Detection of Adverse Reaction to Drugs in Elderly Patients through Predictive Modeling

Rafael San Miguel Carrasco

Researcher, Universidad Internacional de La Rioja

***Abstract* — Geriatrics Medicine constitutes a clinical research field in which data analytics, particularly predictive modeling, can deliver compelling, reliable and long-lasting benefits, as well as non-intuitive clinical insights and net new knowledge. The research work described in this paper leverages predictive modeling to uncover new insights related to adverse reaction to drugs in elderly patients. The differentiation factor that sets this research exercise apart from traditional clinical research is the fact that it was not designed by formulating a particular hypothesis to be validated. Instead, it was data-centric, with data being mined to discover relationships or correlations among variables. Regression techniques were systematically applied to data through multiple iterations and under different configurations. The obtained results after the process was completed are explained and discussed next.**

Keywords — Geriatrics, Medicine, Data Analytics, Statistical Analysis, Predictive Modeling, Knowledge Management, Adverse reactions, Drugs.

I. Introduction

THE availability of big data and data analytics technologies and data analytics tools in the healthcare sector has not been seen as an advantage until recent times.

Today, managers of healthcare providers notice that data analytics can bring improvements in a broad range of business processes, and can also radically increase the effectiveness of service delivered to patients, while allowing for on-the-go research that can produce net new knowledge not intuitively or easily acquired by traditional research.

This research exercise aimed to generate a predictive model to accurately anticipate occurrence of such a relevant event as mortality (Exitus), among elderly patients admitted in a Geriatric Acute Unit.

While big data is typically associated with a high volume of data, variety (amount of data sources involved in an analysis) and velocity (speed at which data is generated) are also common big data features. This research exercise took advantage of using a dataset including variables from multiple data sources, as opposed to typical single-source, ad-hoc approaches often seen in traditional clinical research.

For this purpose, anonymized clinical records were mined with data analytics software. These records contained details about patients demographics, diagnosis, treatment, physical disability, mental disability, blood tests, admission-related complications and administered drugs. Patients' physical and mental disability were measured using CRF and CRM scales, respectively. These scales have been developed by Hospital Central de la Cruz Roja.

II. State of the Art

Previous research performed by professionals in the fields of interest was reviewed prior to starting the project.

This information allowed to gain an understanding of what other researchers discovered in the past, or are currently investigating on, as a useful reference of how this research must be approached.

Data used in this research work included information on drugs administered to patients. Therefore, it became relevant to understand what knowledge was available about adverse interactions between drugs and clinical variables like mortality, LOS (Length of Stay) and, at a higher level, cost associated with healthcare services delivery.

As Grizzle, F. R. [1] points out, average cost of an error in drugs administration is $977. The total cost is $177,5 billion, of which 70% represents the cost of patients' admissions resulting from these errors.

Matthew G. Whitbeck, R. J. [2] demonstrated that patients with atrial fibrillation suffer from multiple adverse reactions to Digoxin, including a higher mortality rate. The same conclusion is reached by Mate Vamos, J. W. [3], which explains that this adverse interaction is independent from other factors as kidney function, cardiovascular comorbidity or adherence to medications.

Also, Mate Vamos, J. W. [4] confirms that this circumstance is not limited to patients with AF (Atrial Fibrillation), but can also be applied to patients with CHF (Congestive Heart Failure), and suggest that this drug must be used with caution.

Finally, Wooten, J. M. [5] concludes that drug-administration errors have a higher impact on elderly patients. It also states that avoiding polypharmacy, rigorous analysis of drug interactions and frequent monitoring of patients' adverse reactions to drugs can dramatically lower risk.

III. Goals

The goal of this research work was developing a use case in which medical knowledge is extracted from a clinical dataset without a prior hypothesis.

In a broader context, this research work attempts to provide a practical example to shed additional light on finding an answer to the following questions:

1. Can data analytics support traditional clinical research in generating valuable medical knowledge while being less dependent upon intuition?
2. Can data analytics improve current clinical research's efficiency by providing additional insights and shortening deadlines?

IV. Methodology

A. Data sources

Clinical records of patients' admissions to Geriatrics Acute Unit from Jan 1, 2006 to Jul 31, 2015 (N=11.795) were extracted from a clinical dataset in Microsoft Access format. These observations were anonymized and saved in an appropriate format to be used in SAS.

DOI: 10.9781/ijimai.2016.368

In order to make the planned analysis affordable through standard computing resources, the dataset was filtered to extract patients admitted from nursing homes in the first half of 2015 (N=138). Having said that, this use case can scale up to millions of records with no changes in implementation.

The resulting dataset was then combined with data extracted from additional Clinical Information Systems to obtain further clinical variables about each patient. These variables related to drugs-administration, medical tests and consults, and visits to emergency units.

The final number of variables considered was 81 (p=81). The set of observations with inputs from multiple data sources constituted a data lake in which multiple queries can be run.

For the sake of simplicity, this article focuses on insights related to mortality, which is a key clinical variable. However, the same data lake can be used with no changes for any other analysis related to these patients' admissions.

B. Data preparation

Several preparation routines were run against the dataset to facilitate agile and effective mining activities once loaded into SAS.

Particularly:

1. Missing values. The following default values were assigned to empty cells: "Missing", for categorical (discrete) variables, and empty string ("") for numeric variables.
2. Deletion of records. Applied to those records where missing values occurred in key fields for the analysis.
3. Review of minimum and maximum values to detect outliers or erroneous values. These were replaced by the average, minimum or maximum value in the field, depending on each particular case.
4. Removal of irrelevant, unused or redundant variables.
5. Transformation of variables. Admission and discharge dates were replaced by length of stay, and birthdate was converted to age.
6. Replacement of numeric codes with meaningful strings, to allow for faster interpretation of results.
7. Replacement of strings with numeric codes, to fine-tune input variables before executing regression techniques.

C. Data analysis

The following statistical analysis and modeling techniques were used:

1. Calculation of descriptive indicators, to understand each field's structure.
2. Transformation of variables, to increase the degree of linear correlation between available inputs and the target variable.
3. Variable selection, to rapidly discard those inputs that show low predictive capabilities.
4. Logistic regression

D. Toolset

SAS Enteprise Miner[1] was used for this research work. The suggested methodology to perform logistic regression analysis described by Sharma, K. S. [6] was implemented.

E. Limitations

The methodology, tools and data used in this research work is subject to several limitations that are described next. This information will help the reader assess whether obtained results are reliable enough for a particular scenario.

Data sampling

The original dataset was filtered to obtain those patients having been admitted from a set of nursing homes in the first half of 2015. Typically, one year is a more appropriate period for inference techniques to be reliable.

In addition to this, the dataset was subject to bias, given that all patients records belonged to a single hospital. Ideally, these records must have been obtained from multiple hospitals.

Data quality

Data was gathered by healthcare professionals, and input into well-designed clinical Information Systems implementing measures to avoid input errors.

However, the risk of having erroneous data is not fully mitigated. Also, certain variables' values are influenced by the subjective perception of the doctor or nurse.

Accuracy of results

Software used is enterprise-class and commonly used in scientific studies and research. Furthermore, criteria applied to assess statistical significance was based on generally accepted practices.

However, this doesn't imply that they are suitable for other scenarios or use cases beyond the context of this research work.

Seasonality

Selected records covered a period of six months. Therefore, seasonality factors couldn't be accounted for. This might result in biased values. However, the resulting deviation won't likely impact the final results that were obtained.

Geographical factors

As stated previously, the source of data was one hospital in Madrid (Spain).

Therefore, conclusions might not be applicable to other geographies. However, this source of bias is common to most clinical trials.

Methodological errors

Data mining procedures used along this research constitute industry best practices. Nevertheless, other context-related factors might not have been taken into account.

Other limitations

No additional limitations were identified.

In addition, conflicts of interest were not found to apply to the author of this work or any of his collaborators. None of the participants will personally benefit from obtained results.

V. Detailed Procedure

A. Background

Patients mortality is a key clinical variable.

Datasets were mined to discover what variables could accurately predict mortality (Exitus) on a given set of patients.

B. Methodology

The following diagram was designed and run in SAS Enterprise Miner to build the model:

The regression node was configured as follows:

- Two factors interaction: No.
- Polynomials terms: No.

1 http://www.sas.com/en_us/software/analytics/enterprise-miner.html

- Regression type: Logistic regression.
- Link function: Logit.
- Model selection: Stepwise.
- Selection criteria: Validation error.
- Optimization technique: Default.

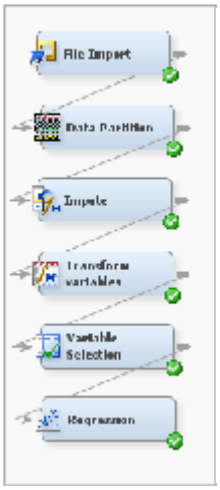


Fig. 1. SAS diagram.

Model fit indicators and relative risk (odds-ratio) values were obtained and analyzed in order to assess reliability and predictive capabilities of the model.

C. Obtained results

The process to build the model was split in several iterations.

In the first iteration, a single variable was found to predict mortality with 100% of accuracy: Place of Exitus. It became obvious that this variable had to be removed from the model.

In the next iteration, Morphine was found to accurately predict mortality. However, since this drug is typically administered to patients when they are about to pass away, the resulting model would offer no predictive capabilities to a doctor. As such, this drug was also removed from the model.

In the third iteration, however, a model containing several meaningful variables was obtained.

These were the following:

1. Digoxin, a drug that has been proved to be associated with higher mortality rates for other populations in previous clinical trials.
2. Number of lab tests requested by the doctor during the admission process.
3. Occurrence of pressure ulcers.

The model was assessed to check whether it was reliable and accurate from a statistical perspective.

Most relevant model fit indicators displayed by SAS are shown in Fig. 2 and discussed next:

- **Global Null Hypothesis**, that tests whether all coefficients in the regression model are zero, was rejected (p-value < 0.0001).
- **Type 3 Analysis of Effects**, that tests whether each individual predictor's coefficient in the model is zero, shows that three of those variables (I_LAB_Number, I_Pharma_Digoxine, TI_GN_Evaluation_Ulceras_p3) exhibit non-zero coefficients when entered in the model (p-values 0.004, 0.0207, 0.0048, respectively).
- **Odds Ratio Estimates**, which are the proportions of observations in the main group (mortality) compared to the control group (no mortality) with respect to each predictor in the model, confirm that the three previous predictors influence patients' mortality (with odds ratio estimates of 1.273, 3.953, 6.280, respectively).

D. Conclusions

The built model turned out to be statistically significant and accurate. Therefore, it would be ready to be implemented in a production environment to predict mortality for a given set of patients.

Furthermore, the confusion matrix depicted below confirms that

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
100.087	63.714	36.3732	3	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
I_LAB_number	1	8.1078	0.0044
I_Pharma_DIGOXINA	1	5.3495	0.0207
TI_G_N_Evaluation_Ulceras_p3	1	7.9718	0.0048

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept	1	-2.0981	0.6488	10.46	0.0012		0.123
I_LAB_number	1	0.2414	0.0848	8.11	0.0044	0.7080	1.273
I_Pharma_DIGOXINA	1	1.3746	0.5943	5.35	0.0207	0.5818	3.953
TI_G_N_Evaluation_Ulceras_p3 0	1	0.9187	0.3254	7.97	0.0048		2.506

Odds Ratio Estimates

Effect	Point Estimate
I_LAB_number	1.273
I_Pharma_DIGOXINA	3.953
TI_G_N_Evaluation_Ulceras_p3 0 vs 1	6.280

Fig. 2. Model fit indicators.

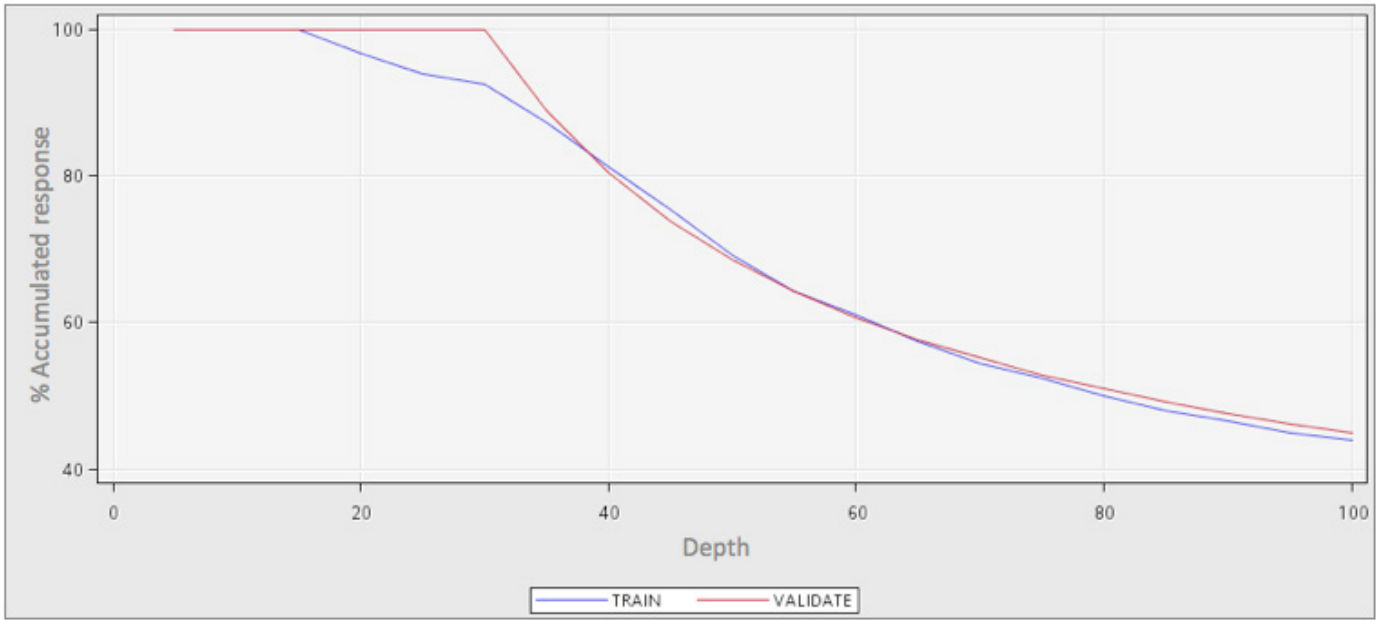


Fig. 3. Percentage of captured response for training and validation data.

the model performed quite well against the train dataset, with no overfitting signals:

Event classification table

Data role=TRAIN Target=B_Exitus_Horus Target label=B_Exitus Horus

False negative	*True negative*	*False positive*	*True positive*
8	*35*	*6*	*24*

Out of 30 cases of mortality in our training data, the model was able to predict 24 of them (80%).

VI. Obtained Results

Digoxin, a regular drug occasionally used in the treatment of various heart conditions, was proved to be linked to patients' mortality. This connection had been demonstrated for other populations but never tested with elderly patients.

Also, other admission-related factors turned out to contribute to increase the likelihood of an elderly patient passing away.

These circumstances were discovered without an initial hypothesis about how available input variables (including drugs administration) could be related to the target variable (mortality).

Beyond actual results, it is also proven that data analytics can uncover non-intuitive or complex relationships in a far more efficient way, with no prior assumptions required to trigger a research exercise, and with multiple potential results brought along the discovery process.

VII. Discussion

Operationalizing a predictive model that is able to anticipate which patients are at a higher risk of mortality allows doctors to adjust their treatments on time and provide further attention to their evolution overtime, hence lowering the mortality rate.

This outcome from a research exercise is highly desirable and fulfills the purpose of enhancing delivery processes in the healthcare environment.

However, one can argue that this result had been proven in a different population by previous studies. What makes this exercise different is the approach taken to get to results: no prior questions were asked and no hypothesis was formulated.

The approach was completely agnostic. As such, the goal was not to uncover an adverse reaction to drugs or to what extent pressure ulcers increases the mortality likelihood. The main purpose was to uncover hidden relationships among clinical variables having an impact on mortality.

VIII. Conclusion

The ultimate goal of this research exercise was to provide a real-life use case in which data analytics added value to traditional clinical research. This added value would translate into insights that were not part of an initial hypothesis, but rather discovered on-the-go while crunching available data.

A doctor's extensive medical knowledge is still limited by his/her professional experience and subjective interpretation of available details. This use case attempts to go beyond this limitation by analyzing data from an agnostic point of view, leveraging all available variables and avoiding prior assumptions that could limit potential results.

Using a data-centric strategy to analyze data, as opposed to generating a custom dataset and focusing on a particular goal, might increase the amount of conclusions derived from the research process and uncover unexpected insights that would have not become a priority otherwise.

The example provided in this article has focused on trying to demonstrate how such data-centric approach would work on a very limited and simplified scenario. Therefore, further research would be highly recommended to gradually prove how far data analytics' contribution could potentially be.

IX. Future Research Paths

Once confirmed that it's feasible to build reliable statistical models to detect adverse reactions to certain drugs and predict mortality, it would be highly suggested to perform additional research in order to go beyond these initial results.

In this research exercise, a limited sample of observations was used. By processing all available patients' data, which can scale up to millions of clinical records in large hospitals, the amount, quality and accuracy of obtained insights would likely be much higher than what was obtained here.

Likewise, in order to realize the benefits of these additional insights, they must be operationalized, that is, made available to doctors through a production-ready Information System implementing a Recommendation Engine (RE).

Lastly, given the inherent computational complexity of such an implementation, RE's performance metrics must be taken into account in the process, as pointed out by Luis F. Chiroque [7].

Acknowledgment

This research was possible thanks to PhD Javier Gómez Pavón, senior doctor, as well as expert and researcher in Geriatrics Medicine.

PhD Beatriz Ares Castro-Conde, senior doctor, also made a contribution to this research.

References

[1] Grizzle, F. R. "Drug-Related Morbidity and Mortality: Updating the Cost-of-Illness Model", J Am Pharm Assoc (Wash). 2001 Mar-Apr;41(2):192-9, 2001.

[2] Matthew G. Whitbeck, R. J. "Increased mortality among patients taking digoxin—analysis from the AFFIRM study", European Heart Journal (2013) 34, 1481–1488 doi:10.1093/eurheartj/ehs348, pp. 1-8, 2013.

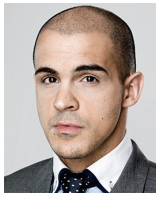
[3] Mate Vamos, J. W. "Increased Mortality Associated With Digoxin in Contemporary Patients With Atrial Fibrillation", Journal of the American College of Cardiology, pp. 1-8, 2014.

[4] Mate Vamos, J. W. "Digoxin-associated mortality: a systematic review and meta-analysis of the literature", European Heart Journal doi:10.1093/eurheartj/ehv143, pp. 2-7, 2015.

[5] Wooten, J. M. "Pharmacotherapy Considerations in Elderly Adults", South Med J. 2012;105(8):437-445 doi: 10.1097/SMJ.0b013e31825fed90, pp. 2-7, 2012.

[6] Sarma, K. S. "Predictive Analytics with SAS Enterprise Miner", SAS, pp. 359-371, 2013.

[7] Luis F. Chiroque. "Empirical Comparison of Graph-based Recommendation Engines for an Apps Ecosystem", IJIMAI , DOI: 10.9781/ijimai.2015.327, pp. 35-36, 2015.



Rafael San Miguel Carrasco has developed his professional career in the Technology industry for the past eleven years. He has taken on roles in the field of research, technology, project management, team leading, business development and middle management. He has worked for multinational firms as Deloitte, Telefónica, Santander and FireEye, engaging on and leading international business initiatives combining technology, management and operations. Rafael works as a Data Scientist at Universidad Internacional de la Rioja (UNIR) in a research initiative in the field of Healthcare Analytics, where big data technologies as SAS, R and Hadoop play a key role.