P. de Toledo, R. Pérez-Rodríguez, P. de Miguel, A. Sanchis and P. Serrano, "Prediction of patient evolution in terms of Clinical Risk Groups form routinely collected data using machine learning," *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Berlin, Germany, 2019, pp. 1721-1724

DOI: 10.1109/EMBC.2019.8857625

Prediction of patient evolution in terms of Clinical Risk Groups form routinely collected data using machine learning

Paula de Toledo, Rodrigo Pérez-Rodríguez, Pablo de Miguel, Araceli Sanchis, Pablo Serrano

Abstract— **Chronicity is a problem that is affecting quality of life and increasing healthcare costs worldwide. Predictive tools can help mitigate these effects by encouraging the patients' and healthcare system's proactivity. This research work uses supervised learning techniques to build a predictive model of the healthcare status of a chronic patient, using Clinical Risk Groups (CRGs) as a measure of chronicity and prescription and diagnosis data as predictors. The model is addressed to the whole population in our healthcare system regardless of the disease, as data used are widely available in a consistent way for all patients. We explore different ways to encode data that are appropriate for machine learning. Results suggest that these data alone can be used to build accurate models, and show that, in our set, prescription information has a higher predictive value than diagnosis.**

I. Introduction

Chronicity an increasing problem in developed countries [1]. Demographic growth and unhealthy life habits are causing a significant increase in the number of chronic patients, both harming people's quality of life and increasing healthcare system costs. Hospitals and regional healthcare providers need to understand the chronicity of the population they attend to, with tools such as predictive models that estimate the evolution of this population for planning, resource allocation and to assess actual performance against expected (predicted) performance. Such models can be as well a clinical tool for individual patients informing them of their predicted risks as a tool to promote healthier lifestyle choices. This paper outlines a preliminary methodology to build such models using health related data gathered for non-clinical purposes by a local healthcare provider in Spain.

In the last decade, Electronic Health Care Records (EHRs) have finally reached widespread adoption in most developed countries. This provides a growing body of data that can be used to build descriptive and predictive models. The named big data revolution is happening in the healthcare sector as well as in all other markets and has the potential to transform medical practice by using information generated every day to improve the quality and efficiency of care [4].

However, even though EHRs are widely available, the way clinical data is stored is not standard yet, as it varies within healthcare providers and even in different departments on an organization. EHR data are not ready yet to be used as the basis of a methodology to build high level models to predict evolution for a population. For this reason we propose the use of healthcare related data gathered for administrative purposes - such as billing - to build our models. In our (Spain) and many other countries, the recording of this information is required by law and has been registered routinely for more than two decades.

The strategies for chronic patient management currently deployed require a prior stratification of patients to estimate necessary human and economic resources. This stratification is performed by using population groupers. A widespread stratification system is the so-called clinical risk groups (CRGs) [2]. CRGs are a categorical clinical model in which each individual is assigned to a single mutually exclusive risk group based on the historical clinical and demographic characteristic of the individual. The CRG is an indicator of the amount and type of healthcare resources he or she consumes. Policy makers use the CRG mix in a population to make decisions for strategic management of healthcare services and financing, such as adjusting capitation payments within an organization or predicting health service use [2]. More details on the CRG model are provided in section II.a.3.

A. Objectives

The objective of this work is to design a methodology to build predictive models to forecast the evolution of patients in terms of CRG group, using clinical data that are routinely collected for all patients within a healthcare setting. For this reason we decided to use prescription and diagnosis information as described later. Specifically, we want to develop two different models: a binary classifier (remain stable vs worsen) and a five level classifier predicting the next year CRG of a patient. To achieve this goal, a key step is to identify the optimal way to represent the diagnostic and prescription information available, a representation that is both meaningful and practical for machine learning.

B. State of the art

Mining of electronic healthcare records using supervised learning techniques for predicting the evolution of patients is an active area of research[3][4]. Some predicted aspects are the risk of hospitalization[5][6], the development of complications [7][8] or the outcome of an episode[9][10]. Most published work focuses on a specific disease, but the wider availability of integrated EHR is now paving the way for broader approaches -like the one proposed here-, that allow for the analysis of the population in an area and are of interest mainly to support management and resource allocation decisions. This approach means working with broad, high-level information, like diagnosis or pharmacy,

P. De Toledo and A. Sanchis are with the Computer Science Department, Universidad Carlos III de Madrid, Spain. (e-mail: mtoledo@inf.uc3m.es).

Rodrigo Perez is with the Biomedical Research Foundation, Getafe University Hospital and with Universidad Carlos III de Madrid.

P. de Miguel-Bohoyo is with Fuenlabrada University Hospital, Madrid.

P. Serrano is with Hospital Universitario 12 de Octubre, Madrid.

This work was supported in part by projects TRA2015-63708-R and TRA2016-78886-C3-1-R (Spanish Government) and Predict-TB (European Union, Innovative Medicines Initiative).

and not with specific disease-related information available in departmental information systems. Current popular research areas in the field are optimal ways to represent EHR information for mining, feature selection [11] and trustworthy reuse and privacy preservation [12][11].

II. MATERIAL AND METHODS

A. Data items

1) The Minimum Basic Data Set

Since the mid 90's, Spanish hospitals are required by law to keep a registry named "Minimum Basic Data Set" (MBDS) for every encounter with the healthcare system at any level (primary care visit, hospital outpatient visit or inpatient stay, emergency care). For every encounter, demographic data, main diagnosis motivating the encounter and additional diagnoses are recorded. The coding system used was the International Classification of Diseases (ICD) in its 9th revision (ICD-9-CM), upgraded to the more specific and wider scope version ICD-10 since 2016.Although the MBDS data is gathered for administrative purposes and is much less specific than a fully deployed Electronic Patient Record or a more specific Departmental Information System, it has several features that make it very useful for building non-disease-specific models as is our aim in this work. The main feature is its span to the complete population, as it is recorded for every patient contact with the public Healthcare System. The second feature is its generalizability to all patients regardless of the disease, as more specific Departmental Information systems capture the specificities of a single disease, but the data gathered are difficult to merge with data from other disease-oriented sets. The third feature is the direct access to temporal information, as Departmental Systems and general purpose EHRs sometimes require pre-processing to generate time-labeled diagnostic data.

2) Drug prescription information

In our dataset, prescription information is codified with the widely adopted drug coding system ATC (Anatomical Therapeutic Chemical classification system). Electronic prescription systems are in place for in-hospital prescription in this setting since 2010 and in the last few years have been deployed in primary care as well.

3) Clinical Risk Groups

For this work we have used the higher level grouping of the CRG stratification approach consisting of nine different groups, out of which, five correspond to chronicity: single minor chronic disease (group 3), minor chronic diseases in multiple organ systems (4), significant chronic disease (5), significant chronic diseases in multiple organ systems (6) and dominant chronic disease in three or more organ systems (7). The other four groups are: Catastrophic condition status; History of major organ transplant; Dominant and metastatic malignancies; History of significant acute disease and Healthy/Non-Users.

B. Description of the source data

This work is based on data collected from a suburban area to the south of Madrid in central Spain, composed of a University Hospital and nine primary healthcare centers. This cluster provides public healthcare services to a population of 225,000 people. An encounter record contains patient demographic information, date, and care setting (primary care, secondary care, pharmacy or emergency) as well as codified diagnostic information (up to 15 diagnostics per encounter using ICD-9) or prescription information. Clinical Related Groups (CGR) score for a patient is generated by the hospital administrative staff on a yearly basis. To support this task an automated tool (3MTM Clinical Risk Grouping Software) using diagnostic and pharmacy data recommends score, which is then validated. Data available encompasses years 2010, 2011 and 2012

C. Data preparation: from encounters to patient profile. Representing data in a way that is meaningful for machine learning

Source data are processed to generate a list of entries corresponding to an encounter with the following items [patient id, date, care setting, list of ICD codes, and list of ATC codes].

1) Selecting patients

As the goal is to predict the evolution of chronic patients, we included only patients with a CRG code indicating chronicity (3-5). Out of the 250,000 patients registered with the healthcare provider, 161,511 had at least one contact with the system in the years of the study, and 31,587 had a CRG indicating chronicity and therefore were included.

2) Representing diagnostic and prescription information:

A naïve approach to build the predictive model would be to encode every diagnosis and drug as a feature (present / not present). This leads to a very sparse matrix not usable as-is as the input for the machine learning task. As a reference there are more than 13,000 ICD 9 codes and 4,482 of them are present in our dataset. We investigated different approaches to reduce the dimensionality of the data. A first approach completely disregards the codes content and uses the number of diagnostics per visit as a proxy for disease complexity, or aggregated information such as number of visits or time between visits. The second approach does take into account the diagnostic and prescription information but simplifies it using different strategies: a) use the different granularity levels inherent to the ICD codes and their hierarchical structure to reduce the number of features (truncate codes to three and four digits) and b) reduce the number of ICD codes by using only the most common ones. We also used a mixed approach where we counted the number of diagnostics from each of the first ICD hierarchical levels per visit (18 different values). Other approach would be to use domain information to group ICD codes into relevant higher-level groups. This was not included here as we were work towards a generic and data driven methodology. We also used different standard feature selection techniques but only once after data were aggregated with these approaches.

3) Representing time

Our goal is to predict the evolution of a patient in terms of CRG code for next year using information of the previous

year. We also tested if including data from the two previous years would increase the predictive ability of the model

4) Model outcome

As already mentioned, we have developed two different models: a) a binary model where the class to predict is remain stable vs worsen; and b) an ordinal multiclass classifier with 5 classes corresponding to the 5 CRG levels to predict. This second model has an important class imbalance, as groups 5 and 6 (significant chronic disease, significant chronic disease in multiple organ systems) are much more common in the dataset than the rest. Class imbalance is a well-known problem that affects the performance of machine learning classifiers [13] and is nearly always present in clinical dataset

5) Datasets

To assess the alternatives to represent information we built 6 datasets including different variables as described in Table I. Each dataset has two versions (.1 and .2) depending on the number of years used to predict (one or two).

TABLE I VARIABLES INCLUDED PER DATASET
ALL DATASETS INCLUDE AGE, GENDER AND PREVIOUS YEAR CRG

	DS1	DS2	DS3	DS4	DS5	DS6
Count of diagnosis per CGR (18 features)		X		X		
Most common diagnostics (200 features)			X		X	
Prescription information (200 features)				X	X	X

D. Building the models

1) Training and validation

Models were validated using a 10-fold cross-validation approach, in which the original data are randomly divided into 10 sub-samples, retaining one for testing and using the remaining 9 as training data. The selection of the most common diagnostics was done individually for each of the 9 subsamples to avoid inappropriately entering test data information into the train set. The number of different diagnostics and drugs to include was heuristically set at 200. To address the severe imbalance problem, a cost sensitive learning strategy was adopted, setting the costs proportional to the imbalance referred to the majority class [13].

2) Performance metrics

Sensitivity, Cohen's kappa statistics (κ) and area under the receive operator curve (AUC) have been used to compare the different classifiers. Kappa statistic [14] corrects the degree of agreement between the classifier's predictions and reality by considering the proportion of predictions that might occur by chance, and has the advantage over the more widely used AUC that it's easier to interpret for multiclass classifiers. Kappa values over 0.40 are considered moderate and over 0.60 good. Significance testing is done at a confidence interval of 95% using a two-tailed student t-test and using matching paired data.

Algorithms and tools

We used the open source tool Weka [15], a collection of state-of-the-art data mining algorithms and data preprocessing methods. The following machine learning algorithms (selected according to their suitability to the problem domain and coverage of different learning approaches) were used: a) bagging [16], as representative of ensemble learning using a fast decision tree learner (REPTree) [17] as base classifier; b) repeated incremental pruning (RIPPER) [18] that generates a set of classification rules; c) C4.5 classification trees [19] and d) Bayesian networks[20]. Each model was evaluated on the 12 datasets.

III. RESULTS

Tables III shows the results for the multiclass and binary model, where statistically significant superior performance values (for a dataset) are bolded. Drug prescription information (included only in DS4 and DS6) significantly improves the sensitivity of the model. Using a two year window for prediction (Dx.2) does also improve the results for both the multiclass and binary models. Results suggest that diagnostic information is not adding discriminant power to the predictive model regardless of it being grouped according to the first hierarchical level of ICD-9-CM or reduced by selecting the most frequent diagnoses. Only demographic information, previous years CRGs and prescriptions seem to be enough to estimate the future chronicity group of a patient. Exploring other alternatives to represent the diagnostic information that are more relevant for the model is needed.

Regarding the different classifiers tested, the ensemble tree learning classifier (bagging) yields the best predictive model. This is consistent with the literature [22].

Table III shows the confusion matrix for the best classifier (bagging) and dataset (DS6.2, two years with prescription information). CRGi stands for each of the chronicity groups, and TPR and FPR for true positives rate and false positives rates respectively). It is important to remark that even though there are classes with a relatively low TPR (CRG4 and CRG7), most of the misclassifications go to adjacent classes, and, being the model intended for resource estimation, this error is less important than classifying patients in non-adjacent classes.

TABLE II
CONFUSION MATRIX : MULTICLASS MODEL (BAGGING)
DATASET DS6.2. (TWO YEARS, PRESCRIPTION DATA)

	CRG_3	**CRG_4**	**CRG_5**	**CRG_6**	**CRG_7**	**TPR**	**FPR**
CRG_3	**94**	22	16	5	0	0,686	0,035
CRG_4	26	**43**	12	15	0	0,448	0,044
CRG_5	81	79	**641**	266	2	0,600	0,120
CRG_6	31	71	327	**1.954**	128	0,778	0,305
CRG_7	0	1	5	188	**59**	0,233	0,034
Avg.	-	-	-	-	-	0,686	0,224

As mentioned before the number of ATC codes and diagnoses to be used in the models using the most common codes was selected heuristically: we built different datasets with 5. 10, 200, 400 and 1000 most common codes and tested them with the bagging classifier for both models (binary and multiclass) statistically significant differences have been found in the binary classifier in terms of AUC, showing 200 codes perform better than the other options. For the multiclass problem the results are similar: datasets containing 200 and 400 ATC codes perform better (with statistical significance), than the rest in terms of hit ratio and kappa

statistics. Further investigation is needed to identify the optimal and minimum number of ATC and diagnostic codes

TABLE III
RESULTS
(S SENSITIVITY, Κ COHEN'S KAPPA, AUC = AREA UNDER ROC CURVE)

	Binary model												Multiclass model							
	BAGGING			RIPPER			C4.5			BAYES NET			BAGGING		RIPPER		C4.5		BAYES NET	
	S	κ	AUC	S	κ	AUC	S(%)	κ	AUC	S	κ	AUC	S	κ	S	κ	S	κ	S	κ
DS1.1	**75.19**	**0.49**	**0.83**	**75.24**	**0.49**	0.76	**75.27**	**0.49**	0.79	74.52	0.47	0.81	53.42	0.33	53.24	0.29	53.80	0.33	53.79	**0.34**
DS2.1	75.57	0.5	**0.84**	**76.3**	**0.52**	0.78	74.13	0.48	0.77	73.89	0.47	0.82	**56.48**	**0.36**	54.78	0.32	53.64	0.31	52.11	0.33
DS3.1	**76.01**	**0.51**	**0.84**	**76.22**	**0.51**	0.78	75.36	0.5	0.8	72.67	0.45	0.8	**55.52**	**0.36**	55.12	0.32	54.61	0.33	52.05	0.32
DS4.1	**78.96**	**0.57**	**0.88**	75.51	0.56	0.8	75.98	0.52	0.78	67.49	0.36	0.76	**63.27**	**0.45**	60.03	0.41	60.58	0.40	60.36	0.42
DS5.1	**79.00**	**0.58**	**0.88**	78.49	0.56	0.8	76.78	0.53	0.81	66.85	0.35	0.76	**63.34**	**0.45**	60.45	0.41	60.76	0.41	60.53	0.43
DS6.1	**78.88**	**0.57**	**0.88**	78.43	0.56	0.8	76.47	0.53	0.8	69.19	0.39	0.78	**63.10**	**0.45**	60.49	0.42	60.68	0.41	60.65	0.43
DS1.2	**78.00**	**0.56**	**0.85**	**78.35**	**0.57**	0.8	**78.55**	**0.57**	0.81	76.73	0.53	0.81	61.79	0.34	**64.23**	0.32	61.96	0.34	62.64	0.36
DS2.2	**78.21**	**0.56**	**0.85**	**78.37**	**0.57**	0.81	74.35	0.49	0.74	75.81	0.8	0.52	**64.57**	**0.36**	62.81	0.33	57.58	0.27	56.43	0.31
DS3.2	**78.45**	**0.57**	**0.86**	**78.66**	**0.57**	0.81	75.15	0.5	0.77	73.08	0.46	0.79	**64.22**	**0.36**	**64.22**	**0.34**	59.13	0.29	57.10	0.30
DS4.2	**80.61**	**0.61**	**0.88**	**80.24**	0.6	0.83	75.53	0.51	0.75	79.63	**0.59**	0.86	**68.37**	**0.43**	64.49	0.39	62.17	0.34	63.58	0.40
DS5.2	**80.88**	**0.62**	**0.88**	**80.38**	**0.61**	0.83	76.71	0.53	0.79	78.06	0.56	0.84	**68.83**	**0.44**	65.46	0.40	61.87	0.35	63.70	0.40
DS6.2	**80.97**	**0.62**	**0.89**	**80.4**	**0.61**	0.83	77.3	0.55	0.79	80.03	0.6	0.86	**68.78**	**0.44**	65.47	0.40	63.42	0.37	64.46	0.42

IV. . CONCLUSION

It is possible to derive sensitive predictive models for chronicity groups in terms of CGR from clinical data routinely acquired for administrative purposes. The main strength of the proposed classifiers is the readily availability of the information used to build the models, given the fact that the use of the Minimum Basic Data Set, prescription data and CRGs is very widespread amongst the healthcare institutions not only in Spain but also worldwide and little pre-processing is needed to prepare these data.

We have found that pharmaceutical information in consonance with the findings obtained by de Jonge et al. [22]and Higdon et al.[23] is more relevant than diagnosis in terms of discriminant power to predict chronicity evolution with our approach. However, only a very preliminary approach to aggregating diagnosis data in a way that is more meaningful for machine learning has been presented here. We plan to use sparse Principal Component Analysis to reduce the dimensionality of the patient vs. diagnosis matrix and to cluster different diagnoses using language processing methodologies, specifically finding an appropriate indicator of distance among diagnoses based on Jaro distance.

An important limitation of the work is that we only addressed patients that are already identified as chronic. For further research we plan to include a random subgroup of non-chronic patients to see if it is possible to predict those patients that will move from group Healthy to any CGR group indicating chronicity.

REFERENCES

[1] Preventing Chronic Diseases, a vital investment. World Health Organization, Geneva, Switzerland, 2005.

[2] JS. Hughes, RF Averill et al. (2004).Clinical Risk Groups (CRGs): a classification system for risk adjusted capitation-based payment and health care management. Medical Care. 42.

[3] HC Koh, G Tan. (2011). Data mining applications in healthcare. Journal of Healthcare Information Management. 19(2).

[4] SH Liao, PH Chu, PY. Hsiao. (2012). Data mining techniques and applications - A decade review from 2000 to 2011. ESWA. 39(12).

[5] G Phillips-Wren, P Sharkey, SM Dy. (2008). Mining lung cancer patient data to assess healthcare resource utilization. ESWA. 35(4).

[6] W Dai, TS Brisimi et al. (2014). Prediction of hospitalization due to heart diseases by supervised learning methods. Int J Med Info. 84(3).

[7] A Azari, VP Janeja, A Mohseni. 2012. Healthcare Data Mining: Predicting Hospital Length of Stay. Int J Knowledge Discovery in Bioinformatics. 3(3).

[8] I Cho, I Park et al. (2013). Using EHR data to predict hospital-acquired pressure ulcers: A prospective study of a Bayesian Network model. Int J Med Info. 82(11),

[9] P de Toledo, PM Rios et al. (2009). Predicting the outcome of patients with subarachnoid hemorrhage using machine learning techniques. IEEE T Info Tech Biomedicine. 13(5).

[10] A Çakir, B Demirel. (2011). A software tool for determination of breast cancer treatment methods using data mining approach. Journal of Medical Systems.35(6).

[11] I Kamkar, SK Gupta, et al. (2014). Stable feature selection for clinical prediction: Exploiting ICD tree structure using Tree-Lasso. Journal of Biomedical Informatics. 53,

[12] A Geissbuhler, A Leese, et al. (2013). Trustworthy reuse of health data: a transnational perspective. Int J Med Info. 82(1).

[13] N. Japkowicz. (2005). Learning from imbalanced data sets: A comparison of various strategies. AAAI workshop on learning from imbalanced data sets.

[14] J Cohen. (1960). A coefficient of agreement for nominal scale. Educational and Psychological Measurement. 20 (1).

[15] IH Witten, E Frank. Data Mining: Practical machine learning tools and techniques, 2nd Edition. Morgan Kaufmann, 2005.

[16] L Breiman. (1996). Bagging predictors. Machine learning. 24(2).

[17] E Frank, I Witten. Generating Accurate Rule Sets Without Global Optimization, 15th International Conference on Machine Learning, San Francisco, CA, 1998.

[18] WW Cohen. Fast effective rule induction. 12th International Conference on Machine Learning, Lake Tahoe, CA, 1995.

[19] J R Quinlan. C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann, 1993.

[20] D Heckerman, D Geiger, DM Chickering. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. Machine learning. 20(3)

[21] JR Quinlan. C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann, 1993.

[22] P de Jonge, I Bauer et al. (2003). Medical inpatients at risk of extended hospital stay and poor discharge health status. Psychosomatic medicine. 65(4).

[23] R Higdon, E Stewart et al. (2013). Predictive Analytics In Healthcare: Medications as a Predictor of Medical Complexity. Big Data. 1(4).