*MEDINFO 2017: Precision Healthcare through Informatics*
*A.V. Gundlapalli et al. (Eds.)*

*doi:10.3233/978-1-61499-830-3-366*

# Development and Validation of Various Phenotyping Algorithms for Diabetes Mellitus Using Data from Electronic Health Records

**Santiago Esteban, Manuel Rodríguez Tablado, Francisco Peper, Yamila S. Mahumud, Ricardo I. Ricci, Karin Kopitowski, Sergio Terrasa**

*Family and Community Medicine Division, Hospital Italiano, Buenos Aires, Argentina*

**Abstract**

*Precision medicine requires extremely large samples. Electronic health records (EHR) are thought to be a cost-effective source of data for that purpose. Phenotyping algorithms help reduce classification errors, making EHR a more reliable source of information for research. Four algorithm development strategies for classifying patients according to their diabetes status (diabetics; non-diabetics; inconclusive) were tested (one codes-only algorithm; one boolean algorithm, four statistical learning algorithms and six stacked generalization meta-learners). The best performing algorithms within each strategy were tested on the validation set. The stacked generalization algorithm yielded the highest Kappa coefficient value in the validation set (0.95 95% CI 0.91, 0.98). The implementation of these algorithms allows for the exploitation of data from thousands of patients accurately, greatly reducing the costs of contructing retrospective cohorts for research.*

***Keywords:***

Diabetes Mellitus, Algorithms, Precision Medicine

## Introduction

Recently, the progression towards precision medicine [1] has sought the development of large databases, allowing assessment of the impact of risk factors or treatments in specific subpopulations. This is usually a problem for classical cohorts, given the difficulty of enrollment and follow-up of a large enough number of patients [5]. Even more difficult is the situation for developing countries, given the usual lack of funds for local research [3].

Electronic health records (EHR) have been proposed as a solution to these two costs problems [10].

Phenotyping algorithms allow, through the combination of different variables extracted from the EHR, classifying patients according to their particular phenotype [8; 11]. Ideally, these algorithms must be validated and a metric should be estimated (accuracy, sensitivity and specificity, coefficient Kappa, F1 score, positive and negative predictive values) of the ability to classify patients compared to a gold standard. This facilitates the classification of large numbers of patients without the intervention of a human.

Boolean or rule-based algorithms are a common strategy for developing these algorithms. A different approach is the development of learners based on statistical learning, such as logistic regression or more recent methods such as decision trees, neural networks or support vector machines. The different families of algorithms explore the multidimensional space of data in different ways so it can be beneficial to combine them. One way to do this is through stacked generalization. This methodology, described by Wolpert [4] for classification problems and by Breiman [7] for regression problems, seeks to improve the predictive power of the individual algorithms by developing a meta-learner incorporating the predictions of all algorithms as input, combining them, and then issuing a final prediction.

Our objective is to compare the performance of different classification strategies (only using standardized problems, rules-based algorithms, statistical learning algorithms and stacked generalization), for the categorization of patients according to their diabetic status (diabetics, not diabetics and inconclusive; diabetes of any type) using information extracted from EHR.

## Methods

### Study population

Patient information was extracted from the EHR of the Hospital Italiano in Buenos Aires, Argentina.

In order to have a training and a validation dataset, two samples of patients from different years (2005-2015; total n = 2463) were extracted. The only inclusion criterion was age (≥40 <80 years old by 1/1/2005 and by 1/1/2015 for each sample). The sampling was carried out using simple randomization. The training set (2005) featured 1663 patients. The validation set (2015) represented roughly 33% of the total sample (n = 800).

### Feature extraction

Six variables were extracted: No. of standardized problems related to Diabetes Mellitus (DM) (inpatient, outpatient and emergency department codes); No. of filled oral hypoglycaemic or insulin prescriptions; No. of outpatient fasting glucose (FG) measurements ≥126 mg/dl; No. of outpatient FG measurements <126 mg/dl; No. of HbA1c measurements ≥6.5%; No. of HbA1c measurements <6.5%. These variables were also used in previous research [6; 12]. Oral glucose tolerance measurements were left out in order to avoid making a diagnosis of gestational diabetes. Random ≥200 mg/dl blood glucose was not considered since it was difficult to establish if it coincided with diabetic symptoms, as indicated by the criteria of the American Diabetes Association (ADA).

**Manual chart review**

Four researchers manually reviewed all records and classified patients, analyzing all available information in the EHR. Patients were classified as:

- Diabetics: The ADA criteria [2] to classify patients as diabetics were used. Also, patients whose records stated that they were diabetics were classified as such.
- Not diabetic: To be classified as a non-diabetic, patients must have at least one FG measurement below 126 mg/dl, and must not have any references in their records regarding being diabetic or fulfill any of the ADA's criteria for DM.
- Inconclusive: Patients without a reference in their EHR regarding their diabetic status, nor a normal FG measurement, were classified as inconclusive. Those who had a single FG value above 126 mg/dl without a subsequent confirmatory measurement were characterized in the same way.

The level of agreement among researchers was assessed using the Kappa coefficient with a value of 0.92 (95% CI: 0.84, 0.99).

**Algorithm development and validation process**

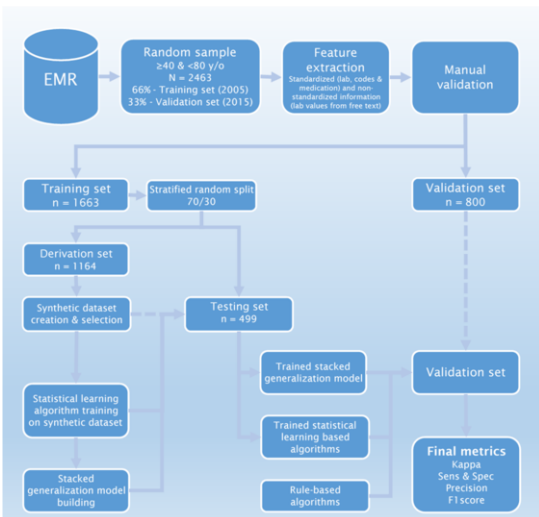


*Figure 1 - General process of developing and validating phenotyping algorithms. EMR: Electronic Medical Records*

***Rules-based algorithms***

In our study, we included two algorithms of this type:

- Classification of patients according to standardized codes: patients were classified as diabetics if they had at least one DM code in their EHR.
- Boolean logic algorithm (Adapted from Kho, Wilke, and Nichols): we used a combination of these three algorithms.

***Algorithms based on statistical learning***

Four of these learners were included individually: multinomial logistic regression, random forests, neural networks and support vector machines with radial kernel. According to their performance on the test set, the best one was evaluated in the validation set.

***The problem of imbalanced datasets***

We opted to use sampling techniques to adjust the imbalance between classes of the dependent variable. To select the best synthetic sampling algorithm, we divided the derivation dataset into a training and a test set. We then used the approach developed by Lopez et al. [9]: from the training set, we generated 19 sets of data by applying a combination of over and under sampling algorithms and analyzed them by means of four algorithms that use different approaches (neural networks, Elastic Net, Gradient Boosting Machine and C5.0). We then applied the trained learners on the test set (which remained unbalanced) and then ranked the datasets according to their performance. The best ranked dataset was used for the training of the statistical learning algorithms.

***Development of the meta-learner***

For the final prediction we selected the Elastic-Net algorithm. As a first step, we discarded those learners with significantly lower performance in the different subsets of the repeated cross-validation (RCV) (set 1). Then, four selection strategies were used: 1. We kept those algorithms whose performance in the different subsets of the RCV were less correlated (Pearson coefficient <0.75; set 2); 2. Using hierarchical clustering (Euclidean distance, complete method), learners were clustered according to their patterns of performance in the subsets of the RCV; the best within each cluster at different height levels were selected (sets 3, 4, 5); 3. We selected those with better performance within each family of algorithms (support vector machines, neural networks, decision trees, instance-based algorithms, algorithms, Bayesian, discriminant analysis, and linear models (set 6). Each of these versions of the meta-learner was evaluated on the test set and the most parsimonious and best performing learner was selected as the final model.

**Validation**

For the validation process, the different algorithms were evaluated in the validation set. The Kappa coefficient was used as the performance metric.

All analyses were performed using R (R Foundation for Statistical Computing, Vienna, Austria.) URL: https://www.R-project.org).

## Results

Table 1 shows the characteristics of the patients included in both datasets. We can observe that patients from the sample of 2015 (validation dataset) generally have a greater number of measurements.

*Table 1 - Characteristics of patients included in the training and validation datasets. DM: Diabetes Mellitus; FG: Fasting glucose.*

| | Training Set (2005) | | | Validation Set (2015) | | |
|---|---|---|---|---|---|---|
| | Non - DM | DM | Unde-fined | Non - DM | DM | Undefined |
| n | 1249 | 121 | 293 | 698 | 58 | 44 |
| Number of DM-related codes (median [IQR]) | 0.00 [0.00, 0.00] | 1.00 [1.00, 1.00] | 0.00 [0.00, 0.00] | 0.00 [0.00, 0.00] | 1.00 [1.00, 2.00] | 0.00 [0.00, 0.00] |
| Number of DM-related prescriptions filled | 0.00 [0.00, 0.00] | 3.00 [0.00, 7.00] | 0.00 [0.00, 0.00] | 0.00 [0.00, 0.00] | 13.50 [1.00, 34.25] | 0.00 [0.00, 0.00] |
| Number of FG ≥126 mg/dl | 0.00 [0.00, 0.00] | 2.00 [1.00, 4.00] | 0.00 [0.00, 0.00] | 0.00 [0.00, 0.00] | 4.50 [2.00, 9.00] | 0.00 [0.00, 0.00] |
| Number of FG <126 mg/dl | 2.00 | 2.00 | 0.00 | 7.00 | 7.50 | 0.00 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | [1.00, 3.00] | [0.00, 4.00] | [0.00, 0.00] | [3.00, 11.00] | [4.00, 15.75] | [0.00, 0.00] |
| **Number of HbA1c ≥6.5%** | 0.00 [0.00, 0.00] | 1.00 [0.00, 2.00] | 0.00 [0.00, 0.00] | 0.00 [0.00, 0.00] | 2.00 [0.00, 7.00] | 0.00 [0.00, 0.00] |
| **Number of HbA1c <6.5%** | 0.00 [0.00, 0.00] | 1.00 [0.00, 2.00] | 0.00 [0.00, 0.00] | 0.00 [0.00, 0.00] | 3.00 [1.00, 6.00] | 0.00 [0.00, 0.00] |
| **Age (mean (sd))** | 60.52 (11.48) | 66.22 (9.09) | 53.94 (10.64) | 60.12 (11.15) | 66.47 (9.89) | 55.00 (10.07) |

The results of all proposed algorithms are presented in figure 2 and table 2.

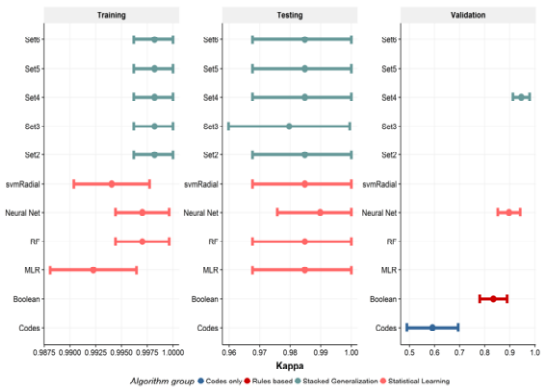


*Figure 2 - Algorithm performance across training, test and validation sets.*

### *Development and selection of the synthetic dataset*

The dataset with best performance was a combination of Synthetic Minority Over-Sampling Technique (SMOTE) and Edited Nearest Neighbors (ENN) and was selected for the training of models. See figure 3.

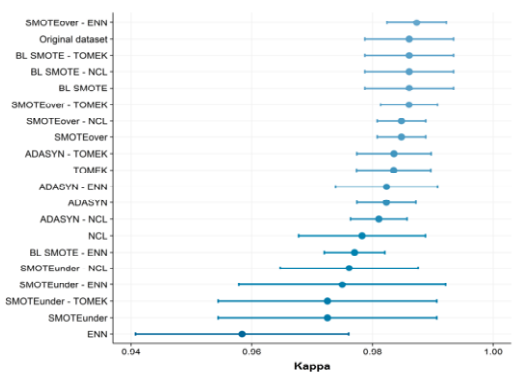


*Figure 3 - Averaged performance of four algorithms trained on different datasets, in the testing set. SMOTE: Synthetic minority over-sampling technique; BL: Borderline; ADASYN: Adaptive synthetic sampling; over: over-sampling; under: undersampling; ENN: Edited nearest neighbors; TOMEK: Tomek links; NCL: Neighborhood cleaning rule.*

### Algorithms based on statistical learning

The network neural presented the best performance in the test set and therefore was evaluated in the validations set (figure 2 and table 2).

### Development of the meta-learner

### *Selection of models with the best performance*

Figure 4 shows the performance of the 16 algorithms under consideration for building the meta-learner. Five algorithms showed Kappa coefficient values clearly below the rest and were excluded. The remaining eleven constitute set 1.

### *Selection of models based on low correlation*

We assessed the level of correlation of performance of the algorithms in the different subsets of RCV (set 1). We detected those correlations greater than 0.75; then the average correlation of both algorithms were compared and the one with the highest mean correlations with all other models was removed. RRF, GBM, EGB - Linear and Elastic Net were removed. The remaining six formed set 2. The results are presented in figure and table 2.

*Table 2 - Performance metrics for four algorithm strategies in the validation set. NDM: Non-Diabetes Mellitus; DM: Diabetes Mellitus; INC: Inconclusive*

| Perf. Measure | Confusion Matrix | | | | Kappa |
|---|---|---|---|---|---|
| | Pred. x Reference | | | | |
| **Codes** | | NDM | DM | INC | 0.59 (0.49, 0.69) |
| | NDM | 697 | 12 | 44 | |
| | DM | 1 | 46 | 0 | |
| | INC | 0 | 0 | 0 | |
| **Boolean** | | NDM | DM | INC | 0.83 (0.78, 0.89) |
| | NDM | 668 | 11 | 19 | |
| | DM | 0 | 58 | 0 | |
| | INC | 4 | 0 | 40 | |
| **Feedforward Neural Net** | | NDM | DM | INC | 0.90 (0.85, 0.94) |
| | NDM | 682 | 0 | 4 | |
| | DM | 16 | 58 | 0 | |
| | INC | 0 | 0 | 40 | |
| **Stacked Generalization (set 4)** | | NDM | DM | INC | 0.95 (0.91, 0.98) |
| | NDM | 693 | 1 | 4 | |
| | DM | 5 | 58 | 0 | |
| | INC | 0 | 0 | 40 | |

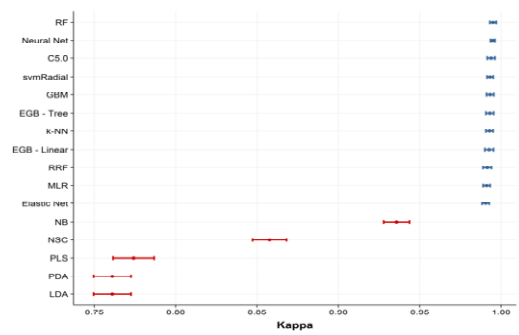


*Figure 4 - Algorithm performance averaged over 50 cross-validation training subsets. RF: Random forests; svmRadial: Support vector machines with radial kernel; GBM: Gradient boosting machine; EGB: Extreme gradient boosting; k-NN: k nearest neighbors; RRF: Regularized random forests; MLR: Multinomial logistic regression; NB: Naïve bayes; NSC: Nearest shrunken centroids; PLS: Partial least squares; PDA: Partial discriminant analysis; LDA: Linear discriminant analysis.*

***Selection based on hierarchical clustering***

Figure 5 shows the dendrogram generated from the hierarchical clustering of the performance of the different algorithms in the subsets of RCV. Three cutoff points were chosen. For cutoff level, the algorithm with best performance by cluster was selected (Set 3: RF, svmRadial, k-NN, GBM; Set 4: 3 Set + MLR; 5 set: Set 4 + Neural Net, EGB - Linear). Each set was then used as input for a different version of the Elastic Net-based meta-learner. Each version was tested in the test set. Set 4 presented the best combination of performance and parsimony and was selected to be applied to the validation set.

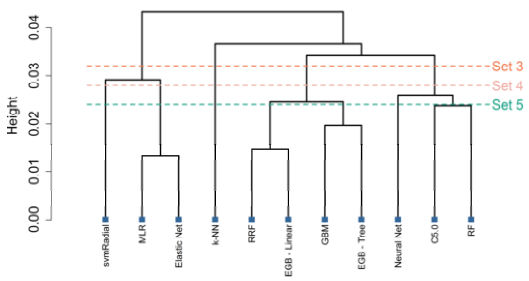


*Figure 5 - Hierarchical clustering of classification algorithms based on their performance in the CV datasets.*

***Salection based on algorithm family***

We selected those algorithms with the best performance within each family, to form set 6 (SVM radial Kernel, neural network with a single layer, EGB - Linear, Random Forests and k-Nearest Neighbors).

***Selection of the best meta-learner***

Finally we compared the performance of different versions of the meta-learner in the test set. The version that used set 4 as input presented the best combination of performance and parsimony. Its ability to classify patients was then evaluated in the validation set (figure 2 and table 2).

## Discussion

Three of the four algorithms evaluated on the validation set showed excellent performance in terms of the Kappa coefficient. Our decision to use this metric above others more commonly used, such as accuracy, was related to the need for high levels of classification within each of the categories given the imbalance of classes in our dataset. This can be affected when the considered metric is accuracy, since it does not consider the agreement for each class but only the level of total error in the confusion matrix.

Each strategy presents pros and cons that are important to consider, since performance is not the only variable to take into account when selecting an algorithm to apply. Algorithms based on rules have the advantage of being simple and easily scalable with minimum processing time. However, we found that their performance is clearly lower than those based on statistical learning and stacked generalization. The neural network showed a high level of optimism (the difference in performance between the training and the validation set). This was less significant for the stacked generalization, which would go in hand with the main objective of this strategy — to reduce overfitting to the training set. The version of the meta-learner that used set 4 as input showed the best metrics of classification in the validation set. Its implementation for research would be helpful, but probably not so for real-time applications given the higher processing time compared to simpler approaches.

## Conclusion

We evaluated the performance of different strategies for the development of diabetes phenotyping algorithms using data extracted from an EHR from Argentina. The stacked generalization strategy showed the best metrics of classification in the validation set. The implementation of these algorithms enables the exploitation of the data of thousands of patients accurately, and a reduction of costs compared to traditional ways of collecting data for research. Thus, millions of patients from developing countries could benefit from local and specific data that could lead to treatments that take into account all their characteristics (genetic, environmental, habits, etc.) as it is the objective of precision medicine.

## References

[1] The Precision Medicine Initiative, in.
[2] A.D. Association, Standards of Medical Care in Diabetes--2014, *Diabetes Care* **37** (2014), S14-S80.
[3] A. Barceló and S. Rajpathak, Incidence and prevalence of diabetes mellitus in the Americas, *Revista Panamericana de Salud Publica* **10** (2001), 300-308.
[4] W. DH, Stacked Generalization, *Neural Networks* **5** (1992), 18.
[5] J.E. Fradkin, M.C. Hanlon, and G.P. Rodgers, NIH Precision Medicine Initiative: Implications for Diabetes Research, *Diabetes Care* **39** (2016), 1080-1084.
[6] A.N. Kho, M.G. Hayes, L. Rasmussen-Torvik, J.A. Pacheco, W.K. Thompson, L.L. Armstrong, J.C. Denny, P.L. Peissig, A.W. Miller, W.-Q. Wei, S.J. Bielinski, C.G. Chute, C.L. Leibson, G.P. Jarvik, D.R. Crosslin, C.S. Carlson, K.M. Newton, W.A. Wolf, R.L. Chisholm, and W.L. Lowe, Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study, *Journal of the American Medical Informatics Association: JAMIA* **19** (2012), 212-218.
[7] B. L, Stacked Regressions, *Machine Learning* **24** (1996), 15.
[8] K.P. Liao, A.N. Ananthakrishnan, V. Kumar, Z. Xia, A. Cagan, V.S. Gainer, S. Goryachev, P. Chen, G.K. Savova, D. Agniel, S. Churchill, J. Lee, S.N. Murphy, R.M. Plenge, P. Szolovits, I. Kohane, S.Y. Shaw, E.W. Karlson, and T. Cai, Methods to Develop an Electronic Medical Record Phenotype Algorithm to Compare the Risk of Coronary Artery Disease across 3 Chronic Disease Cohorts, *PLoS One* **10** (2015), e0136651.
[9] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics, *Information Sciences* **250** (2013), 113-141.
[10] S.N. Murphy, G. Weber, M. Mendis, V. Gainer, H.C. Chueh, S. Churchill, and I. Kohane, Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2), *J Am Med Inform Assoc* **17** (2010), 124-130.
[11] L.V. Rasmussen, W.K. Thompson, J.A. Pacheco, A.N. Kho, D.S. Carrell, J. Pathak, P.L. Peissig, G. Tromp, J.C. Denny, and J.B. Starren, Design patterns for the development of electronic health record-driven phenotype extraction algorithms, *J Biomed Inform* **51** (2014), 280-286.
[12] R.A. Wilke, R.L. Berg, P. Peissig, T. Kitchner, B. Sijercic, C.A. McCarty, and D.J. McCarty, Use of an Electronic Medical Record for the Identification of Research Subjects with Diabetes Mellitus, *Clinical Medicine and Research* **5** (2007), 1-7.

**Address for correspondence**

Corresponding author: Dr. Santiago Esteban

E-mail : santiago.esteban@hospitalitaliano.org.ar; santiagoesteban@gmail.com