Greg, ML: Automatic Diagnostic Suggestions

Humanity is Overrated. Or not.

Paola Lapadula [1], Giansalvatore Mecca [1], Donatello Santoro [1],
Luisa Solimando [2], and Enzo Veltri [2]

[1] Università della Basilicata – Potenza, Italy
[2] Svelto! Big Data Cleaning and Analytics – Potenza, Italy

(Discussion Paper)

Abstract. Recently machine-learning techniques have been applied in a variety of fields. One of the most promising and challenging is handling medical records. In this paper we present Greg, ML, a machine-learning tool for generating automatic diagnostic suggestions based on patient profiles. At the core of our system there are two machine learning classifiers: a natural-language module that handles reports of instrumental exams, and a profile classifier that outputs diagnostic suggestions to the doctor. After discussing the architecture we present some experimental results based on the working prototype we have developed. Finally, we examine challenges and opportunities related to the use of this kind of tools in medicine, and some important lessons learned developing the tool. In this respect, despite the ironic title of this paper, we underline that Greg should be conceived primarily as a support for expert doctors in their diagnostic decisions, and can hardly replace humans in their judgment.

1 Introduction

The larger availability of digital data related to all sectors of our everyday lives has created opportunities for data-based applications that would not be conceivable a few years ago. One example is medicine: the push for the widespread adoption of electronic medical records [9, 5] and digital medical reports is paving the ground for new applications based on these data.

Greg, ML [8] is one of these applications. It is a machine-learning tool for generating automatic diagnostic suggestions based on patient profiles. In essence, Greg takes as input a digital profile of a patient, and suggests one or more diagnosis that, according to its internal models, fit the profile with a given probability. We assume that a doctor inspects these diagnostic suggestions, and takes informed actions about the patients.

We notice that the idea of using machine learning for the purpose of examining medical data is not new [7, 11, 10]. In fact, several efforts have been taken in this direction [1, 6]. To the best of our knowledge, however, all of the existing tools concentrate on rather specific learning tasks, for example identifying a single pathology – like

heart disease [14, 12], or pneumonia [13], or cancer, where results of remarkable quality have been reported [15]. On the contrary, Greg has the distinguishing feature of being a broad-scope diagnostic-suggestion tool. In fact, at the core of the tool stands a generic learning model that allows to suggest large numbers of pathologies, currently several dozens, and in perspective several hundreds.

Greg is a research project developed by Svelto!, a spin-off of the data-management group at University of Basilicata.

The rest of the paper is devoted to introduce Greg, as follows. We discuss the internal architecture of the tool in Section 2. Then, we introduce the methodology and the additional tools in Section 3. We introduce some experimental results based on the current version of the tool in Section 4.

Finally, in Section 5 we conclude by discussing the possible applications we envision for Greg, and discuss a few crucial lessons learned with the tool, which, in turn, have inspired the title of this paper.

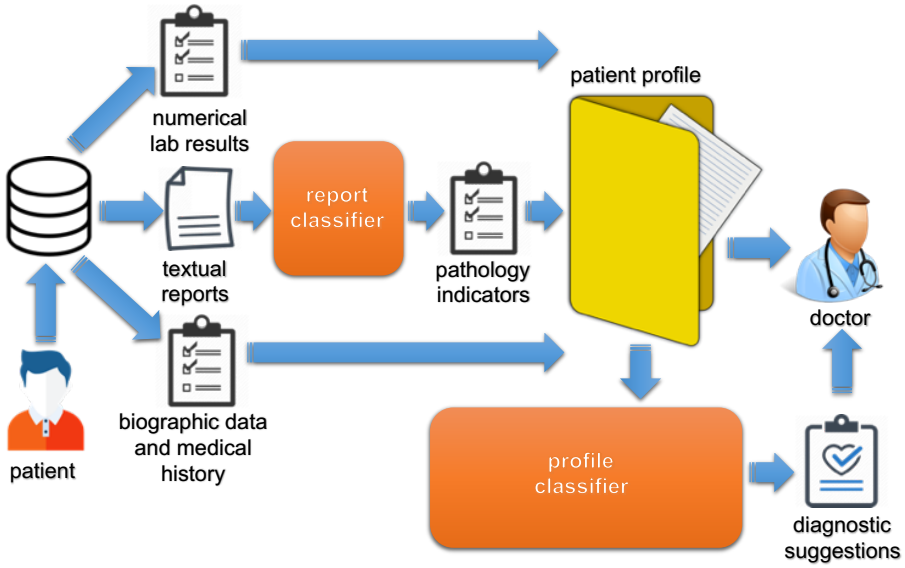

Fig. 1: Architecture of Greg.

2 Architecture of Greg

The architecture and the overall flow of Greg is depicted in Figure 1.

As we have already discussed, at the core of Greg stands a classifier for patient profiles that provides doctors with diagnostic suggestions. Profiles are entirely anonymous, i.e., Greg does not store nor requires any personal data about patients, and are composed of three main blocks:

- anonymous biographical data, mainly age and gender, and medical history of the patient, i.e., past medical events and pathologies, especially the chronic ones;
- result of lab exams, in numerical format;
- textual reports from instrumental exams, like RX, ultrasounds etc.

These items compose the patient profile that is fed to the profile classifier in order to propose diagnostic suggestions to doctors. Notice that, while biographic data, medical history and lab exam results are essentially structured data, and therefore can be easily integrated into the profile, reports of instrumental exams are essentially unstructured. As a consequence, Greg relies on a second learning module to extract what we call *pathology indicators*, i.e., structured labels indicating anomalies in the report that may suggest the presence of a pathology.

The report classifier is essentially a natural-language processing module. It takes the text of the report in natural language and identifies pathology indicators that are then integrated within the patient profile.

The report classifier is, in a way, the crucial module for the construction of the patient profile. In fact, reports of instrumental exams often carry crucial information for the purpose of identifying the correct diagnostic suggestions. At the same time, their treatment is language-dependent, and learning is labor-intensive, since it requires to label large set of reports in order to train the classifier.

Once the profile for a new patient has been built, it is fed to the profile classifier that outputs diagnostic suggestions to the doctor. There are a few important aspects to be noticed here.

- First, Greg is trained to predict only a finite set of diagnoses. This means that it is intended primarily as a tool to gain positive evidence about pathologies that might be present, rather than as a tool to exclude pathologies that are not present. In other terms, the fact that Greg does not suggest a specific diagnosis does not mean that that can be excluded, since it might only be the case that Greg has not be trained for that particular pathology. It can be seen that handling a large number of diagnoses is crucial, in this respect.
- Second, Greg associates a degree of probability with each diagnostic suggestion, i.e., it ranks them with a confidence measure. This is important, since the tool may provide several different suggestions for a given profile, and not all of them are to be considered as equally relevant.

It can be seen how a tool like Greg has an effective as seamless integration with the everyday procedures of a medical institution is. To foster this kind of adoption, Greg can be used as a stand-alone tool, with its own user-interface, but it has been developed primarily as an engine-backed API, that can be easily integrated with any medical information system that is already deployed in medical units and wards. Ideally, with this kind of integration, accessing medical suggestions provided by Greg should cost no more than clicking a button, in addition of the standard procedure for patient-data gathering and medical-record compilation.

3 The Greg Workflow and Ecosystem

As we have discussed in the previous sections, the effectiveness of a system like Greg is strongly related to the number of pathologies which it can provide suggestions for. We therefore put quite a lot of effort in structuring the learning workflow in order to make it lean and easily reproducible. In this section we summarize a few key findings in this respect, that led us to the development of a number of additional tools, which compose the Greg ecosystem.

A first important observation we make is that a system like Greg needs to make reference to a standardized set of diagnosis. As it is common, we rely on the international classification of diseases, *ICD-10 (DRG)* [3]. This, however, poses a challenge when dealing with large and heterogeneous collections of medical records coming from disparate sources, which do not necessarily are associated with a DRG. This poses a standardization problem for diagnosis labels. In fact, standardizing the vocabulary of pathologies and pathology indicators is crucial in the early stages of data preparation. To this end, we leveraged the consolidated suite of data-cleaning tools developed by our research group over the years [2–4].

A second important observation is that we need to handle large and complex amounts of data gathered from medical information systems, including biographical data, admissions and patient medical history, medical records, multiple lab exams, and multiple reports. These data need to be explored, selected and prepared for the purpose of training the learning models. In order to streamline the data-preparation process, we decided to develop a tool to explore the available data. The tool is called Caddy and is essentially a data warehouse build on top of the transactional medical databases. This allowed us to adopt a structured approach to data exploration and data selection, that proved essential in the development of the tool.

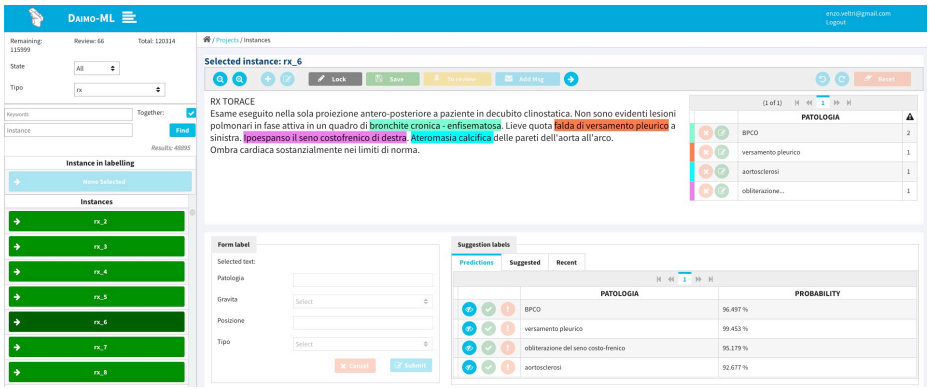

Fig. 2: DAIMO, the ML Labeling Tool.

However, the tool that proved to be the most crucial in the development of Greg is DAIMO, our instance labeler. DAIMO stands for *Digital Annotation of Instances and*

[3] http://www.who.int/classifications/icd/icdonlineversions/en/

Markup of Objects. It is a tool explicitly conceived to support the labeling phase of machine learning projects. A snapshot of the system is shown in Figure 2.

DAIMO is a semi-automated tool for data labeling. It provides a simple and effective interface to explore pre-defined collections of samples to label. Samples may be either textual, or even structured – for example, in tabular format– or even of mixed type. Users that are tasked with labeling can cooperatively explore the samples, pick them, explore existing labels and add more. Figure 2 shows the process of labeling one report. Labels associated with the report are on the right. Each corresponds to a colored portion of the text.

We believe that even only the availability of an intuitive tool to support cooperative labeling-work significantly increases productivity. In addition to this, DAIMO provides additional functionalities that further improve the process.

First, it allows to define *label vocabularies*, in order to standardize the way in which labels are assigned to samples. Users usually search labels within the vocabulary, and add new ones only when the ones they need are not present. When dealing with complex labeling tasks with many different labels, such a systematic approach is crucial in order to get good-quality results.

Second, DAIMO is able to *learn* labeling strategies from examples. After some initial training, it does not only collects new labels from users, but actually suggests them, so that users need only to accept or refuse DAIMO's suggestions. This approach really transforms the labeling process from the inside out, since after a while it is DAIMO, not the user to do most of the work.

In fact, in our experience, working with DAIMO may lower text-labeling times up to one order of magnitude with respect to manual, unassisted labeling.

4 Experimental Results

ID	Diagnosis	GT - D.L.	GT - Doct.
D1	Iron Deficiency	0,18	1,00
D2	Atherosclerosis	0,09	0,98
D3	Cardiomegaly	0,46	0,95
D4	Hepatic Cirrhosis	0,18	0,97
D5	Hypercholesterolemia	0,10	1,00
D6	Prostatic Hyperplasia	0,18	0,97
D7	Hypokalemia	0,10	0,97
D8	Pneumonia	0,63	0,96

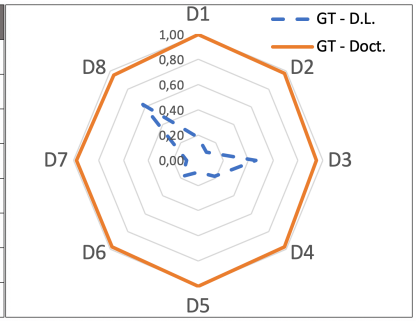

Fig. 3: Experimental Results prior and after Diagnosis Review in terms of F-Measure

We developed an advanced prototype of Greg, used to conduct a number of experiments to assess the feasibility of the overall approach.

We conducted a first preliminary experimental evaluation using 200 medical records over a small set of diagnosis (pneumonia, cirrhosis, anemia, urological infection) [8].

Lately. we used a bigger dataset, containing a total of 22160 medical records and we were able to learn about 50 diagnosis. We used the discharge letters for each medical record as labels. We called this annotated dataset *GT - D.L.* (Ground Truth - Discharge letters). As usually data were split in training, cross-validation and test sets and we measured the F-Measure of Greg predictions. On *GT - D.L.* we obtained poor results, not as good as we would have expected. Some of the results are shown in Figure 3, in terms of F-Measure. Our investigation of the data, however, suggested that in many cases the quality of the results have been underestimated. In essence, in several cases Greg suggested a more thorough set of diagnoses than the one indicated by the doctor in discharge letters. As an example, this happened frequently with patients suffering from anaemia, which is often associated with cirrhosis, even though doctors had not explicitly mentioned that specific diagnosis in the discharge letter.

We therefore conducted a second experiment. We asked our team of doctors to review the set of diagnoses associated with patient profiles used for the test. In essence, our doctors made sure that all relevant diagnoses were appropriately mentioned, including those that the hospital doctors had omitted in the discharge letter. We called this manually annotated dataset *GT - Doct.* (Ground Truth - Doctors). Figure 3 reports Greg's results over this revised dataset. As it can be seen, we obtained an F-Measure for each diagnosis always above the 95%.

To summarize, our preliminary tests show that Greg can effectively achieve high accuracy in its predictions. In addition, it may effectively assist doctors in formulating their diagnoses, by providing systematic suggestions.

5 Conclusions: Opportunities and Lessons Learned

We believe that Greg can be a valid and useful tool to assist doctors in the diagnostic process. Given its ability to learn diagnostic suggestions at scale, we envision several main scenarios of use for the system in a medical facility:

- We believe Greg can be of particular help in ER, during the triage and first diagnostic phase; in particular, based on first evidences about the patient, it may help the ER operator to identify a few pathologies to it is worth exploring, perhaps with the help of specialized colleague.
- Then, we envision interesting opportunities related to the use of Greg in the diagnosis of rare pathologies; these are especially difficult to capture by a learning algorithm, because, by definition, there are only a few training examples to use, and therefore a special treatment is required. Still, we believe that supporting doctors – especially younger ones, that might have less experience in diagnosing these pathologies – in this respect is an important field of application.
- In medical institutions that rely on standardized clinical pathways or *integrated care pathways (ICPs)* – PDTAs in Italy – Greg may be used to quickly suggest which parts of a pathway need to be explored, and which ones can be excluded based on the available evidence.
- Finally, Greg may be used as a second-opinion tool, i.e., after the doctor has formulated her/his diagnosis, for the purpose of double checking that all possibilities have been considered.

While in our opinion all of these represent areas in which Greg can be a valid support tool for the doctor, we would like to put them in context by discussing what we believe to be the most important lessons we have learned so far.

On the one side, the development of Greg has taught us a basic and important lesson: in many cases, probably the majority, the basic workings of the diagnostic process employed by human doctors is indeed reproducible by an automatic algorithm.

In fact, it is well known that doctors tend to follow a decision process that looks for specific indicators within the patient profile – e.g., values of laboratory tests, or specific symptoms – and decides to consider or excludes pathologies based on them. As fuzzy as this process may be, as any other human-thinking process, to our surprise we learned that for a large number of pathologies this process provides a perfect opportunity for the employment of a machine learning algorithm, which, in turn, may achieve very good accuracy in mimicking the human decision process, with the additional advantage of scale – Greg can be trained to learn very high numbers of diagnostic suggestions. In this respect, ironically quoting Gregory House, we might be tempted to state that "Humanity is overrated", indeed.

However, our experiences also led us to find that there are facets of the diagnostic process that are inherently related to intuition, experience, and human factors. These are, by nature, impossible to capture by an automatic algorithm. Therefore, our ultimate conclusion is that humanity is not overrated, and that Greg can indeed provide useful support in the diagnostic process, but it cannot and should not be considered as a replacement of an expert human doctor.

References

1. R. C. Deo. Machine learning in medicine. *Circulation*, 132(20):1920–1930, 2015.
2. F. Geerts, G. Mecca, P. Papotti, and D. Santoro. Mapping and Cleaning. In *Proceedings of the IEEE International Conference on Data Engineering - ICDE*, 2014.
3. F. Geerts, G. Mecca, P. Papotti, and D. Santoro. That's All Folks! LLUNATIC Goes Open Source. In *Proceedings of the International Conference on Very Large Databases - VLDB*, 2014.
4. J. He, E. Veltri, D. Santoro, G. Li, G. Mecca, P. Papotti, and N. Tang. Interactive and deterministic data cleaning. In *Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016*, pages 893–907, 2016.
5. T. Heinis, A. Ailamaki, et al. Data infrastructure for medical research. *Foundations and Trends in Databases*, 8(3):131–238, 2017.
6. A. Holzinger. Machine learning for health informatics. In *Machine Learning for Health Informatics*, pages 1–24. Springer, 2016.
7. I. Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, 23(1):89–109, 2001.
8. P. Lapadula, G. Mecca, D. Santoro, L. Solimando, and E. Veltri. Humanity Is Overrated. or Not. Automatic Diagnostic Suggestions by Greg, ML. In *New Trends in Databases and Information Systems*, pages 305–313. Springer International Publishing, 2018.
9. R. H. Miller and I. Sim. Physicians' use of electronic medical records: barriers and solutions. *Health affairs*, 23(2):116–126, 2004.
10. O. Mohammed and R. Benlamri. Developing a semantic web model for medical differential diagnosis recommendation. *Journal of medical systems*, 38(10):79, 2014.

11. N. Peek, C. Combi, R. Marin, and R. Bellazzi. Thirty years of artificial intelligence in medicine (aime) conferences: A review of research themes. *Artificial intelligence in medicine*, 65(1):61–73, 2015.
12. P. Rajpurkar, A. Y. Hannun, M. Haghpanahi, C. Bourn, and A. Y. Ng. Cardiologist-level arrhythmia detection with convolutional neural networks. *arXiv preprint arXiv:1707.01836*, 2017.
13. P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.
14. J. Soni, U. Ansari, D. Sharma, and S. Soni. Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications*, 17(8):43–48, 2011.
15. I. Steadman. IBM's Watson is better at diagnosing cancer than human doctors. *WIRED*, 2013.