Predictive AI Models for the Personalized Medicine

Luigi Lella[1], Ignazio Licata[2], Gianfranco Minati[3], Christian Pristipino[4], Antonio Giulio De Belvis[5] and Roberta Pastorino[5]

[1]*ASUR, Regional Health Agency of Marche, AN, Italy*
[2]*ISEM, Inst. for Scientific Methodology, PA, Italy*
[3]*AIRS, Italian Association for Systems Research, MI, Italy*
[4]*ASSIMSS, Italian Association for Systems Medicine & Healthcare, Rome, Italy*
[5]*Section of Hygiene, Inst. of Public Health, Università Cattolica del Sacro Cuore, Rome, Italy*

Keywords: e-Health, e-Health Applications, Pattern Recognition and Machine Learning, Decision Support Systems.

Abstract: Innovative information systems which enable personalized medicine are presented. The designed decision support systems are expected to infer with an excellent level of accuracy the outcome of a therapeutic intervention through the analysis of biometric, genetic and environmental data. They are also capable to motivate their predictions according to a dynamic knowledge base, which is kept updated with new analysed cases. These systems can be used by researchers to identify useful correlations between biometric, genetic and environmental data with potential risks and benefits of certain therapeutic choices. They can also be used by the patients to choose the most appropriate therapeutic intervention according to their needs and expectations. In other words the presented decision support tools can realize the vision of the predictive, preventive, personalized and participatory (P4) medicine pursued by the systemic medicine.

1 INTRODUCTION

As reported in (Personalized Medicine, 2013), personalized medicine has opened a new rapidly growing market in the European industry, also creating new job opportunities.

The purpose of personalized medicine is essentially to contain healthcare expenditure at a time when the cost of healthcare delivery is growing throughout Europe along with the prevalence of chronic diseases and disorders, and more than 6% of readmission cases hospital due to acute conditions are caused by serious adverse drug reactions.

Research on the correlations between biological mechanisms, environmental interactions and the development or evolution of certain diseases and disorders will have a significant impact throughout the health care chain, from the research world to the provision of health care services (Saqui M. et al. 2016).

Despite the development of some personalized medicine approaches, we are still in one of the first stages of implementation of this intervention strategy (Nimmersgern E., 2017). According to a recent review of the personalized medicine literature presented by (Di Paolo A. et al. 2017), focused on research carried out within the European Union, there would not seem to be even sufficient consensus on the definition and conception of personalized medicine itself.

Some articles correlate its definition to the concept of stratification or subdivision of patients into subgroups, depending on the probability of receiving benefits from the adoption of a specific pharmaceutical therapy or clinical treatment. Others instead frame it as the assignment of a tailored therapy to patients on the basis of new individual and dynamical classifications of diseases based on their molecular basis and networking characteristics rather than only on clinical grounds.

As pointed out by the authors, the initial state of the patients is almost always evaluated considering mainly their genetic data and their biological markers together with the outcome of some specialized examinations. Instead, other factors such as the clinical evolution over time, as well as the needs and preferences of the patient should be considered as also required by a recent European recommendation (Personalised Medicine 2010) (Sagner M. et al 2017). Also according to (Di Paolo

Lella, L., Licata, I., Minati, G., Pristipino, C., De Belvis, A. and Pastorino, R.
Predictive AI Models for the Personalized Medicine.
DOI: 10.5220/0007472203960401
In *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2019)*, pages 396-401
ISBN: 978-989-758-353-7

A. et al. 2017), further research work is aimed both at predicting the individual outcomes of certain treatments and the probability of incurring collateral effects (Baumbach J. et al. 2018).

Regarding the technologies used in these deep learning tasks, literature seems to converge in recent years on the use of recurrent neural networks, in particular those models based on the Long Short-Term Memory (LSTM) paradigm for the analysis of genetic data (Vohradsky J. 2009; Xu R. et al. 2007) or the analysis of data contained in electronic medical records (Lipton Z.C. et al. 2017; Pham T. et al.2017).

This paper presents some solutions based on machine learning systems able to infer the outcome of a given treatment together with any side effects on the basis of patients status (genetic data, biometric data, environmental data), the chosen therapies, their needs and preferences.

In agreement with (Di Paolo A. et al. 2017) we believe that just by the adoption of a holistic approach, that does not consider only genetic data and biological markers but also the environment and the needs of the patients, it is possible to effectively deal with the problem of personalized medicine, adapting it exactly to the profile of patient. Through the development of outcome measures co-developed between researchers, patients and subject experts we will cover what really matters to patients, embracing the cognitive, self-cognitive, psychological, symbolic, social, ecological and environmental dimensions.

Most machine learning techniques are oriented towards a kind of structural representation of knowledge. This can be symbolic or subsymbolic. Sub-symbolic models can achieve the best results in problems that are difficult to solve if a static knowledge base consisting of simple logical production rules is adopted. Sub-symbolic models can be further subdivided into classification learning algorithms (Kohonen T., 1988; Rumelhart D.E. and McClelland J. L., 1986), association learning algorithms (Kohonen T., 1989) and clustering learning algorithms (Van Hulle M. M., 2012; Kohonen T. 1990; Fritzke B., 1994; Licata I. and Lella L., 2007).

In classification learning, the system is trained to provide a given output (a class) from a set of classified examples. This type of model, to which LSTMs belong, is only effective if the correlations between non-class attributes and all the possible classes are known in advance. This model does not therefore adapt to the case of the predictions of therapeutic choices in personalized medicine, since it can be very complicated to define the rules of association between individual profiles of patients and possible therapeutic interventions.

In association learning there are no specific classes, the system only tries to find an interesting scheme or a correlation between the data. Association rules can be used to predict attributes of any kind, not just class ones. Since we are interested in predicting the therapeutic choice, the duration of therapy, the risks and the results that can be achieved, association learning models are not suited to solve the problem.

Finally clustering algorithms are unsupervised, meaning that there is no set of classified examples that can be used to train the system. If we choose the duration of therapy, the achievable results and possible side effects as class attributes, the system can extrapolate several clusters related to class attributes. In this way it is possible to avoid the presence of human experts making this solution more interesting and easy to implement.

Among the algorithms belonging to this last family the SOM (Kohonen T., 1989) have been widely used in healthcare, but we believe that the best results can be obtained using more adaptive models. In this type of unsupervised learning activity there is no clear correlation between class attributes and the other ones. In other words, the exact topology of the input space is unknown.

B. Fritzke in one of his articles showed that his network model called GNG (Fritzke B., 1994) is able to identify exactly the local dimension of the input space, i.e. a GNG can find how many attributes in the defined input space are needed to accurately predict the considered class attributes.

As a further model to be compared with the self-organizing neural networks and the LSTMs, we will test the self-organizing symbolic model of the Non Organized Turing Machine (A-Type) (Turing A., 1948) consisting essentially of a network of NAND gates by which it is possible to construct a sort of knowledge base modelling the problem. This network will evolve through the use of various algorithms that encode the network configuration by means of fixed-length bit sequences. In particular we will consider Genetic Algorithms (Eiben A. E. and Smith J. E., 2015; Mitchell A.E., 2015) and Swarm Intelligence algorithms (Praveena S., 2018).

A Non-Organized Turing Machine is a symbolic model from which we expect a lower performance in terms of prediction accuracy than the considered subsymbolical models, but an A-type may be able to justify the inferred answers by resorting to a dynamic logical formalism.

The trained machine learning models will allow professionals and assistants to select the most suitable therapeutic regimen to treat the clinical case taken care of. The patient will have the opportunity to evaluate the outcome of a pharmacological or specialist therapy by selecting it from the list of those already used to treat similar cases.

This will lead not only to the patient's empowerment, but it will also lead to the realization of the long-awaited therapeutic alliance between caregiver and doctor who has taken care of him, limiting inappropriate interventions.

To achieve this result it is important to define diseases more precisely and to stratify patients into subgroups, based on their likelihood of responding to a given treatment, and also to stratify healthy citizens according to their risk of disease.

The classic approach of diagnosis and treatment must be overcome through specific omics data acquisition, the individual profile of the subject is assessed, enabling the choice of a specific therapeutic strategy. It is thus possible to minimize the "toxic cost" of the therapy, improving the patient's quality of life and optimizing the management of the available economic resources.

The described models will be tested at the the University Polyclinic Foundation Agostino Gemelli Hospital Center with the help of ad hoc resources collecting information which is not already collected routinarily. The expected pathology specific clinical, economic, quality and humanistic outcomes will be suggested by the involved multidisciplinary team.

2 METHODS

As input data to encode patients status, a binary vector will be assembled that encodes the genetic information, the molecular fingerprints (e.g. -omics), the biometric information, the clinical data, the therapeutic choice, the exposome, and the needs and the psychological dimensions of the patient and of his/her social networks.

As far as genetic information is concerned, a selection could be made, at least during the test phase of the developed decision support system, of all the possible about 30,000 human genes, considering only those that research considers useful for predicting the onset of disorders or diseases. For cancer alone, for example, large-scale studies (Hill S., 2018) have confirmed that there are about 450 "key genes" to be considered in the prediction of the onset or evolution of different forms of cancer.

In order to reduce the training and processing times of the chosen machine learning models, the considered cases could be limited only to a set of tumor forms that are particularly incident on the territory.

The encoding of such data will be accompanied by the codification of the outcome of some related specialist examinations. For example, the key gene for breast cancer called HER2 (Perez E.A. et al., 2014) is associated with the IHC (Immuno Histo Chemistry) exam that identifies the percentage of HER2 proteins in tumor cells, and with FISH (Fluorescence In Situ Hybridization), SPoT-Light HER2 CISH test and Inform HER2 Dual ISH test to identify if there are too many pairs of HER2 genes in tumor cells. The outcome of all these specialized examinations must be appropriately coded using a simple binary coding in the case of results that can be simply positive or negative or a "one-hot" coding, having as many bits as all the possible outcomes, and with only one of these coded as 1. For the IHC test of the HER2 gene, for example, the code "1000" can be used for the "negative" result, the code "0100" for the result "also negative", the code "0010" for the result "borderline" and the code "0001" for the "positive" result. A one-hot code should also be used to codify the choice of treatment regimen, the status and the needs of the patient.

The indicators to be taken into consideration to define the patient's status and needs will be taken from the information systems for the measurement of the outcomes reported by the patients as the one developed within the PROMIS project (Cella D. et al., 2010) or other outcome measures reported from the patients studied in literature (Black N., 2013) (Donabedian A., 1988). The outcome measures reported by the patients (PROM) are measures of functionality and well-being in the sphere of the patient's physical, mental and social health (Black N., 2013).

To codify the output of the chosen forecasting models, a vector with the one-hot coding of the duration of the therapy will be assembled (duration divided into classes or periods, for example: 0-6 months, 6-12 months> 1 year), together with a one-hot vector with the possible pathology specific outcomes (also in this case we will adopt the PROMIS coding system), and a sequence of binary codes (present or not present) associated with possible side effects.

To improve the learning process of the chosen self-organizing networks (SOM and GNG) as well as the Non-Organized Turing Machine, we will adopt the methodology suggested by Kohonen

(Kohonen T., 1990). The input vector of the chosen models will be constructed by concatenating a contextual part that represents the class attributes of the instance and a symbolic part composed of the other attributes. The part of the symbol and the part of the context will therefore be represented by two orthogonal vectors such that the norm of the second is larger than that of the first. In this way, in the subsimbolic prediction models taken into account the symbols can be coded in a topological order (connection between neural units) that reflects the logical analogies.

The implemented evolutionary algorithms (genetic and swarm intelligence) will instead be able to make more accurate predictions by selecting them from the considered space of the solutions.

The data set will be divided into a part equal to the 66% of the samples used as a training set, and a part equal to the remaining 34% of the samples used as a test set to evaluate the predictive accuracy of the model. All of these models have already been successfully tested in computationally similar contexts like the length of hospital stay prediction on the basis of the data contained in the patients admission forms (Lella L. and Licata I.,2017; Lella L. and Licata I., 2018).

Finally, it has to be noticed that these models must be trained with a large number of data, or rather, following the definition of big data provided by (Anderson C., 2008; Mayer-Schonberger V. and Cukier K., 2017; Godsey B., 2018), automatically collecting, storing and analysing all the clinical data, managing them as soon as they become available. As expressed by (Naimi A.I. and Westreich D. J., 2014) we will not consider the automatic analysis of all the data as the best adoptable scientific approach. According to the book review, we believe that all the available data will never be completely free of bias and in any case it will be necessary to adopt preprocessing techniques including resampling.

Instead, it will be fundamental to monitor in real time all the patients available data in order to follow the evolution of their clinical picture, suggesting possible prevention and treatment pathways.

In a future in which the personal, health-related and environmental information of each individual will be contained within a "personal data cloud" it will be possible to analyse in real time all this amount of data in order to provide people with useful coaching suggestions on how to improve their health preventing chronic disorders.

It will be possible, for example, to suggest to an individual, who has a genetic variant associated with a high predisposition to type 2 diabetes and a rapid increase in blood glucose level, to undergo a series of tests and to adopt certain dietary regimens and levels of physical activity to avoid the devastating effects of this disorder.

By activating the participatory component of medicine, patients will be more involved by making them aware of the possible consequences of their behaviour. This will reduce the onset of chronic disorders through self-monitoring and self-assessment leading to improved quality of life for patients and their caregivers.

3 CONCLUSIONS

Data mining and knowledge discovery processes do not follow precise rules. There is no model or method capable of producing useful results in any context of use.

In the case of the prediction of the duration of the therapy, of the outcome and the side effects of a personalized medicine case, it may be useful to use models such as GNG that perform the so-called dimensionality reduction. These models can find a sub dimensional space that contains most of all input data. The GNG model has the potential to adapt effectively to the input space, but it must be trained through the use of appropriate preprocessing techniques. We believe that the GNG model will perform better than other considered self-organizing networks, achieving a greater prediction accuracy.

We will also test a second symbolic model that implements the Non-Organized Turing Machines that is able to justify its predictions and to autonomously evolve its knowledge base over time.

The development of these systems is perfectly in line with two of the objectives specified in the European Union report on personalized medicine (Personalized Medicine, 2013), which are primarily to reduce the number of unnecessary interventions and adverse events by maximizing the added value perceived by patients, but also to favour a containment of welfare costs.

The use of artificial intelligence models in forecasting the outcomes of therapeutic choices can contribute to implement the predictive, preventive, personalized and participatory (P4) vision predicted and desired by some pioneers of systems medicine (Flores M. et al. 2013; Auffray C. et al. 2017).

Decision support systems supported by AI models, such as those presented in this work, will also make it possible to improve the effectiveness of medical decisions by moving from symptom-focused medicine to medicine focused on causes,

highlighting the therapies with the highest probability of success with the lower level of risk for each individual wherein the participation of the patient remains pivotal (Leyens L. et al. 2014).

REFERENCES

Anderson C., 2008. The end of theory: the data deluge makes the scientific method obsolete, Wired.

Auffray C., Sagner M., Abdelhak S., Adcock I., 2017. Viva Europa, a Land of Excellence in Research and Innovation for Health and Wellbeing. Progress Prev Med 2017; 2(3):e006. doi: 10.1097/pp9.0000000000000006

Baumbach J., Schmidt H., 2018. The End of Medicine as We Know It: Introduction to the New Journal, Systems Medicine. *Syst Med 2018*; 1: https://doi.org/https://doi.org/10.1089/sysm.2017. 28999.jba

Black N., 2013. Patient reported outcome measures could help transform healthcare, *BMJ n.346, p.167*.

Cella D., Yount S., Rothrock N., Gershon R., Cook K., Reeve B., Ader D., Fries J. F., Bruce B., Rose M., 2010. The patient-reported outcomes measurement information system (PROMIS), PMC.

Di Paolo A., Sarkozy F., Ryll B., Siebert U., 2017. Personalized medicine in Europe: not yet personal enough?, *BMC Health Services Research, vol.17, Issue 289*.

Donabedian A., 1988. The quality of care. How it can be assessed?, JAMA n.260, pp.1743-1748.

Eiben A. E., 2015. Smith J. E. Introduction to Evolutionary Computing, *Springer: Natural Computing Series, 2nd ed.*

Flores M., Glusman G., Brogaard K., Price N.D., Hood L, 2013. P4 medicine: how systems medicine will transform the healthcare sector and society, *Personalized Medicine vol.10, issue 6, pp. 565-576.*

Fritzke B., 1994. A Growing Neural Gas Network Learns Topologies. *Part of: Advances in Neural Information Processing Systems 7, NIPS.*

Godsey B., 2018. Think like a data scientist: tackle the data science process step by step, *Manning Publications Co*.

Hill S., 2018. Introducing genomics into cancer care, *BJS n.105, pp. e14-e15*.

Kohonen T., 1988. An introduction to neural computing, *Neural Networks, vol.1, pp.3-16.*

Kohonen T., 1989. Self-Organization and Associative Memory, *Berlin: Springer-Verlag*.

Kohonen T., 1990. The Self Organizing Map, *Proc. of the IEEE, vol.78, n.9*.

Lella L., Licata I., 2017. Prediction of length of hospital stay using a growing neural gas model, *Proc. of IMCIC 2017*.

Lella L., Licata I., 2018. Length of hospital stay prediction through unorganised Turing machines, *Proc. of BIOSTEC 2018, vol.5: HEALTHINF, pp. 402-407.*

Leyens L., Hackenitz E., Horgan D., Richer E., Brand A., Bubhoff U., Ballensiefen W., 2014. CSA Permed: Europe's commitment to personalised medicine, Eurohealth.

Licata I., Lella L., 2007. Evolutionary Neural Gas (ENG): A model of self-organizing network from input categorization", *EJTP, vol.4, n.14*.

Lipton Z.C., Kale D.C., Elkan C., Wetzel R., 2017. Learning to Diagnose with LSTM Recurrent Neural Networks, Available online: *arXiv:1511.03677*.

Mayer-Schonberger V., Cukier K., 2017. Big Data: A revolution that will transform how we live, work and think, *John Murray Publishers.*

Mitchell M., 1996. An Introduction to Genetic Algorithms, Cambridge, *MA: MIT Press.*

Naimi A. I., Westreich D. J., 2014. Big Data: A revolution that will transform how we live, work and think – book review, *American Journal of Epidemiology*, Available online: https://academic.oup.com/aje/article/179/9/1143/2739247.

Nimmesgern E., Benediktsson I., Norstedt I., 2017. Personalized Medicine in Europe, *Clin.Transl Sci. n.10, pp.61-63.*

Perez E. A., Cortes J., 2014. Gonzalez-Angulo A. M., Bartlett J. M. S. HER2 testing: current status and future directions, *Cancer Treatment Reviews n.40, pp. 276-284*.

Personalised Medicine – opportunities and challenges for European healthcare, 2010. Available online: http://ec.europa.eu/research/health/pdf/13th-european-health-forum-workshop-report_en.pdf.

Personalized Medicine, 2013. Available online: https://ec.europa.eu/research/health/index.cfm?pg=policy&policyname=personalised

Pham T., Tran T., Phung D., Venkatesh S., 2017. DeepCare: A Deep Dynamic Memory Model for Predictive Medicine, Available online: *arXiv:1602.00357*.

Praveena S., 2018. Review on Swarm Intelligence Algorithms, *International Journal of Scientific Research in Computer Science, Engineering and Information Technology, Vol.3, Issue 4*.

Rumelhart D. E., McClelland J. L., Eds., 1986. Parallel Distributed Processing, vol.1, *Cambridge, MA: MIT Press*.

Sagner M., McNeil A., Puska P., Auffray C., Price N. D., Hood L., Lavie C. J., Han Z.G., Chen Z., Brahmachari S. K., McEwen B. S., Soares M. B., Balling R., Epel E., Arena R., 2017. *The P4 Health Spectrum - A Predictive, Preventive, Personalized and Participatory Continuum for Promoting Healthspan. Prog Cardiovasc Dis.2017;59(5):506-521*.

Saqi M., Pellet J., Roznovat I., Mazein A., Ballereau S., De Meulder B., Auffray C., 2016. *Systems Medicine: The Future of Medical Genomics, Healthcare, and Wellness. Methods Mol Biol.2016;1386:43-60.*

Turing A., 1948. Intelligent Machinery, in Collected Works of A.M.Turing:Mechanical Intelligence. Edited by D.C.Ince. *Elsevier Science Publishers, 1992*.

Van Hulle M.M., 1989. Self Organizing Maps, *Handbook of Natural Computing, pp. 585-622*.

Vohradsky J., 2009. Neural network model of gene expression, *the FASEB Journal, vol.19, no.2, pp. 320-329.*

Xu R., Wunsch II D., Frank D., 2007. Inference of genetic regulatory networks with recurrent neural network models using particle swarm optimization, *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), vol. 4, n.4, pp. 681-692.*