20

Predicting healthcare cost of diabetes using machine learning models

Javier-Leonardo González-Rodríguez[♭1], Javier Díaz Carnicero[♮], David Vivas-Consuelo[♮], Silvia González de Julian[♮] and Olga Lucia Pinzon Espitia[♯]

(♭) Business Management School,
Universidad del Rosario, Bogotá,
(♮) INECO,
Universitat Politécnica de València,
(♯) Universidad Nacional de Colombia.

1 Introduction

Diabetes mellitus (DM) describes a group of metabolic disorders characterised by high blood glucose levels. People with diabetes have an increased risk of developing several serious life-threatening health problems resulting in higher medical care costs, reduced quality of life and increased mortality [1] DM, like the majority of non-contagious chronic diseases, is associated with multimorbidity, defined in the growing literature as the existence of two or more chronic conditions [2,3]. Multimorbidity causes a negative impact on both clinical and health indicators and primary health care costs [1,4]. While true that the analysis of multimorbidity in this type of population is relatively new, the tendency towards this approach to the study of chronic diseases is ever increasing [5, 6].

This co-occurrence of diseases has implications from a disease management point of view, as the features of comorbid diseases can be much more complicated than a simple aggregation of individual illnesses [7, 8]. Previous studies have related DM to a set of diseases such as cardio-vascular, renal, obesity and the metabolic syndrome.

Diabetes mellitus Type II (DM2) [9] is among the chronic diseases that generate the most health expenditure and clinical risk, due to the comorbidities that it frequently deals with. For this reason, it is very important to determine a total risk index calculated based on the variability determined by the number and severity of the associated morbidities.

Based on this risk index, a predictive model of pharmaceutical expenditure can be developed, applicable not only to DM2, but also to other chronic diseases.

[1]e-mail: jagonro1@upvnet.upv.es

2 Materials and methods

Objective
To design a predictive model of the pharmaceutical expenditure of DM2 patients, derived from the risk index determined by the associated comorbidities, in a health district of Valencian Community Spain.

- Cross-sectional descriptive and analytical study, and predictive models of total healthcare expenditure for application in clinical management.
    - **Population**: 28.345 DM2 patients in a public health district from Valencian Community.
    - **Sample**: 13.820 patients, equivalent to 40% who had complete data to asses, according to de defined variables.
    - **Variables**: Age, sex, Primary and secondary diagnosis, (Comorbidity and Multimorbidity), Clinical Risk Groups - CRG, Glycosylated haemoglobin – HbA1c, Average Glycemia, Creatinine, Microalbuminuria, Lipid profile, (total cholesterol, Triglycerides, Glomerular Filtering.

3 Modelling

Prediction of events and complexities related to DM2 based on clinical information using logistic regression models:

Main Risk:

- Acute Myocardial Infarction.
- Brain Vascular Stroke.

Complications:

- Chronic Kidney Disease - Kidney Failure - Transplant.
- Diabetic Retinopathy - Secondary Blindness.

Prediction of the pharmaceutical expenditure using the calculated risk of events and complexities, comparing between classical linear regression a machine learning models.

4 Results

Descriptive Analysis The descriptive analysis is presented in the following tables, highlighting the most relevant aspects, such as the distribution of patients by age and hospital stay, on the one hand, and on the other hand, the distribution of patients from the perspective of the most significant events or comorbidities related to diabetes, these are: retinopathy, chronic kidney disease - CKD, myocardial infarction and stroke BV (Table 3).

The variables used for the design of the predictive model respond to explicative aspects, both patients themselves and of the conditions of their illness and allow us to assume both the behaviour in the variation of the risk and its impact on the costs of care.

Thus, age and sex can be related to variations in days of hospital stay and therefore in cost. The existence of a primary diagnosis, in this case diabetes or some secondary diagnoses or comorbidities, affect the Clinical Risk Group CRG index, glycosylated haemoglobin - HbA1c, and average Glycemia, account for the state of diabetes (controlled or not controlled), Creatinine and Microalbuminuria allow the calculation of GFR, with which the risk classification KDIGO is obtained; and finally the lipidic profile (total cholesterol, Triglycerides, gives an idea of the state of cardiovascular risk, (total cholesterol, Triglycerides, gives an idea of the state of cardiovascular risk), (total cholesterol, Triglycerides, gives an idea of the state of cardiovascular risk), and finally the lipid profile, (total cholesterol, Triglycerides, gives an idea of the state of cardiovascular risk).

In the table 1, the predominance of male patients can be appreciated, with 52.7% of the cases as opposed to 47.3% of female sex. This higher frequency of male patients is also evident in the distribution by hospital stay, which is shown in (table 1, 2).

AGE / **G age**	**SEX**	**SIP_ RECOD** / **YS_RECOD**	
< 40 years	M	185	
	F	224	
	Total	**409**	**3,0%**
> 85 years	M	413	
	F	749	
	Total	**1162**	**8,4%**
$40 - 55$ years	M	978	
	F	555	
	Total	**1533**	**11,1%**
$55 - 70$ years	M	2997	
	F	2013	
	Total	**5010**	**36,3%**
$70 - 85$ years	M	2715	
	F	2991	
	Total	**5706**	**41,3%**
Total	**M**	**7288**	**52,7%**
	F	**6532**	**47,3%**
	Total	**13820**	

Table 1: Age distribution of patients.

It is striking that 86% of the cases correspond to patients over 55 years of age, so it is worth reflecting on whether age is a decisive factor in the presentation of greater association with comorbidities. As well as these patients over 55 years of age, they explain 80% of the 16831 days of hospital stay.

On the other hand, it was found that the distribution of the data has a non-parametric character, which is why, to evaluate the level of significance of the distributions, the binomial test was used, which yields a highly significant result, (table 4).

SEX	G. AGE	N	MEDIA	SUM	PERCENTAGE
M	< 40 years	185	0,23	43	0,3%
	> 85 years	413	2,09	864	5,1%
	$40 - 55$ years	978	0,62	606	3,6%
	$55 - 70$ years	2997	0,96	2887	17,1%
	$70 - 85$ years	2715	1,74	4711	27,9%
	Total	**7288**	**1,25**	**9111**	**54,0%**
F	< 40 years	224	1,06	237	1,4%
	> 85 years	749	1,82	1361	8,1%
	$40 - 55$ years	555	0,5	279	1,7%
	$55 - 70$ years	2013	0,69	1395	8,3%
	$70 - 85$ years	2991	1,50	4498	26,6%
	Total	**6532**	**1,19**	**7770**	**46,0%**
Total	< 40 years	409	0,68	280	
	> 85 years	1162	1,91	2225	
	$40 - 55$ years	1533	0,58	885	
	$55 - 70$ years	5010	0,85	4282	
	$70 - 85$ years	5706	1,61	9209	
	Total	**13820**	**1,22**	**16881**	

Table 2: Distribution of patients by stay.

SEX	N	RETINOPATHY	CDK	INFARCT	BVS	PIELONEFRITHYS
MALE	**7288 (54%)**	239	168	329	106	66
FEMALE	**6532 (46%)**	211	88	117	85	149
TOTAL	**16881 (100%)**	450	256	446	191	215
STAY	16881	918	1181	1132	4722	353

Table 3: Event summary morbidity.

Event Prediction. Logistic Regression Results. In order to the prediction of events and complexities related to DM2 (Infarction, Stroke, Retinopathy, Renal failure) we propose different logistic regression models, using available clinical information. A linear combination of the variables with their corresponding coefficients is transformed via the logistic function, presented below, in order to obtain the probability of an event occurring.

$$P(y = 1|x) = \frac{\exp(x)}{1 + \exp(x)}$$

For instance, the results obtained for the prediction of an infarction event occurring are shown hereafter. The results were obtained for the rest of the variables in a similar way, varying the correspondent coefficients in order to maximize the predictive value of the model.

$$x = -341,93 + 0,37\ State\,of\,health + 0,29\ Severity - 20,53\ Filtr - 97,25\ Album - 1579,92\ HbA1C - 2,95\ Cholestherol$$

The resulting ROC for our example curve can be seen in graph 2, and its corresponding area under the curve is 0.767, which determines a satisfactory predictive power of the model. Additionally, the calibrations of the model allow to have a high negative predictive value, over 75%, without trading off the overall results of the model. The results are similar for all the logistic regression prepared.

		Category	N	Observed Prop	Test Prop	Bilateral Exact Sign
SEX	Group1	1	3150	0,60	0,50	0,000
	Group 2	0	2083	0,40		
	Total		5233	1,00		

Table 4: Event summary morbidity.

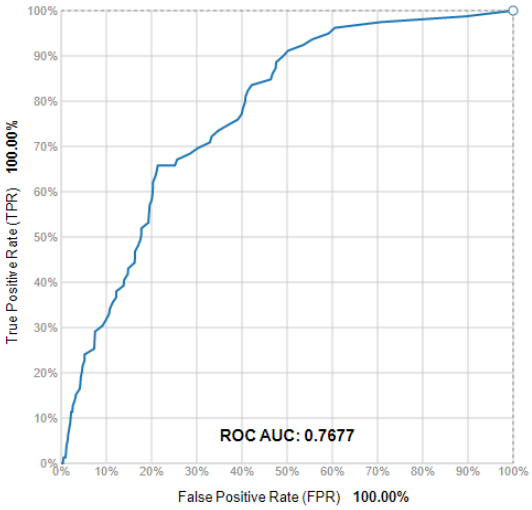

Figure 1: Predictive Power of the Regression Model.

Pharmaceutical expenditure prediction models. Now we try to create the model for the prediction of the pharmaceutical expenditure using the calculated risk of events and complexities. Firstly, we calculated the linear regression, and the equation proposed is as follows:

$$\begin{aligned} Expenditure = & -1,50 \times 10^5 - 172,41\ Age - 6537,10\ Gender + 41001\ State\,of\,Health \\ & + 2468,60\ Severity + 46377\ Retinopathy + 13946,3\ Renal\,failure \\ & + 34143\ Infarction + 22452,9\ Stroke \end{aligned}$$

This R^2 value obtained was $0,32$, it is not significant enough for a practical use, but values close to 50% would be expected according to the studies published.

Secondly, we prepare for the prediction model a machine learning approach. For this purpose, we prepare a neural network with 1 hidden layer and the ADAM algorithm for training. In order to compare the results with the classical linear regression, we selected the same variables. In this case the R^2 resultant is $0,35$. This is slightly higher than that obtained by linear regression, but there is no noticeable difference in practice.

5 Conclusions

The risk management model is based on the study of expenditure on health services caused by DM and its comorbidities, which have a significant impact on the health services budget, with pharmaceutical expenditure being the most relevant.

It is shown how expenditure increases significantly as the number of associated diseases increases, so it can be deduced that the financial risk index is definitively associated with the comorbid-based risk class. These elements provide some basis for the design of the prescriptive spending model. The risk prediction models are particularly valid for their negative predictive value.

While Machine learning models lightly improve the result, their computational cost is significantly higher, so the linear regression is globally a more effective alternative.

It would be necessary to collect more variables in order to improve the predictive outcome of our model.

References

[1] Sundararajan, V., Henderson, T., Perry, C., Muggivan, A., Quan, H. and Ghali, WA., New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality. *J Clin Epidemiol*, 57(12): 1288–94, 2004.

[2] Fox, R. and Fletcher, J., Alarm symptoms in primary care. *Br Med J.*, 334(7602): 1013–4, 2007.

[3] Valderas, JM., Starfield, B., Sibbald, B., Salisbuty, C. and Roland, M., Understanding Health and Health Services. *Ann Fam Med.*, 7(4): 357–63, 2009.

[4] Glynn, LG., Valderas, JM., Healy, P., Burke, E., Newell, J., Gillespie, P., et al. The prevalence of multimorbidity in primary care and its effect on health care utilization and cost. *Fam Pract.*, 28(5): 516–23, 2011.

[5] Barnett, K., Mercer, SW., Norbury, M., Watt, G., Wyke, S. and Guthrie, B., Epidemiology of multimorbidity and implications for health care, research, and medical education: A cross-sectional study. *Lancet* [Internet], 380(9836):37–43, 2012. Available from: `http://dx.doi.org/10.1016/S0140-6736(12)60240-2`.

[6] Holden, L., Scuffham, PA., Hilton, MF., Muspratt, A., Ng, S. and Whiteford, HA., Patterns of multimorbidity in working Australians. 1–5, 2011.

[7] Stavem, K., Hoel, H., Skjaker, SA. and Haagensen, R., Charlson comorbidity index derived from chart review or administrative data: Agreement and prediction of mortality in intensive care patients. *Clin Epidemiol* [Internet], 9:311–20, 2017. Available from: `https://www.scopus.com/inward/record.uri?eid=2-s2.0-85020468676&doi=10.2147%2FCLEP.S133624&partnerID=40&md5=9b93e44559138d782f578b59f99cafc4`.

[8] Fritzen, K., Heinemann, L. and Schnell, O,. Modeling of Diabetes and Its Clinical Impact. *J Diabetes Sci Technol*, 12(5):976–84, 2018.

[9] Caballer-Tarazona, V., Guadalajara-Olmeda, N. and Vivas-Consuelo, D., Predicting healthcare expenditure by multimorbidity groups. *Health Policy* (New York) [Internet], 123(4):427–34, 2019. Available from: `https://doi.org/10.1016/j.healthpol.2019.02.002`