An Empirical Analysis of Predictors for Workload Estimation in Healthcare

Roberto Gatta[1], Mauro Vallati[2(✉)], Ilenia Pirola[3], Jacopo Lenkowicz[1], Luca Tagliaferri[4], Carlo Cappelli[3], and Maurizio Castellano[3]

[1] Universitá Cattolica del Sacro Cuore, Istituto di Radiologia, Rome, Italy
roberto.gatta.bs@gmail.com
[2] School of Computing and Engineering, University of Huddersfield, Huddersfield, UK
m.vallati@hud.ac.uk
[3] Universitá degli Studi di Brescia, Spedali Civili di Brescia, Brescia, Italy
[4] Fondazione Policlinico Universitario A. Gemelli IRCCS, Rome, Italy

Abstract. The limited availability of resources makes the resource allocation strategy a pivotal aspect for every clinical department. Allocation is usually done on the basis of a workload estimation, which is performed by human experts. Experts have to dedicate a significant amount of time to the workload estimation, and the usefulness of estimations depends on the expert's ability to understand very different conditions and situations. Machine learning-based predictors can help in reduce the burden on human experts, and can provide some guarantees at least in terms of repeatability of the delivered performance. However, it is unclear how good their estimations would be, compared to those of experts.

In this paper we address this question by exploiting 6 algorithms for estimating the workload of future activities of a real-world department. Results suggest that this is a promising avenue for future investigations aimed to optimising the use of resources of clinical departments.

Keywords: Workload estimation · Machine learning · Predictors

1 Introduction

Global spending on health is consistently growing worldwide [7], and it is expected to grow in the near future. This is the result of two main driving forces:

- in countries where the economy is developing, the increase is due to the improvement of services overtime.
- In the so-called first-world countries, the growing life expectancy and the low birth rate are already increasing the pressure on the healthcare (see for instance [10]).

Remarkably, the problem faced in developed countries envisages a scenario where optimising the available resources will be a mandatory way to increase the efficiency of the healthcare system, and to optimise delivered services.

© Springer Nature Switzerland AG 2020
V. V. Krzhizhanovskaya et al. (Eds.): ICCS 2020, LNCS 12137, pp. 304–311, 2020.
https://doi.org/10.1007/978-3-030-50371-0_22

There are many different aspects and perspectives that can be subject to resource optimisation: optimisation can focus on different levels of the organisational charts, can focus on geographical clusters, can be tuned for the type of delivered services or clinical domains, and can address both administrative and clinical issues. Examples of approaches aimed at optimising the use of resources include a dynamic appointment scheduling system to cope with no-show patients and appointments deletion [5]; a scheduler for Radiology Departments [9]; and a chemotherapy appointment scheduling model under uncertainty [2]. There is a growing interest in optimisation approaches, thanks to the potentially large benefits that their application would result in for an hospital or a clinical department.

Notably, most of the existing optimisation approaches deal with the allocation of resources, as soon as appointment requests are received or an estimation of future workload has been performed. In a sense, this is a kind of *reactive* optimisation. Intuitively, optimising resources on the basis of estimated workload can lead to better resource optimisation, due to the fact that there is no need to wait for actual appointments to be made. This would allow a shift from *reactive* to *pro-active* otpimisation. However, this kind of pro-active optimisation is very sensitive to the quality of predictions that are provided. Despite being pivotal for the allocation and exploitation of available resources, the workload estimation is still mostly performed manually by human experts, that have to devote a usually significant amount of their time to perform such task. Moreover, the usefulness of estimations depends on the expert's ability to understand very different conditions and situations, and is very hard to verify. In fact, the same expert can provide both very accurate and very inaccurate estimations, undermining the subsequent allocation processes.

Machine learning-based predictors may help to overcome some of the aforementioned issues. In particular, their use can reduce the burden on experts, and provide some general guarantees on the quality of the predictions. Furthermore, machine learning can be used to quickly generate multiple scenarios, that can then be compared by experts to select the most appropriate. However, in order to understand the usefulness of well-known machine learning approaches for this task, it is mandatory to assess their ability in estimating future workloads in real-world circumstances.

In order to address the above issue, in this paper we present the results of a large empirical analysis aimed at comparing the performance in workload estimation of a number of algorithms on real-world data obtained from a Centre of study on Thyroid. To minimise the risk of providing results that are only specific for the case taken into account, we trained the considered algorithms on a restricted set of information, commonly available on the vast majority of Electronic Health Records (EHR) or appointment booking systems.

The remainder of this paper is organised as follows. First, we describe material and methods of the performed analysis. Then, in Sect. 3 we present results and a discussion. Next, we provide the conclusion of this paper and we envisage future steps.

2 Materials and Methods

From the EHR of a Centre of study on Thyroid we extracted the complete clinical pathways of patients treated for acute and chronic thyroid diseases (e.g., cancer, age related, genetic hyper-hypo thyroidism, etc.). The investigated events were:

- (i) oncological examinations: ambulatory visits aimed at staging the Thyroid neoplasm, to assess the progression during the treatment or follow-up visits;
- (ii) Non-oncological examinations: ambulatory visits for generic consultations for specific non-oncologic diseases such as hypo or hyper-thyroidism (e.g. due to physiological ageing or more specific reasons, such as the Basedow diseases).
- (iii) Free triiodothyronine (fT3), a thyroid hormone. This analysis only requires a blood sample; for this reason it tends to be relatively cheap to perform and it is commonly prescribed.
- (iv) Free thyroxine (fT4), a thyroid hormone similar to fT3. It can be analysed as the fT3 and, together, they are primarily responsible for regulation of metabolism.
- (v) Parathyroid hormone (PTH), an hormone secreted by the parathyroid glands with a relevant role in the regulation of the serum calcium.
- (vi) Thyroglobulin (Tg), a protein produced and consumed within the Thyroid.
- (vii) Other common laboratory exams, such as complete blood count, cholesterol, etc.
- (viii) Thyroid ultrasound investigation,
- (ix) Fine Needle Aspiration Cytology (FNAC): the aspiration of some thyroid cells with a fine needle guided via ultrasound. Due to the invasive nature of the procedure, it require specific clinical skills and can be considered the most demanding event.

We decided to focus on this level of granularity because, also as a result of discussions with human experts of the considered medical field, these are key events with regards to human resources of a department (e.g. FNAC, Ultrasound, Medical examinations) or with regards to lab time and costs (e.g. fT3, fT4, Tg, PTH). Furthermore, those events are commonly recorded in EHRs, and would therefore provide a general ground to exploit workload estimation predictors in different units or departments.

Other clinical variables, such as co-morbidities, drugs or biomarkers was not considered: such kind of data are not always present in the EHR and when present are often represented without any specific reference to a shared ontology. For this reason, even if their inclusion had increased the performances of the predictions, it would also had reduced the reproducibility.

We considered a total of $5,941$ patients treated by the thyroid centre, which lead to $42,839$ events. The available data has been processed as follows. For each of the 9 clinical events analysed in this study, and considering all the patients involved in the event, we divided the logs in two parts, corresponding to an observation time window of at least 18 months before and 18 months after. The predicting task is to estimate the number of events that will occur in the 18

months after the event, given information about the 18 months before. It should be noted that a different predictor is built for each of the 9 events, and such predictor is only used to predict the number of future occurrences of such event. We then trained and tested the predictors exploiting a cross-validation jackknife approach, where 90% of the available data is used for training purposes, and the remaining 10% for testing predictors.

The 18-months time window reflects, to some extent, the nature of the treatments performed in the considered centre. This represents the common follow-up time, and includes a prudential margin to allow enough informative content for the prediction of the following 18. Of course, for different departments, this value can be straightforwardly adapted.

2.1 Algorithms

For the sake of this experimental analysis, we considered six well-known algorithms for building predictors, spreading from naive approaches –exploited as baselines– to widely-exploited Machine Learning techniques.

- **Mean**: considering the entire training set it calculates the density of each kind of event during the time (how many, on average, per month) and use this density to predict how many events are expected in the future.
- **TipOver**: each prediction is simply made by replicating the past recorded events. More specifically, for each patient, the kind and number of events of the next x months are exactly the same of the previous x months.
- k-nearest neighbours algorithm (**kNN**) [4]: uses the neighbourhood of the 8 most similar clinical cases and uses them to estimate the future, exploiting the mean of events occurred in the past 18 months. The metric is built on an n dimensional space where n is the number of kind of events. In this way, any patient can be seen as a point and the euclidean distance is used to select the neighbourhood. The axes are normalised between 0 and 1 to avoid overweighting the most frequent events.
- Generalised linear model (**lm**) [11]: uses generalised linear regression to estimates the next 18 months, adopting all the entire training data set.
- Random Forest (**rf**) [3]: Random forests are a combination of predictors such that each predictor is randomly generated, and all the predictorsa have the same weight. We built a Random Forest-based models using 500 random trees;
- Support-vector machine (**svm**) [1]: Support Vector Machines-based models the exploits a Gaussian kernel to perform the prediction.

2.2 Domain Expert

The director of the Centre of study on Thyroid is the human expert that is in charge of estimating the workload of the unit. She has some 20+ years experience in the specific domain. In order to make the comparison as fair as possible, she was asked to make estimations on the basis of the same data that is made available to the considered algorithms. Notably, this is not the usual amount of

data that is provided to human experts. In most of the cases, they are required to estimate future workloads by relying on a significantly smaller amount of explicit knowledge; however, they can leverage on their extensive experience in the field. For this reason, we believe we put the human expert in the best possible condition to perform her work: a large amount of available data that can provide a good and compelling overview of the past months.

3 Results

Results of the performed comparison are shown in Fig. 1. For each considered event, we provide a box-and-whisker plot showing the distribution of the error percentage, measured as the percentage of events as follows:

$$\frac{(predicted - actual)}{actual} \tag{1}$$

An average percentage error value of 0.0 indicates that the predictor has always provided the perfect estimation. In each box, the mid-line is the median of the performance, with the upper and lower limits of the box being the third and first quartile respectively. The whiskers will extend up to 1.5 times the interquartile range from the top (bottom) of the box to the furthest datum within that distance. In predicting the expected numbers of Other Lab exams, for example, the algorithm tipOver has a median error close to 100% with the 50% of the measured performances included approximately between and 80% and 110%. Admittedly, tipOver is not a good approach to estimate the workload for that type of clinical events.

Dispersion is also to take into account, as a high value indicates that the corresponding predictor's performance can vary greatly according to the considered circumstances. The solid horizontal (red) line represents the performance of the human expert. In this case, we could not show any dispersion value, as the expert made only a limited number of estimations, due to the complexity of the task when performed manually.

3.1 Discussion

The results presented in Fig. 1 indicate that the machine learning-based predictors tend to estimate better than the very basics *mean* and *tipOver* approaches. However, even such naive approaches can deliver good performance in a couple of cases, indicating that the corresponding events are trivially easy to predict, given a suitable amount of available information. Notably, in some cases the *mean* approach is able to deliver prediction that outperform human experts: it can indeed be the case that even such naive approaches can be useful in supporting humans, by clearly highlighting regular patterns that would otherwise be hard to identify. On the other hand, more sophisticated ML approaches tend to consistently deliver better performance also on more complex cases.

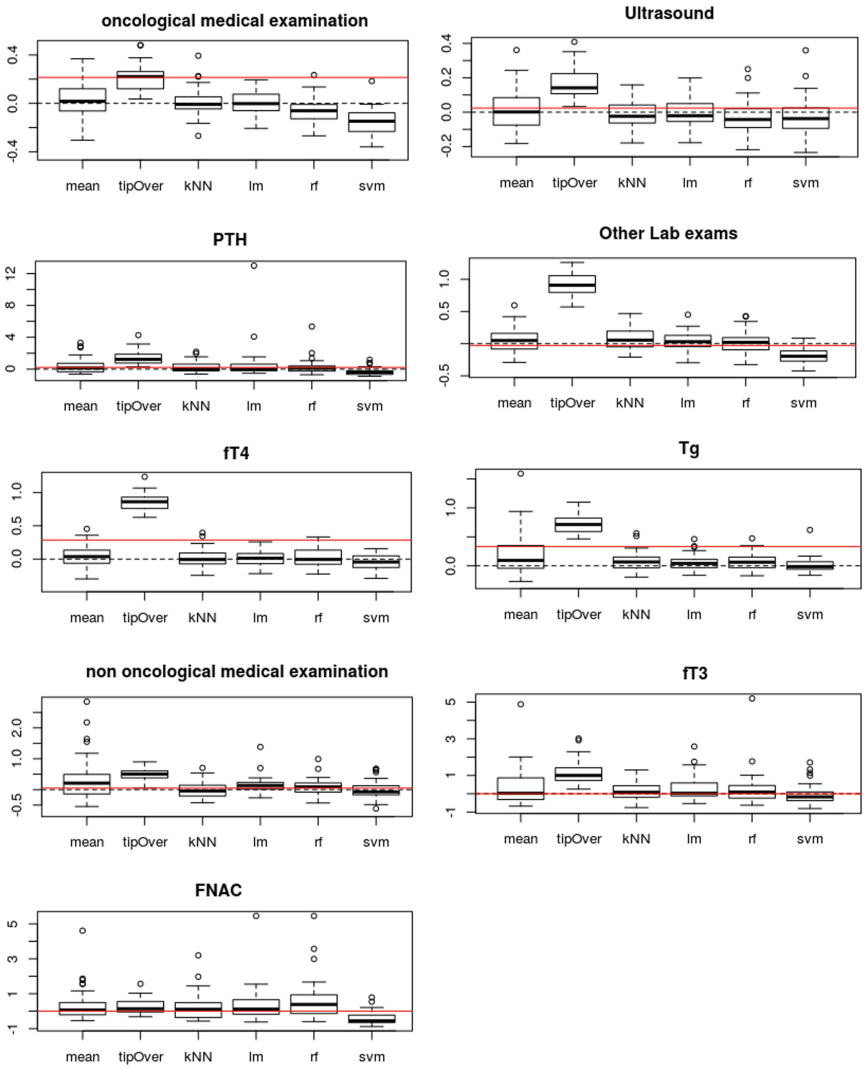

Fig. 1. Performance, in terms of average estimation error percentage (y-axis) of the considered algorithms when predicting the number of occurrences of the 9 clinical events. In each box, the mid-line is the median of the performance, with the upper and lower limits of the box being the third and first quartile respectively. The whiskers will extend up to 1.5 times the interquartile range from the top (bottom) of the box to the furthest datum within that distance. The solid horizontal indicates the performance of the human expert. (Color figure online)

In most of the considered cases, the performance of the human expert are impressive, even though ML-based techniques can still help in reducing mistakes and improving predictions. Noteworthy, the human expert has been making

workload estimations for the considered centre for more than 20 years. Therefore, it is safe to assume that the delivered predicting performance is a very accurate representation of the best performance that can be achieved by a human. Further, the workload estimation task is very time consuming, and the results can significantly vary according to the experience of the human expert. The more experienced the expert is, the best are expected to be the predictions: however, there is also to factor in that fact that more experienced humans are extremely valuable resources that should spend their precious time on more critical tasks. Given this perspective, ML-based approaches can deliver generally good performance for estimating the workload for all the considered clinical events, and are extremely quick.

Interestingly, there is not a single algorithm that is able to outperform all the others in all the considered prediction tasks. On the one hand, this suggests that the clinical events we focused on are suitable for empirically comparing approaches as they pose very different challenges to predictors. On the other hand, results also point to the fact that an ensemble predictor may best suit the needs of a clinical department. An ensemble approach where a different predictor is trained for each event may therefore deliver robust and reliable performance.

4 Conclusions

Workload estimation is pivotal for optimising the use of resources in modern hospital departments. However, despite its importance, this task is mostly performed by human experts. Experts require a significant amount of time for performing this task, and results are highly dependent on the experience of the human. In this paper, we investigated the use of machine learning approaches for efficiently performing this tedious yet pivotal task.

The experimental analysis we performed demonstrates that it is possible to exploit machine learning-based predictors to accurately estimate workload of a clinical department, in terms of occurrences of a number of personnel or lab/cost intensive clinical events. In other words, human experts can be relieved by the burden of performing such time-consuming task: this has significant implications in terms of optimisation. Firstly, senior experts will have more available time to dedicate to more relevant matters. Secondly, quick and accurate ML-based predictions can be used as input to schedule-optimiser, in order to optimise the allocation of resources via more robust and better informed scheduling.

We see several avenues for future work. Firstly, we are interested in investigating the use of ensemble-based approaches for maximising the predicting performance of a wide range of clinical events. Secondly, we plan to extend our analysis to different departments, in order to evaluate how general the presented results are. Thirdly, we are interested in evaluating whether sharing information between departments of different hospitals can help improving the performance of predictors, by leveraging on privacy-preserving approaches [6].Finally, we will focus on approaches aimed at integrating the strengths of machine learning with the capabilities of human experts, possibly using an overarching framework that encompasses all the relevant steps of the process [8].

References

1. Aizerman, M.A.: Theoretical foundations of the potential function method in pattern recognition learning. Autom. Remote Control **25**, 821–837 (1964)
2. Alvarado, M., Ntaimo, L.: Chemotherapy appointment scheduling under uncertainty using mean-risk stochastic integer programming. Health Care Manage. Sci. **21**(1), 87–104 (2016). https://doi.org/10.1007/s10729-016-9380-4
3. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
4. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Trans. Inf. Theory **13**, 21–27 (1967)
5. Creps, J.R., Lotfi, V.: A dynamic approach for outpatient scheduling. J. Med. Econ. **20**, 786–798 (2017)
6. Damiani, A., et al.: Distributed learning to protect privacy in multi-centric clinical studies. In: Holmes, J.H., Bellazzi, R., Sacchi, L., Peek, N. (eds.) AIME 2015. LNCS (LNAI), vol. 9105, pp. 65–75. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-19551-3_8
7. Dieleman, J.L., et al.: National spending on health by source for 184 countries between 2013 and 2040. Lancet **387**(10037), 2521–2535 (2016)
8. Gatta, R., et al.: Towards a modular decision support system for radiomics: a case study on rectal cancer. Artif. Intell. Med. **96**, 145–153 (2019)
9. Gatta, R., et al.: On the efficient allocation of diagnostic activities in modern imaging departments. In: Pereira, F., Machado, P., Costa, E., Cardoso, A. (eds.) EPIA 2015. LNCS (LNAI), vol. 9273, pp. 103–109. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23485-4_10
10. Guzman-Castillo, M., et al.: Forecasted trends in disability and life expectancy in england and wales up to 2025: a modelling study. Lancet **2**(1), e307–e313 (2017)
11. Nelder, J.A., Wedderburn, R.W.: Generalized linear models. J. Roy. Stat. Soc. Ser. A (Gen.) **135**(3), 370–384 (1972)