such a statistical approach, it is widely used in the medical community and thus a tool to communicate with doctors.

Rather than computing an ATE on the whole dataset, we hope to find with our prescriptive approach a sub-population on which the ATE yields significant results. Let us recall to the reader that we expect a negative ATE on a responsive sub-population since the "positive" outcome Y has value 0 and $Y \in \{0, 1\}$. Responsiveness being defined as an negative ATE, we wished to cluster our dataset in a way which allows us to compute ATE. Therefore, this implied having balanced enough clusters in size, but also diverse enough in terms of treatment and outcome. More precisely, the prescription value will be associated to each patient and used for categorizing patients according to their level of prescription

3.2 A first clustering with 5 Optimal Prescriptive Trees

To meet the balancedness criterion, we started by fitting 5 prescriptive trees with various values of training hyperparameters i.e. minbucket, maximum depth, treatment minbucket and prescription factor. Playing with the training hyperparameters is the alternative we found since adding constraints directly in the OPT was not an option. We associated patients with their number of prescriptions in [0,5] and thus categorized them according to the level of prescription.

Prescriptive tree parameters	**1**	**2**	**3**	**4**	**5**
max depth	5	5	5	5	5
minbucket		10	10	60	40
treatment minbucket				10	10
prescription factor			0.8		0.8

Table 2: 5 prescriptive trees

3.3 Robustness of categories

One can question the stability of the 6 categories created in the previous subsection. Indeed, two sources of randomness might affect the outcome of this prescriptive approach: missing values imputation and randomness inherent to Optimal Prescriptive Trees. In order to robustify this first categorization, we built a pipeline as depicted on figure 1, leading to 10 000 prescriptive trees by means of 100 imputations and 20 prescriptive trees fitted on each of the imputed dataset. In this setting, each patient is then attributed a value between 0 and 10 000 corresponding to the number of different trees which prescribed fibrinogen.

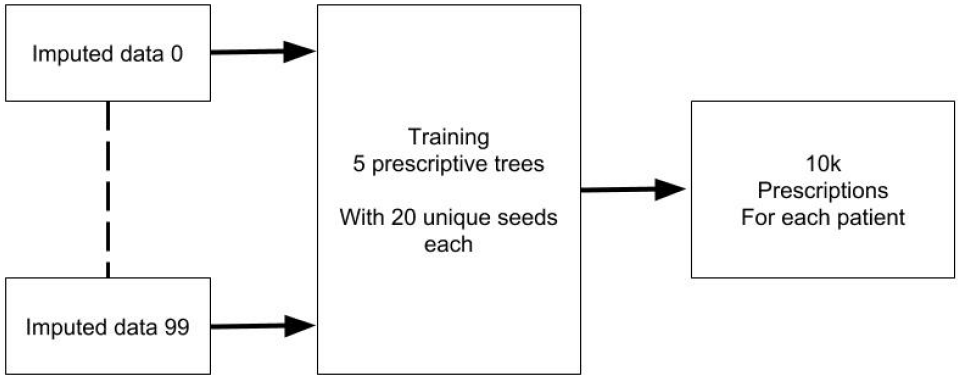

Figure 1: Our pipeline to ensure robustness of the analysis

4 Results

4.1 A robust categorization leading to three categories of patients

The first clustering using 5 different prescriptive trees led to 6 categories of patients according to the number of prescriptions. It is worth noting that all 5 trees agreed on prescribing fibrinogen to a group of 69 patients with rather interesting characteristics. Indeed, these patients present a low average value of treatment i.e. $\approx 40\%$ against $\approx 80\%$ for the remaining groups. Moreover, they show a significantly greater mortality i.e. $\approx 80\%$ against $\approx 20\%$.

This first categorization also highlights a small group of patients with 4 trees agreeing to prescribe fibrinogen. Figure 2 suggest that this group acts as buffer between two broad categories of patients as transition matrix 4 illustrates. Randomness being a part of initial imputation and Optimal Prescriptive Trees [3], it is now meaningful to question the stability of this first categorization by implementing the pipeline described in section 3.3. The observed results in figure 3 brought us to build three categories of patients according to the number of trees which prescribed them fibrinogen:

1. Group 0: patients with low number of prescriptions i.e. < 2000
2. Group 1: patients for which trees did not reach any consensus regarding fibrinogen prescription.
3. Group 2: patients with very high number of prescriptions i.e. > 8000

Using the same quantiles as in the first clustering with 5 trees, we are able to reproduce a 6-category clustering with this last distribution. Very interestingly, the transition matrix 4 shows strong stability for category 0 and 5 – nearly $\approx 80\%$ – and diffusion to close categories for 1 to 4.

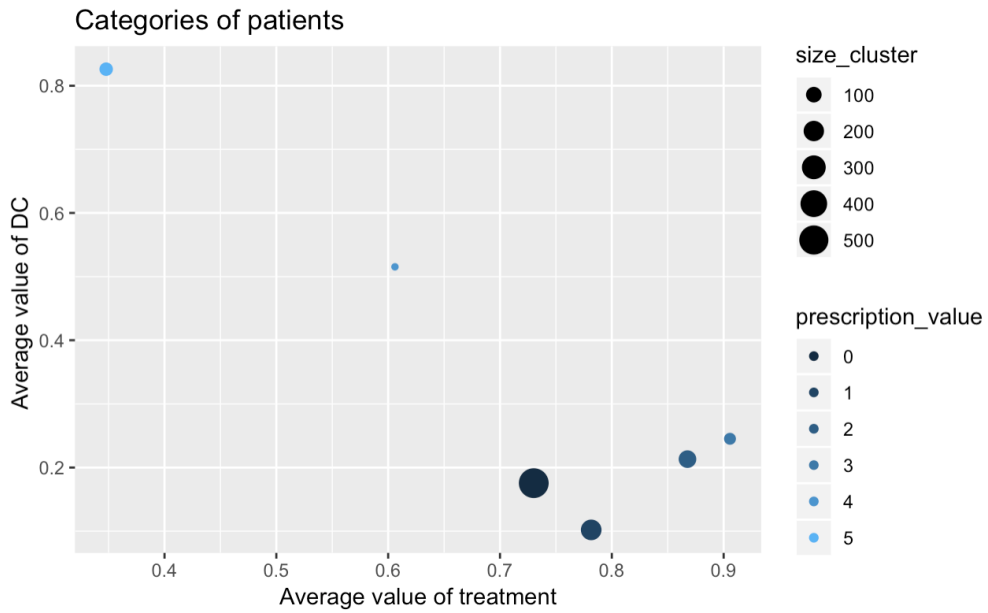

Figure 2: Plot of the 6 categories resulting from the 5 prescription

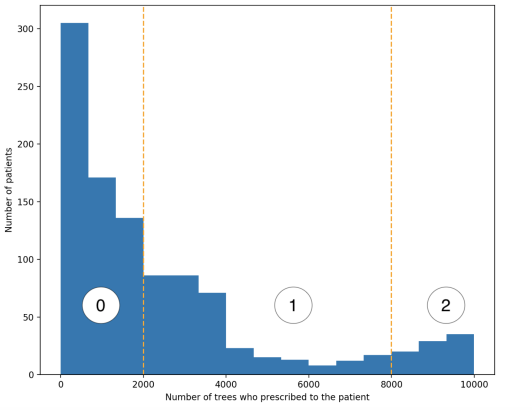

Figure 3: Distribution of # prescriptions on patients

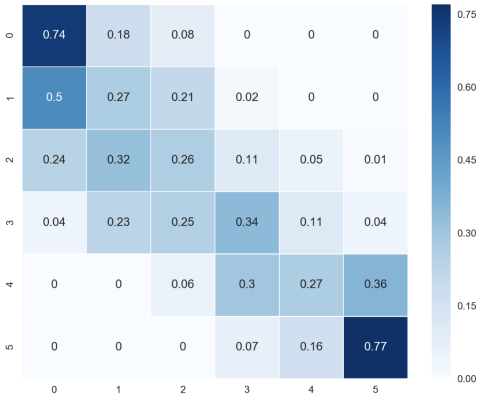

Figure 4: Transition matrix for the 6 initial groups

4.2 Classification for interpretability

In order to recover the interpretability lost by using 10k prescriptive trees, we trained an Optimal Classification Tree [2] to classify the sample in categories 0, 1 and 2 suggested by cuts on figure 3.

This results in figure 5 that can help doctors better understand the medical rationale behind the classification of each patients in each category.

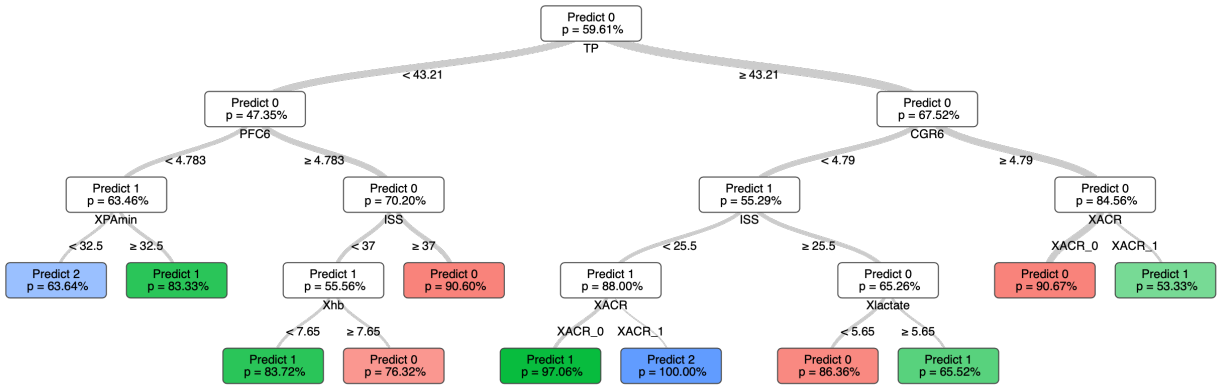

Figure 5: Optimal Tree Classification

4.3 Formal measure of ATE

Using the same process as in the previous study, we computed two ATEs:

- On **group {0 + 1}**:

$$ATE = 7.35\% \text{ with } 95\% \text{ CI: } [2.5\%, 12.2\%] \tag{1}$$

- On **group {1 + 2}**:

$$ATE = -16.7\% \text{ with } 95\% \text{ CI: } [-23.6\%, -9.94\%] \tag{2}$$

We included group 1 in both ATE in order to provide sufficient variability to GRF [7], so that it could successfully compare patients to reach an unbiased estimate of the average treatment effect.

These results suggest that fibrinogen injection decreases mortality for group 1 + 2 but increases it for group 0 + 1. However, these conclusions should be tempered because of:

- Survival bias: some patients die to soon to receive fibrinogen
- Prescription bias: some doctors might prescribe fibrinogen as an ineffective last resort solution

Conclusion

In this project we have used Optimal Prescriptive Trees to extract a subpopulation of patients potentially most responsive to fibrinogen injection. While the results of the ATE seem to confirm this intuition, such a complex affirmation should be tempered. Still, this work suggest a new way to formulate medical strategies without a clinical trial. Future developments:

- In order for this study to be directly actionable by doctors, we need to take a very close look to the temporal aspect of the variables. Indeed, for a protocol to be suggested, only variables available before the injection should be taken into account. Therefore, a key direction of development would be to repeat this study on a subset of such variables.
- Secondly, the survival bias should be addressed. One possible way would be to estimate the time of death by looking at the timeline of measurements.
- Finally, this should be replicated with synthetical data for which we know the counter-factual.

References

[1] Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests, 2016.

[2] Dimitris Bertsimas and Jack Dunn. Optimal classification trees. *Mach. Learn.*, 106(7):1039–1082, July 2017.

[3] Nishanth Mundru Dimitris Bertsimas, Jack Dunn. Optimal prescriptive trees. *INFORMS*, 2018.

[4] Traumabase Group. About section.

[5] Julie Josse and François Husson. missMDA: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, 70(1):1–31, 2016.

[6] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[7] Julie Tibshirani, Susan Athey, Stefan Wager, Rina Friedberg, Luke Miner, and Marvin Wright. *grf: Generalized Random Forests (Beta)*, 2018. R package version 0.10.0.

[8] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011.