Population Health Management exploiting Machine Learning to identify High Risk Patients

Silvia Panicacci, Massimiliano Donati, Luca Fanucci
Dept. of Information Engineering,
University of Pisa,
Pisa, Italy
s.panicacci@studenti.unipi.it
massimiliano.donati@for.unipi.it
luca.fanucci@unipi.it

Irene Bellini, Francesco Profili, Paolo Francesconi
Agenzia Regionale Sanità,
ARS,
Florence, Italy
irene.bellini@ars.tocana.it
francesco.profili@ars.toscana.it
paolo.francesconi@ars.toscana.it

***Abstract*— Population aging and the increase of chronic conditions incidence and prevalence produce a higher risk of hospitalization or death. This is particularly high for patients with multimorbidity leading to a great consumption of resources. Identifying as soon as possible high-risk patients becomes an important challenge to improve health care service provision and to reduce costs. Nowadays, population health management, based on intelligent models, can be used to assess the risk and identify these "complex" patients. The aim of this study is to validate machine learning algorithms (Naïve Bayes, Cart, C5.0, Conditional Inference Tree, Random Forest, Artificial Neural Network and LASSO) to predict the risk of hospitalization or death starting from administrative and socio-economics data. The study involved the residents in the Local Health Unit of Central Tuscany.**

Keywords—population health management; machine learning; decision support system;

I. INTRODUCTION

Nowadays, the population in industrialized countries is getting older and older and the number of people aged 65+ years is expected to grow over the next decades, becoming around the 30% of the overall population by the 2060. Additionally, increases of more than 50% are projected in the number of older people affected by most relevant individual diseases (e.g. hypertension, diabetes, stroke, respiratory, etc.) and multi-morbidities over the next 20 years [1]. The provision and funding of the health care services for this growing group of "complex" patients, with one or more chronic conditions, have become an important challenge.

The implementation of the health care model for chronic patients is mainly in charge of the general practitioners (GPs). They usually react to patients' symptoms prescribing medical analysis or laboratory exams, often involving medical specialists, in order to be able to make diagnosis and decisions. The poor propensity of that model for prevention and early diagnosis often leads to ineffective and inefficient treatments [2]. In particular, it leads to increases of risk of hospitalization, length of stay, readmission and mortality, raise of healthcare costs and reduction of quality of life. For these reasons, the health care model is evolving from a reactive to a proactive approach between the healthcare staff and the patients, with the latter becoming an integral part of the care process.

The correct identification of the patients to be enrolled in such a proactive model of care is crucial to treat them with the most appropriate care plan and, in general, to improve the allocation of the available resources. The choice operated by the doctors can be biased or non-objective [3]. In addition, the selection of the cohort of patients to be monitored could be a difficult work for the GPs alone. Firstly, there are many variables to be considered: besides the medical situation, also biology/genetics, socio-economic factors, culture, environment and behaviour are some determinants of health [4]. Secondly, for each patient a huge amount of data could be available, also considering the emerging wearable and IoT medical devices, the telemedicine services and the digitalization of the informative flows.

Population health management can be very useful to identify the target patients. It is intended as a risk assessment process for defining patients cohorts and stratifying members by the risk of preventable hospitalizations in order to deliver specific treatment programs according to the individual needs, with the final aim of improving the health outcomes [5]. Such a process is based on big data analysis techniques.

There are several institutions and companies which are studying and testing models to support the GPs in selecting patients for specific care programs or to predict the risk of hospitalization or death. The existing models are based on different approaches, from statistics to machine learning, and they use administrative and/or socio-economic data. Statistical models are the most used so far. For example, in Tuscany region [6], an ad-hoc algorithm based on resource consumptions is used to identify complex patients, starting from administrative data; in Veneto region [7], the ACG system is used to stratify the population; in Emilia-Romagna region [8], logistic regression models are applied in order to predict risk of hospital admissions or death. The same methods are used in some countries all around the world such as Sweden [9], Germany [10], Holland [11], Canada [12] and the United Kingdom [13]. Thanks to the growing computational power, machine learning methods are gradually replacing statistical methods in this field, because of their capability of analysing huge amount of data and learn from experience. Some studies were conducted in the USA, using machine learning algorithms to predict future healthcare expenditures. In that case, since the healthcare expenses are chargeable to the patients themselves through their insurance companies or as out-of-pocket-expenses, the predictive models become important not only for the country, to organize the available resources, but also for the insurance companies, to provide proper policy to their clients and to calculate the insurance premium for the following year [14]-[16]. Even if the Italian Healthcare System is very different from the American one (almost free for the chronic patients), the same algorithms used in the USA to predict costs can be applied in Italy to predict risks, according to the consideration that high-cost patients inevitably correspond to high-risk patients.

The aim of this study is twofold. First, to assess the performances of some machine learning algorithms: Naïve Bayes [17], decision trees such as CART [18], C5.0 [19] and Conditional Inference Tree [20], Random Forest [21], Artificial Neural Network [22] and LASSO [23], the same used in the literature [14]-[16], to predict avoidable hospital admissions or death and to identify the involved patients. Secondly, to select the subset of the most important features to be considered for the patient identification, to increase the speed of the analysis of the population. The final goal will be to develop a first level screening tool to identify high-risk patients, the ones that the GPs should monitor with specific treatment programs, in order to reduce the hospitalization rate and/or postpone death.

II. Methods

Starting from administrative and socio-economic available data, different machine learning algorithms were used to solve the binary classification problem of identifying in the initial population those patients with avoidable hospitalizations or death during the following year. In particular, this study used data collected from 2010 to 2014, to make predictions for 2015. For each algorithm, the tuning of the parameters was done with the goal of maximizing the Positive Predictive Ratio (PPR) [24]. This metric was already used for the algorithm currently in use in Tuscany region.

The data mining process included the data pre-processing, in which the dataset was constructed and analyzed, the training of the algorithms and finally the evaluation of the results [25]. These phases will be described in the following paragraphs.

A. *Data Pre-Processing*

The source of the data used as input for the algorithms was the mARSupio database of the Agenzia Regionale Sanità (ARS), in Florence, Tuscany, Italy [26]. Here, patients privacy is protected, since personal data are hidden and each patient is identified by the IDUNI (a univocal identification code of 24 characters) [27]. Data in mARSupio are collected from the principal informative flows coming from the Tuscany Regional Health Services and from the national ISTAT census on the resident population: data coming from the hospitals (i.e. diagnoses and procedures), from the outpatients (i.e. assistive, diagnostic and rehabilitation performances), from the pharmacies (e.g. prescribed drugs, etc.), data regarding the exonerations (both for income or diagnosis) and data coming from the last census (2011). Usually these administrative flows are almost complete because Regional Health System covers the expenses, subjected to code reporting. Providers receive a specific reimbursement for each codified supply. The complete architecture of the system for the collection of data is shown in Fig. 1.

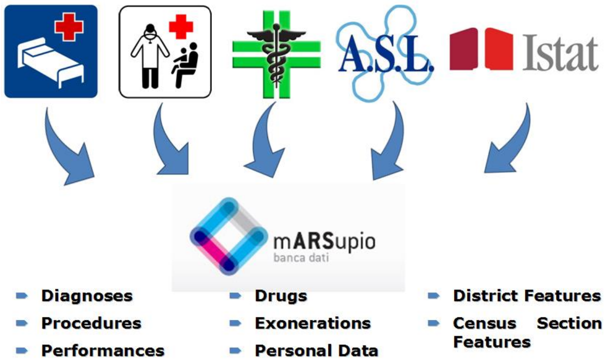

Fig. 1 Architecture of the system for the collection of data

The people involved in the study were all the residents in the Local Health Unit of Central Tuscany, alive on 1st January 2015, who have lived in Tuscany at least the 80% of the days between 2010 and 2014. These restrictions were due to the fact that ARS collects medical information only for the residents in Tuscany and that, before 2010, the available data were partially incomplete. The total number of people considered in the study was 1553397.

The complete set of features was selected as follows:

- the diagnoses were grouped by Aggregated Clinical Codes (ACC) [28] and, for each ACC two different variables were considered: the number of admissions and the number of days of hospitalization, for a total of 566 attributes (ACC are 283);
- the procedures were grouped by ACC and the number of procedures done for each ACC was considered as a variable, for a total of 231 attributes;
- all the types of performances were divided into 76 groups and the number of performances done for each group was considered as a variable, for a total of 76 attributes;
- concerning the drugs, they were grouped according to the third level of the Anatomical Therapeutic Chemical classification system (ATC3) [29], which indicates the therapeutic/pharmacological subgroup of the drug itself, and the number of drugs taken for each ATC3 code was considered as a variable, for a total of 265 attributes;
- the exonerations were partitioned into 28 groups and each group was considered as a variable assuming the following values: ‘0’ meant that the patient has never had an exoneration for that group, ‘1’ meant that the patient has an exoneration for that group, ‘2’ meant that the patient was exonerated for that group in the past. The exonerations provided a total of 28 attributes;
- age and gender were included (2 attributes);
- regarding the municipality of residence, the density of the district, the characteristic of inner area (with values 0, 1, 2, 3, and 4 respectively for centre, middle, belt,

outlying, and outermost) and the classification as fragile area (0 for non-fragile and 1 for fragile) were selected, for a total of 3 attributes;

- concerning the census section, the dependency index, the median level of education (with values 4, 3, 2, 1 and 0 respectively for degree, high school, secondary school, primary school and nothing), the median marital status (0 for single, 1 for married, 2 for divorced and 3 for widowed), the percentage of the working population, the percentage of strangers, the median number of family members and the percentage of rented houses were chosen, for a total of 7 attributes.

Therefore, the total number of features was 1178 and they included both administrative and socio-economic data.

Because of some input attributes referred to the census sections (education, marital status, etc.) presented missing values, the adopted strategy was to replace them with the mean value for continuous variables and the median value for categorical variables.

The outcome measure was a dichotomous variable, where 'B' value meant that the patient will have an avoidable admission or/and will die the following year, while 'G' value meant that the patient will not have an avoidable hospitalization nor die the following year. Since the historical variables were taken until 2014, the predictions have been made for 2015. The problem was thus a supervised binary classification, with the class 'B' considered as the positive one.

In the final dataset, each person was represented by a row and each variable was a column. Its dimension, after the deletion of duplicated rows, was 1529714x1179, considering also the output variable.

B. Modelling

The whole dataset was split into the training set (70%) and the test set (30%). Therefore, the training set had 1070801 samples, while the test set was composed by 458913 samples.

Since the initial dataset was very unbalanced towards the negative output class 'G' (positive class 'B' occurred in less than the 1.5% of the samples), the training set was balanced taking one random sample every twenty samples belonging to the 'G' group. At the end of this process, the training rows were reduced to 67978, where the positive samples were the 22.36% of the total ones (TABLE I). On the contrary, the test set was not modified, in order to evaluate the performances on a real sample of the Tuscan population.

Some different machine learning algorithms were tested in order to find the best model in terms of PPR and, alternatively, F1-Score metrics [30]. For each algorithm, the tuning of the parameters was done using the balanced training set. The models were trained in 10-fold cross-validation with grid search, in order to choose the best combination of parameters, starting from a pre-decided set. The aim was the maximization of the PPR. The results of the training phase highlighted that:

- Naïve Bayes (NB) [17] performed in the same way with and without Laplacian smoothing;

TABLE I POSITIVE 'B' AND NEGATIVE 'G' CLASSES DISTRIBUITON FOR THE TWO DIFFERENT TRAINING SETS AND FOR THE TEST SET

Class	**Initial Training Set**	
	Frequency	***Percentage (%)***
B	15198	1.419311
G	1055603	98.580689
	Balanced Training Set	
	Frequency	***Percentage (%)***
B	15198	22.35723
G	52780	77.64277
	Test Set	
	Frequency	***Percentage (%)***
B	6513	1.419223
G	452400	98.580777

- CART decision tree [18] performed best with the complexity parameter (cp) equal to 0.0001 and with the minSplit (minimum number of samples in a node to attempt a split) equal to 200;

- For C5.0 decision tree [19], the confidence factor did not affect the results (except the computation time), while the best minCases (smallest number of samples that must be put in at least two of the splits) was 50;

- Conditional Inference Tree (CTree) [20] reached the best PPR when minSplit was equal to 200, taking fixed at 0.05 the value that must be exceeded to implement a split;

- Random Forest (RF) [21] performed best with 1000 trees and 34 variables randomly selected as candidates at each split round (mtry);

- Artificial Neural Network (ANN) [22] was built with a single hidden layer and the best number of hidden neurons (from 2 to 15) was 10;

- LASSO's [23] best lambda parameter was 0.001520083.

In order to speed-up and optimize memory consumption and execution time of the training process, the Boruta algorithm was run on the entire training set to select the most important features to predict the outcome [31]. It performed a top-down search for the most predictive attributes, using random forests, comparing each original variable's importance to the importance reachable when that variable is randomly shuffled, and iteratively deleting the less relevant attributes. The Boruta algorithm confirmed a group of 280 attributes and rejected 898 variables of the complete set of 1178 features, producing a reduced training set:

- both age and sex were selected;

- among the 566 features regarding the diagnoses, 99 variables were chosen (both number of admissions and number of days were usually selected for the same ACC);

- 58 of 231 attributes were confirmed for the procedures;

- about the performances, 35 features on a total of 76 were kept in the subset;

- 80 of 266 variables regarding the drugs were selected;
- as regard the exoneration groups, 6 attributes on a total of 28 were chosen;
- the features not referred to the single person, but instead to the municipality of residence and to the census section, were all rejected.

After this step, the tuning of the parameters was done again with the resulting reduced balanced dataset. The best combinations of parameters were the same of the complete dataset for all the models, with some exceptions: CART performed best with the same cp (0.0001) and minSplit equal to 100 (instead of 200), ANN's best number of hidden neurons became 13 and LASSO's best lambda value was 0.0004977588.

All algorithms ran on a Linux server with 64 GB of RAM, using dedicated libraries and a program written in R language.

C. *Evaluation of the results*

The performances were evaluated using the PPR and the F1-Score metrics. On the contrary, accuracy (the proportion of correctness in a classification system) was not considered as a good metric in this context, since it assumes that the a priori probability of the classes are constant and almost balanced [32]. This is not the case, because the classes distribution was very skewed and the test set was really unbalanced. Using PPR and F1-Score metrics, we avoided to incur the "accuracy" paradox.

The main metrics used to compare the performance are:

$$\mathrm{PPR} = \frac{PPV}{(1 - NPV)}$$

$$\mathrm{F1Score} = \frac{2 * (PPV * Sensitivity)}{(PPV + Sensitivity)}$$

where Positive Predictive Value (PPV), Negative Predictive Value (NPV) and sensitivity are defined as:

$$PPV = \frac{TruePositive}{(TruePositive + FalsePositive)}$$

$$NPV = \frac{TrueNegative}{(TrueNegative + FalseNegative)}$$

$$Sensitivity = \frac{TruePositive}{(TruePositive + FalseNegative)}$$

Another interesting metric is specificity, defined as:

$$Specificity = \frac{TrueNegative}{(TrueNegative + FalsePositive)}$$

III. Results

The results concern the application of the algorithms to the test set. Comparing the performance achieved by all the models in terms of both PPR and F1-Score, it is possible to find that Random Forest and LASSO behave better than the others for the given classification problem. Such a result is the same when they are trained both with the complete balanced dataset or the balanced dataset with a reduced number of features. Fig. 2 shows the results in terms of performance of the algorithms trained with the reduced dataset. In case of the complete training dataset, the performance are in general slightly worse, except for the ANN. TABLE II summarizes the final performance evaluation results considering the two golden metrics PPR and F1-Score.

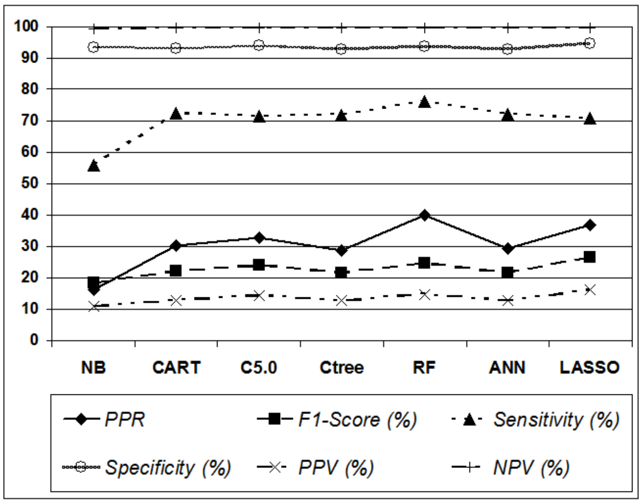

Fig. 2 Comparison of the performance of all the models trained with the reduced balanced dataset, using the best parameters

IV. Discussion

The 280 features selected by the Boruta algorithm as the most predictive variables are suitable for the identification of the target chronic patients. For example, regarding the diagnoses, both the number of admissions and the number of days of hospitalization are usually selected for the same ACC, meaning that these variables are important together (i.e. one hospital admission of 20 days can be different from 4 hospitalizations of 5 days each). Additionally, when only one of the two variables is confirmed, the number of days is chosen, meaning that this variable is more relevant with respect to the number of admissions. Moreover, among the biographic and socio-economic attributes, only the ones referred to the single person, namely age and gender, are selected, while the ones connected to the municipality of residence and to the census section are not so relevant because they refer to a group of people with different characteristics, and thus excluded. The variables like the level of

TABLE II PPR and F1-Score achieved by all the models vs training balanced datasets

Model	**Complete Dataset**		**Reduced Dataset**	
	PPR	***F1-Score (%)***	***PPR***	***F1-Score (%)***
Naïve Bayes	11.97	15.17	16.12	18.18
CTree	28.63	21.44	28.92	21.44
ANN	29.60	22.86	29.23	21.48
CART	29.94	21.89	30.24	21.91
C5.0	32.43	23.79	32.74	23.87
LASSO	36.64	26.36	36.75	26.39
Random Forest	37.90	25.34	39.83	24.54

education, the marital status or the income are very important as determinants of health [4], but unfortunately they are difficult to collect for every patient and approximated with the values from the census section.

The reduced dataset contains less than the 25% of all the features. Those are the most significative for the classification. In this case, obviously, the training time, the processing time and also the required memory decrease for all the models, as shown in TABLE III. These improvements in computational time and memory requirements do not account the performance of the classification, that are even lightly better with the reduced dataset. The only exception is represented by the ANN. In view of the future use of these methods to extract in advance the lists of high-risk patients, the fastest and lightest (in terms of memory) algorithms are the best candidates. These algorithms will allow for suggesting to GPs which are the patients that require specific programs of care to avoid or postpone hospitalizations or death.

Among all the tested algorithms, Random Forest and LASSO result the best models for the target problem; the first for PPR and the second for F1-Score. Analyzing their confusion matrixes and the other metrics (TABLE IV and Fig. 2), both algorithms feature high sensitivity (76.02% and 70.50%, respectively) and high specificity (93.62% and 94.75%, respectively). This represents a point of strength of these models, since they perform a very high prediction of true positives and true negatives. Conversely, there is a high number of false positives (about 4/5 of the samples classified as positive), due to the very low positive outcome prevalence (less than 1.5%). This is not considered a major problem, because the tool is projected to be used for a first level screening, and so it is better to include more people than necessary in the positive class rather than exclude patients having really need of specific treatments. On the other hand, it is important that the false negatives are as few as possible with respect to the total of people classified as negatives. This is confirmed by the high NPV of the algorithms (99.63% and 99.55%, respectively).

TABLE III COMPUTATIONAL REQUIREMENTS VS TRAINING BALANCED DATASETS FOR ALL THE MODELS

Model	**Dataset**	**Memory (bytes)**	**Pre-Processing Time (seconds)**	**Processing Time (seconds)**
Naïve Bayes	Complete	1058000	3.214	10693.22
	Reduced	253644	1.144	2313.150
CART	Complete	11022800	953.224	50.327
	Reduced	5075008	287.455	10.698
CTree	Complete	141570872	225.069	1022.345
	Reduced	139193296	31.277	31.810
C5.0	Complete	6586056	295.188	438.140
	Reduced	571232	53.992	105.938
ANN	Complete	15446976	2179.410	76.909
	Reduced	9318256	505.687	16.996
Random Forest	Complete	864979912	229.007	178.983
	Reduced	685776728	179.605	148.482
LASSO	Complete	561248	692.801	10.672
	Reduced	159048	164.922	2.839

TABLE IV CONFUSION MATRIXES OF RANDOM FOREST AND LASSO MODELS, TRAINED WITH THE REDUCED DATASET

	Random Forest		**LASSO**	
Prediction	***Reference***		***Reference***	
	B	G	B	G
B	4951	28879	4614	23845
G	1562	423521	1899	428555

During the test, the probability threshold for discriminating the negative class from the positive one was set to 0.5. Raising this threshold the number of false positives decreases, because the models tend to predict very high-risk patients. On the other hand, also the true positives could decrease and the false negatives increase, worsening the classification performance. Since GPs usually have limited resources dedicated to follow chronic patients, this scenario could be taken into consideration because it could produce a list with a limited number of patients. The final tool would be used to select the highest risk patients maximizing sensibility and, among them, the GPs should make another selection (second level screening) to restrict again the number, taking into consideration also behavioral, social or other factors that may influence people’s health.

The algorithm currently used in Tuscany [6] to identify high-risk patients in the group of 60+ aged people uses a limited number of variables and only administrative data. It reaches a PPR near 7 in predicting all the hospital admissions (not only the avoidable ones) and approximately 6 for the death during the following year. With machine learning methods, it becomes possible to work with large amount of data and features, also outperforming the results of the previous methods. For example, Random Forest and LASSO have a PPR greater than 36 for the prediction of avoidable hospitalizations or death.

In order to further improve the performance of the classification, Random Forest, LASSO and C5.0 algorithms could be combined together, along with a voting logic: the final predicted class is the one “voted” by the majority of the models.

One of the main limitation for the application of these models is the lack of socio-economic and behavioral data for each patient. These data, in fact, could increase the predictability of the models, since the 80% of factors that affect the health outcome and the clinical phenotype are associated to health behaviors, social and economic factors and physical environment [33]. Moreover, the use of only administrative data leads to a decrease of classification performance because the models are prone to generalizations [34].

V. CONCLUSION

Population is getting older and the number of people suffering from multiple chronic conditions is increasing. For GPs and healthcare providers in general, it becomes crucial to identify as soon as possible the complex patients to treat them with specific program of care, in order to reduce or postpone hospitalizations or death. A possible solution to support this selection process is the development of population health management tools based on machine learning methods.

This paper presents the performance evaluation of several machine learning algorithms to solve the binary classification problem of identifying high-risk patients in the population, by analyzing different sources of administrative and socio-economic data. Among the tested algorithms, the best models in terms of PPR and F1-Score result to be Random Forest and LASSO. These models outperform the methods currently used in Tuscany region for the identification of high-risk patients (7 vs 39 for the PPR metrics). The main limitation of this approach is a quite high number of false positives. This does not represent an issue since these tools are considered for a first level of screening, and the resulting list of patients is expected to be further analyzed by the GPs to extract the final list of patients to be enrolled in dedicated treatment programs.

REFERENCES

[1] Andrew Kingston, Louise Robinson, Heather Booth, Martin Knapp, and Carol Jagger, "Projections of multi-morbidity in the older population in England 2035: estimates from the Population Ageing and Care Simulation (PACSim) model", *Age and Ageing*, https://doi.org/10.1093/ageing/afx201.

[2] Progetto CCM 2015 Paziente Complesso.

[3] Efrat Shadmi and Tobias Freund, "Targeting patients for multimorbid care management interventions: the case for equity in high-risk patient identification", *International Journal for Equity in Health*, vol. 12, article 70, 2013.

[4] Monika M. Safford, Jeroan J. Allison, and Catarina I. Kiefe, "Patient Complexity: More Than Comorbidity. The Vector Model of Complexity", *Journal of General Internal Medicine*, vol. 22, no. 3, pp. 382–390, 2007.

[5] How to get started with a population health management program, https://healthitanalytics.com/features/how-to-get-started-with-a-population-health-management-program.

[6] Irene Bellini, Valentina Barletta, Francesco Profili, Alessandro Bussotti, Irene Severi, Maddalena Isoldi, Maria Bimbi, and Paolo Francesconi, "Identifying High-Cost, High-Risk Patients Using Administrative Databases in Tuscany, Italy", *BioMed Research International*, 2017. https://doi.org/10.1155/2017/9569348.

[7] Measuring the territory to increase equity and efficiency of sanitary services: the ACG Project, http://acg.regione.veneto.it/risultati-preliminari/final-report-2012/.

[8] Daniel Z Louis, Mary Robeson, John McAna, Vittorio Maio, Scott W Keith, Mengdan Liu, Joseph S Gonnella, and Roberto Grilli, "Predicting risk of hospitalisation or death: a retrospective population-based analysis", *BMJ Open*; 4:e005223. doi:10.1136/ bmjopen-2014-005223.

[9] Gerd Fridh and Ingvar Ovhed, "Validating the Johns Hopkins ACG Case-Mix System of the elderly in Swedish primary health care Anders Halling", *BMC Public Health*, 6:171, 2006.

[10] Tobias Freund, Cornelia Ursula Kunz, Dominik Ose, Joachim Szecsenyi, and Frank Peters-Klimm, "Patterns of Multimorbidity in Primary Care Patients at High Risk of Future Hospitalization", *Population Health Management*, vol. 15, no. 2, pp. 119–124, 2012.

[11] Mieke Rijken, Marion van Kerkhof, Joost Dekker, and Francois G. Schellevis, "Comorbidity of chronic diseases. Effects of disease pairs on physical and mental functioning", *Quality of Life Research*, vol. 14, no. 1, pp. 45–55, 2005.

[12] Lisa M. Lix, Joykrishna Sarkar, MSc, Sharon Bruce, and T. Kue Young, "Ethnic and Regional Differences in Prevalence and Correlates of Chronic Diseases and Risk Factors in Northern Canada", *Preventing Chronic Disease*, vol. 7, no. 1, article A13, 2010.

[13] John Billings, Theo Georghiou, Ian Blunt, and Martin Bardsley, "Choosing a model to predict hospital admission: an observational study of new variants of predictive models for case finding", *BMJ Open* 2013;3:e003352. doi:10.1136/bmjopen-2013003352.

[14] Bibudh Lahiri and Nitin Agarwal, "Predicting Healthcare Expenditure Increase for an Individual from Medicare Data".

[15] Shanu Sushmita, Stacey Newman, James Marquardt, Prabhu Ram, Virendra Prasad, Martine De Cock, and Ankur Teredesai, "Population Cost Prediction on Public Healthcare Dataset".

[16] Seyed Abdolmotalleb Izad Shenas, Bijan Raahemi, Mohammad Hossein Tekieh, Craug Kuziemsky, "Identifying high-cost patients using data mining techniques and a small set of non-trivial attributes".

[17] Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", *Elsevier*, chapter 6.4, 2006.

[18] Leo Breiman, Jerome Friedman, Charles J. Stone, R. A. Olshen, "Classification and Regression Trees", *Wadsworth International Group*, 1984.

[19] J. Ross Quinlan, "C4.5: Programs for Machine Learning", *Morgan Kauffamnn Publishers*, 1993.

[20] Torsten Hothorn, Kurt Hornik, Achim Zeileis, "Unbiased Recursive Partitioning: A conditional Inference Framework", 2006.

[21] Adele Cutler, D.Richard Cutler and John R. Stevens, "Random Forests", *Machine Learning*, 2011.

[22] Sun-Chong Wang, "Artificial Neural Network", *The Springer International Series in Engineering and Computer Science*, vol. 743, 2003.

[23] Robert Tibshirani, "Regression Shrinkage and Selection via the Lasso", *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no.1, pp 267-288, 1996.

[24] New patient-oriented summary measure of net total gain in certainty for dichotomous diagnostic tests, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1635036/.

[25] David Hand, Heikki Mannila and Padhraic Smyth, "Principles of Data Mining: a comprehensive, highly technical look at the math and science behind extracting useful information from large databases", *The MIT Press*, 2001.

[26] mARSupio database, https://www.ars.toscana.it/marsupio/database/.

[27] Italian Law, no. 675/1996, Tutela delle persone e di altri soggetti rispetto al trattamento dei dati personali, [Protection of persons and other subjects with regard to personal data processing], http://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/28335.

[28] Aggregated Clinical Codes (ACC), http://www.salute.gov.it/imgs/C_17_pubblicazioni_1006_allegato.pdf

[29] Anatomical Therapeutic Chemical classification system (ATC), https://www.whocc.no/atc/structure_and_principles/.

[30] Accuracy, precision, recall and F1 score: interpretation of performance measures, http://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/.

[31] Miron B. Kursa, Aleksander Jankowski, Witold R. Rudnicki, "Boruta – A System for Feature Selection", *Fundamenta Informaticae*, vol. 101, pp. 271-285, 2010.

[32] Why accuracy alone is a bad measure for classification tasks, and what we can do about it, https://tryolabs.com/blog/2013/03/25/why-accuracy-alone-bad-measure-classification-tasks-and-what-we-can-do-about-it/.

[33] Population Health Management: Systems and Success, https://www.healthcatalyst.com/population-health/.

[34] Challenges of Applying Predictive Analytics to Population Health, https://healthitanalytics.com/news/challenges-of-applying-predictive-analytics-to-population-health.