*The Human Side of Service Engineering,* Vol. 62, 2022, 219–226
https://doi.org/10.54941/ahfe1002561

# Artificial Intelligence Algorithms in Precision Medicine: A New Approach in Clinical Decision-Making

**Pietro Campiglia, Valeria D'Amato, and Clara Bassano**

Department of Pharmacy, University of Salerno, Fisciano, SA, 84084, Italy

## ABSTRACT

US National Institutes of Health described the precision medicine as 'an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment and lifestyle for each person.' In other words, on the basis of the definition, the precision medicine allows to treat patients based on their genetic, lifestyle, and environmental data. Nevertheless, the complexity and rise of data in healthcare arising from cheap genome sequencing, advanced biotechnology, health sensors patients use at home, and the collection of information about patients' journey in healthcare with hand-held devices unquestionably require a suitable toolkit and advanced analytics for processing the huge information. The artificial intelligence algorithms (AI) can remarkably improve the ability to use big data to make predictions by reducing the cost of making predictions. The advantages of artificial intelligence algorithms have been extensively discussed in the medical literature. In this paper based on the collection of the data relevant for the health of a given individual and the inference obtained by AI, we provide a simulation environment for understanding and suggesting the best actions that need to be performed to improve the individual's health. Such simulation modelling can help improve clinical decision-maing and the fundamental understanding of the healthcare system and clinical process.

**Keywords**: Artificial intelligence, Precision medicine, Random forest

## INTRODUCTION

Precision medicine represents a new interesting frontier in healthcare as it customizes medical care to an individual's unique disease state. In general, the precision medicine has remarkably evolved in prevention, diagnosis, intervention and treatment, potentially changing dramatically the landscape of medicine. Based on patient-individual data, including medical diagnoses, clinical phenotype (severity of disease, amount of functional impairment, etc.), biologic investigations including laboratory studies and imagines, the novel approach allows to understand health and disease and make patient-tailored decisions in clinical field. Large volumes of information from multiple sources and multiple domains, including epigenetics "epigenomics", protein "proteomics", metabolics "metabolomics",radiology "radiomics", pharmacology "pharmacomics",microbiome studies "microbiomics", "environmental omics" (Mac Eachern et al. 2020), and others, determine the creation of

complex data, impossible to analyse without the help of data science and increasing computing power and suitable technologies. The widespread adoption of electronic health records has resulted in a tsunami of the collected data that have to help to dissect clinical heterogeneity and aid the healthcare practitioners in targeted decision-making (Saviano et al.2018). In order to extract precious information from the complex data under consideration, artificial intelligence algorithms and data science significantly contribute to identify complex patterns in data, to make predictions or classifications onnew unseen data or for advanced exploratory data analysis. For instance, the recent literature shows promising studies on the topic. In pharmacogenomics, which is a relatively new research field, the data science allows for mechanistic prediction of drug response and may help inform personalized drug design (Kalinin et al. 2018). In Dong et al. (2015) in order to develop anticancer drug sensitivity prediction using genomic data, a support vector machine model has been developed, by demonstrating that the response to cancer treatment could be predicted based on genomics. In particular, the authors show that the algorithms allow to avoid the unnecessary treatments in non-responders, in favour of the most effective treatment based on the patient's genome. In Type I diabetes analysis, Wei et al (2009) show the that a genotype-based disease risk assessment may be possible for diseases for which single nucleotide polymorphism (SNP) arrays by capturing a large risk proportion. As regards a study on Crohn's disease, Romagnoni et al. (2019) classify patients with Crohn's disease using artificial intelligence algorithms by obtaining the identification of future patients who may benefit from a specific treatment. The biological, psychological and environmental factors that influence brain development and mental health have been investigated using brain imaging and genetics with the quantitative support of the data science as in Mascarell et al. (2020). Artificial intelligence algorithms, in particular, machine learning have also been implemented for treatment response prediction. In Lin et al. (2018), the authors developed a deep learning model to predict antidepressant treatment response in patients with major depressive disorder based on SNP, demographic, and clinical data. Actually strengths and weaknesses of the different data science approaches depend on the specific clinical problem to face with (Lo Vercio et al. 2020). In our paper, we propose a random forest algorithm, which is an important tool in the field of the artificial intelligence, to solve a clinical decision –making process. We set up a robust classifier to select the "right" diagnostic-therapeutic protocol, based on the personalised frailty index calculated on the complex clinical data records. The rest of the paper is organised as follows. Section 2 presents methodological issues on the random forest and the frailty index for developing the classifier. In section 3 the main empirical results are illustrated. The final Section concludes.

## METHODOLOGICAL ISSUES

The RF tree base learner is typically grown using the methodology of CART (classification and regression tree), a methodology in which binary splits recursively partition the tree into homogeneous or near-homogeneous

terminal nodes (the ends of the tree). The algorithm was introduced by Breiman (2001) and consists of many individual trees which grow by recursively performing binary splits on the training dataset.

We assume $\left[(x_1,\ y_1),\ldots,(x_n,\ y_n)\right]$ is the training set training, which is considered as a sample of independent random variables distributed as pair $(X, Y)$ from an unknown distribution. The algorithm is addressed to predict the response $Y$ by estimating the regression function $m(x) = E[Y|X = x]$. The mean-squared generalization error for any numerical predictor h(x) is as in the following:

$$E_{X,Y} = \left(Y - h(X)\right)^2 \tag{1}$$

being the random forest predictor the average over $k = 1,\ldots,n$ trees (Breiman, 2001). The estimator of the target variable $\widehat{y}_{R_j}$ is the function of the regression tree estimator:

$$\widehat{f}^{tree}(X) = \sum_{j\epsilon J} \widehat{y}_{R_j} 1_{\{X\epsilon R_j\}} \tag{2}$$

$1_{\{.\}}$ being the indicator function and $(R_j)_{j\in J}$ the region of the predictor space which is divided into $J$ distinct and non-overlapping $R_1, R_2, \ldots, R_J$ and obtained by minimizing the Residual Sum of Square:

$$\widehat{f}^{tree}(X) = \sum_{j\epsilon J} \widehat{y}_{R_j} 1_{\{X\epsilon R_j\}} \tag{3}$$

Let us denote $B$ the number of bootstrap samples, the random forest estimator is calculated as

$$\widehat{f}^{\mathrm{RF}}(X) = \frac{1}{B}\sum_{b=1}^{B} \widehat{f}^{\mathrm{tree}}(X|b) \tag{4}$$

The personalised frailty index is built on the patient-data and we express by the following:

$$Fc(x,t) = F(x,t)\, risk\ factors I(x,t)$$

Being $Fc(x,t)$ the relative frailty of each individual obtained on the basis of the clinical information at age x and at t calendar year. It depends on the general frailty estimated on the aggregated population at age x and at t calendar year multiplied a corrective factor computed on the score of the different comorbidities or risk factors.

## NUMERICAL APPLICATIONS

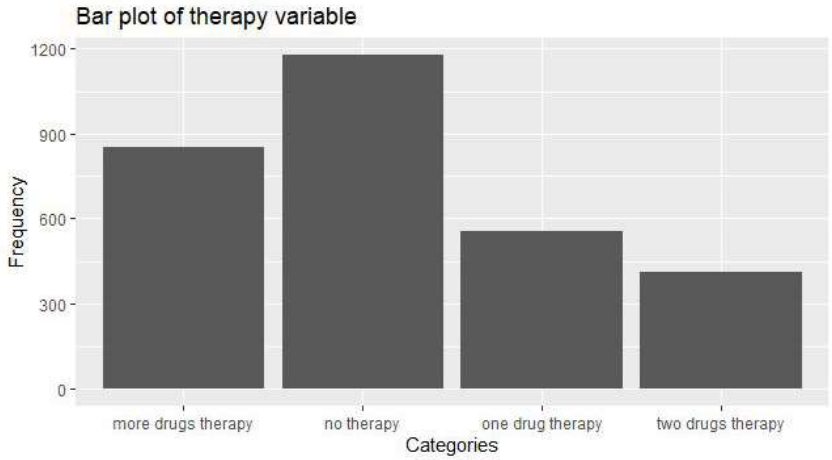
The empirical applications we perform in this section is based on confidential data, we describe with regards to the main structural features.

The dataset is a non-random sample of 3001 hospitalized people due to Covid-19, located in 26 hospitals and health centres of 13 Italian regions,

gathered from March to November 2020. For each individual, the following variables are detected: age, gender, the geographical area of hospitalization, the Charlson Comorbidity Index, a series of dummy variables that detect the presence of co-morbidities linked to Covid-19 and six different dummy variables for each medicine used in different combinations for therapies. Drug-related variables are combined into a categorical variable that considers the number of medicines an individual takes: "no therapy", "one drug therapy", "two drugs therapy", "more drugs therapy". The following table describes the co-morbidity variables:

| Variable | Description |
|---|---|
| hypert | Hypertension |
| obesity | Obesity |
| cad | Coronary artery disease |
| bpco | Chronic obstructive pulmonary disease |
| hf | Heart failure |
| ckd | Chronic kidney disease |
| diabetes | Diabetes |

Before proceeding with the random forest, the target variable is analysed to evaluate the best strategy. The following figure shows a bar plot of the target variable:




As can be seen from the figure, some categories of the target variable have a much lower frequency than others. It is therefore an unbalanced sample that requires the use of stratified sampling for the construction of the classification trees of the random forest. To evaluate the effectiveness of the stratification, the case of the unweighted and weighted random forest is compared, both in the in-sample and in the out-of-sample forecasts. To do this, a training set given by 70% of the sample is built, in order to use the remaining 30% as a test set. Each random forest will be estimated with 500 trees and a selection of 3 variables for each split.

The following table shows the results for in-sample forecasts, comparing Random Forest with unweighted scheme, a 1:1 weighting scheme, that is all the categories have the same number of observations within a tree, and a 1:2 weighting scheme, that is the less represented categories have half the observations of the others:

| **Unweighted Random Forest** | | | | | |
|---|---|---|---|---|---|
| **OOB estimate of error rate: 39.14%** | | | | | |
| **Confusion matrix** | | | | | |
| | **no therapy** | **one drug therapy** | **two drugs therapy** | **more drugs therapy** | **class error** |
| no therapy | 719 | 17 | 13 | 79 | 0.132 |
| one drug therapy | 63 | 46 | 33 | 241 | 0.880 |
| two drugs therapy | 33 | 24 | 38 | 188 | 0.866 |
| more drugs therapy | 32 | 48 | 51 | 475 | 0.216 |
| **Weighted Random Forest with 1:1 weights** | | | | | |
| **OOB estimate of error rate: 41.81%** | | | | | |
| **Confusion matrix** | | | | | |
| | **no therapy** | **one drug therapy** | **two drugs therapy** | **more drugs therapy** | **class error** |
| no therapy | 683 | 43 | 69 | 33 | 0.175 |
| one drug therapy | 55 | 85 | 136 | 107 | 0.778 |
| two drugs therapy | 28 | 37 | 161 | 57 | 0.431 |
| more drugs therapy | 21 | 95 | 197 | 293 | 0.517 |
| **Weighted Random Forest with 1:2 weights** | | | | | |
| **OOB estimate of error rate: 43.67%** | | | | | |
| **Confusion matrix** | | | | | |
| | **no therapy** | **one drug therapy** | **two drugs therapy** | **more drugs therapy** | **class error** |
| no therapy | 640 | 15 | 111 | 62 | 0.227 |
| one drug therapy | 53 | 14 | 138 | 178 | 0.963 |
| two drugs therapy | 15 | 7 | 177 | 84 | 0.375 |
| more drugs therapy | 17 | 21 | 216 | 352 | 0.419 |

As can be seen from the table, with the same global OOB error or at the limit with a minimum loss of accuracy, the use of weights makes it possible to drastically reduce the forecast error for the under-represented categories. In particular, for the 1:2 scheme, there is a noticeable improvement in the

| Unweighted Random Forest | | | | |
|---|---|---|---|---|
| **Accuracy: 0.5882 - 95% CI: (0.5553,0.6206)** | | | | |
| **Confusion matrix** | | | | |
| | **no therapy** | **one drug therapy** | **two drugs therapy** | **more drugs therapy** |
| no therapy | 292 | 32 | 13 | 15 |
| one drug therapy | 10 | 19 | 10 | 16 |
| two drugs therapy | 5 | 15 | 19 | 20 |
| more drugs therapy | 44 | 105 | 86 | 200 |
| | **no therapy** | **one drug therapy** | **two drugs therapy** | **more drugs therapy** |
| Sensitivity | 0.832 | 0.111 | 0.148 | 0.797 |
| Specificity | 0.891 | 0.951 | 0.948 | 0.639 |
| **Weighted Random Forest with 1:1 weights** | | | | |
| **Accuracy: 0.5594 - 95% CI: (0.5263, 0.5921)** | | | | |
| **Confusion matrix** | | | | |
| | **no therapy** | **one drug therapy** | **two drugs therapy** | **more drugs therapy** |
| no therapy | 274 | 27 | 11 | 11 |
| one drug therapy | 20 | 36 | 18 | 34 |
| two drugs therapy | 38 | 51 | 73 | 85 |
| more drugs therapy | 19 | 57 | 26 | 121 |
| | **no therapy** | **one drug therapy** | **two drugs therapy** | **more drugs therapy** |
| Sensitivity | 0.781 | 0.211 | 0.570 | 0.482 |
| Specificity | 0.911 | 0.901 | 0.775 | 0.843 |
| **Weighted Random Forest with 1:2 weights** | | | | |
| **Accuracy: 0.5527 - 95% CI: (0.5196, 0.5855)** | | | | |
| **Confusion matrix** | | | | |
| | **no therapy** | **one drug therapy** | **two drugs therapy** | **more drugs therapy** |
| no therapy | 259 | 23 | 9 | 6 |
| one drug therapy | 8 | 11 | 4 | 5 |
| two drugs therapy | 52 | 58 | 79 | 91 |
| more drugs therapy | 32 | 79 | 36 | 149 |
| | **no therapy** | **one drug therapy** | **two drugs therapy** | **more drugs therapy** |
| Sensitivity | 0.738 | 0.064 | 0.617 | 0.594 |
| Specificity | 0.931 | 0.977 | 0.740 | 0.774 |

prediction for the "two drugs therapy" category, but not for "one drug therapy" one, which even worsens the performance. On the contrary, for the 1:1 weighting scheme there is an improvement in performance for both categories, more marked for the "two drugs therapy" category.

The following table shows the result of out-of-sample forecasts:

As the table shows, all three methods used show a total accuracy of just under 60%. However, if we compare the statistics for the individual categories, there are notable differences in terms of sensitivity and specificity. In particular, we can observe that for unweighted random forest, the specificity is quite high for all categories, but the sensitivity is low for the under-represented categories. So, although the algorithm is able to identify with a certain degree of accuracy who does not need a certain treatment, it cannot identify who needs it. Instead, using a weighted random forest, it is possible to improve the sensitivity of the less represented categories, with a negligible loss of accuracy for the others.

The improvement of the performance of the random forest is for obvious reasons closely linked to the correct choice of weights.

## CONCLUDING REMARKS

Precision medicine is changing significantly the landscape of the medicine by taking advantage by the massive information extractable form the complex clinical data in a reliable and accurate risk management (Barile et al. 2021). The artificial intelligence algorithm can considerably support clinicians with diagnosis, prognosis and treatment on the basis of the patient-individual data analysis. We show promising results on setting out a robust classifier for selecting the "right" diagnostic-therapeutic protocol, based on the personalised frailty index calculated on the complex clinical data records.

On the basis of a personalised frailty indicator built up on the specific patient data, a robust decision rule which addresses the clinician toward the most suitable diagnostic-therapeutic protocol. In further studies, we will compare different data-driven methods for detecting advantages and drawbacks in clinical decision-making process.

## REFERENCES

Barile, S.; Bassano, C.; Piciocchi, P.; Saviano, M.; Spohrer, J. C (2021). Empowering value co-creation in the digital age, The Journal of Business & Industrial Marketing pp. 1–14. DOI:10.1108/JBIM-12-2019-0553. ISSN:0885-8624.

Collins F Precision Medicine Initiative | National Institutes of Health (NIH) [Internet]. National Institutes of Health. 2015. Accessed online on the 25th of July, 2017 from: https://www.nih.gov/precisionmedicine-initiative-cohort-program

Dong, Z., Zhang, N., Li, C., Wang, H., Fang, Y., Wang, J., and Zheng, X. 2015. Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection. BMC Cancer. 15(1): 489. doi:10.1186/s12885-015-1492-6. PMID:26121976.

Jiang F., Jiang Y., Zhi H., Dong Y., Li H., Ma S., Wang Y., Dong Q., Shen H., Wang Y., 2017, Artificial intelligence in healthcare: past, present and future, Stroke and Vascular Neurology, vol 2, doi:10.1136/svn-2017-000101

Filipp, F.V., 2019, Opportunities for Artificial Intelligence in Advancing Precision Medicine, Current Genetic Medicine Reports, 208–213 (2019). https://doi.org/10.1007/s40142-019-00177-4

Lo Vercio, L., Amador, K., Bannister, J., Crites, S., Gutierrez, A., MacDonald, M., and Forkert, N., 2020. Supervised machine learning tools: a tutorial for clinicians. J. Neural. Eng. 17(6): 062001. doi:10.1088/1741-2552/abbff2. PMID:33036008.

Lin, E., Kuo, P.-H., Liu, Y.-L., Yu, Y.W.-Y., Yang, A.C., and Tsai, S.-J. 2018. A deep learning approach for predicting antidepressant response in major depression using clinical and genetic biomarkers. Front. Psychiatry 9: 290. doi:10.3389/fpsyt.2018.00290. PMID:30034349.

Mascarell Maricic, L., Walter, H., Rosenthal, A., Ripke, S., Quinlan, E.B., Banaschewski, T., et al. 2020. The IMAGEN study: a decade of imaging genetics in adolescents. Mol. Psychiatry, 25: 2648–2671. doi:10.1038/s41380-020-0822-5. PMID:32601453.

Mesko B., 2017,The role of artificial intelligence in precision medicine, Expert Review of Precision Medicine and Drug Development, 2:5, 239-241, DOI: 10.1080/23808993.2017.1380516

Romagnoni, A., Jégou, S., Van Steen, K., Wainrib, G., Hugot, J.-P., and Peyrin-Biroulet, L. International Inflammatory Bowel Disease Genetics Consortium. 2019. Comparative performances of machine learning methods for classifying Crohn Disease patients using genome-wide genotyping data. Sci. Rep. 9(1): 10351. doi:10.1038/s41598-019- 46649-z. PMID:31316157.

Saviano, M., Bassano, C., Piciocchi, P., Di Nauta, P., Lettieri M. (2018) Monitoring viability and sustainability in healthcare management, Sustainability, Special Issue in Sustainability for Healthcare, 10 (10), 3548; doi:10.3390/su10103548, ISSN 2071–1050.

Wei, Z., Wang, K., Qu, H.-Q., Zhang, H., Bradfield, J., Kim, C., et al. 2009. From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. PLoS Genet. 5(10): e1000678. doi:10.1371/journal.pgen.1000678. PMID:19816555.