RESEARCH ARTICLE

REVISED Collaboration between a human group and artificial intelligence can improve prediction of multiple sclerosis course: a proof-of-principle study [version 2; referees: 1 approved, 2 approved with reservations]

Andrea Tacchella[1*], Silvia Romano[2*], Michela Ferraldeschi [2], Marco Salvetti[2,3], Andrea Zaccaria[1], Andrea Crisanti[4], Francesca Grassi [5]

[1]Institute for Complex Systems, National Research Council - UOS Sapienza, Rome, 00185, Italy
[2]Center for Experimental Neurological Therapies (CENTERS), Dept. of Neurosciences, Mental Health and Sensory Organs, Sapienza University of Rome, Rome, 00189, Italy
[3]IRCCS Neuromed , Istituto Neurologico Mediterraneo, Pozzilli, 86077, Italy
[4]Department of Physics, Sapienza University of Rome, Rome, 00185, Italy
[5]Institute Pasteur-Cenci Bolognetti Foundation, Dept. Physiology and Pharmacology, Sapienza University of Rome, Rome, 00185, Italy

[*] Equal contributors

v2 **First published:** 22 Dec 2017, **6**:2172 (doi: 10.12688/f1000research.13114.1)
Latest published: 01 Aug 2018, **6**:2172 (doi: 10.12688/f1000research.13114.2)

Abstract

Background: Multiple sclerosis has an extremely variable natural course. In most patients, disease starts with a relapsing-remitting (RR) phase, which proceeds to a secondary progressive (SP) form. The duration of the RR phase is hard to predict, and to date predictions on the rate of disease progression remain suboptimal. This limits the opportunity to tailor therapy on an individual patient's prognosis, in spite of the choice of several therapeutic options. Approaches to improve clinical decisions, such as collective intelligence of human groups and machine learning algorithms are widely investigated.
Methods: Medical students and a machine learning algorithm predicted the course of disease on the basis of randomly chosen clinical records of patients that attended at the Multiple Sclerosis service of Sant'Andrea hospital in Rome.
Results: A significant improvement of predictive ability was obtained when predictions were combined with a weight that depends on the consistence of human (or algorithm) forecasts on a given clinical record.
Conclusions: In this work we present proof-of-principle that human-machine hybrid predictions yield better prognoses than machine learning algorithms or groups of humans alone. To strengthen and generalize this preliminary result, we propose a crowdsourcing initiative to collect prognoses by physicians on an expanded set of patients.

Keywords

Multiple sclerosis, Machine learning, Random Forest, collective intelligence, Hybrid predictions, Crowdsourcing

Open Peer Review

Referee Status: ✓ ? ?

	Invited Referees		
	1	**2**	**3**
REVISED **version 2** published 01 Aug 2018			
version 1 published 22 Dec 2017	✓ report	? report	? report

1 **Bruno Bonetti**, Azienda Ospedaliera Universitaria Integrata, Italy

2 **Roger Tam**, University of British Columbia, Canada

3 **Bjoern Menze**, Technische Universität München, Germany

Discuss this article

Comments (0)

This article is included in the INCF gateway.

Corresponding authors: Andrea Crisanti (andrea.crisanti@phys.uniroma1.it), Francesca Grassi (francesca.grassi@uniroma1.it)

Author roles: Tacchella A: Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Romano S**: Data Curation, Investigation, Supervision, Writing – Review & Editing; **Ferraldeschi M**: Data Curation, Investigation; **Salvetti M**: Conceptualization, Funding Acquisition, Project Administration, Supervision, Writing – Review & Editing; **Zaccaria A**: Conceptualization, Formal Analysis, Methodology, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Crisanti A**: Conceptualization, Formal Analysis, Funding Acquisition, Methodology, Supervision, Writing – Review & Editing; **Grassi F**: Conceptualization, Project Administration, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: CENTERS is a special project of, and is supported by, Fondazione Italiana Sclerosi Multipla. AT and AZ acknowledge funding from the “CNR Progetto di Interesse CRISIS LAB”. All funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Tacchella A, Romano S, Ferraldeschi M *et al.* **Collaboration between a human group and artificial intelligence can improve prediction of multiple sclerosis course: a proof-of-principle study [version 2; referees: 1 approved, 2 approved with reservations]** *F1000Research* 2018, **6**:2172 (doi: 10.12688/f1000research.13114.2)

First published: 22 Dec 2017, **6**:2172 (doi: 10.12688/f1000research.13114.1)

REVISED **Amendments from Version 1**

This version addresses most of the points raised by the Reviewers, to whom we are grateful for useful criticisms.

We included several references and statements to justify why we only considered clinical data, not MRI and to document the use of crowdsourcing initiatives in the diagnostic realm.

We give more details on how we built the database and on the procedure used to obtain hybrid forecasts, which represent the core of the paper.

Finally, we repeatedly state that this work represents a proof of concept of the value of hybrid forecasts, a preliminary step towards deeper studies.

See referee reports

Introduction

The natural course of multiple sclerosis (MS) is extremely variable, ranging from extremely mild to very aggressive forms. Most patients experience an initial relapsing-remitting (RR) phase, in which symptoms appear and fade. Eventually, remissions fail and the disease proceeds to a secondary progressive (SP) form, leading to incremental disability. The palette of disease-modifying treatments is becoming relatively large, in principle opening the possibility to tailor the therapy to meet the specific needs of each patient. Unfortunately, the accuracy of parameters to predict the rate of disease progression remains suboptimal.

Being all the above therapies preventive, in the absence of exact prognostic indicators we have to accept that a proportion of patients is either under- or over-treated. This is a serious concern as the disease can be severely disabling, and some of the available therapies can lead to adverse events that can be worse than the disease itself. Thus, the possibility to formulate a prognosis as exact as possible is becoming increasingly appealing.

In the clinics, as in any other fields of human knowledge, innovative approaches based on machine learning and collective reasoning methods are used in an attempt to succeed where traditional methods of forecasting failed. Machine learning algorithms catch complex relations among existing data to an extent beyond standard regression models. Good performances have been obtained for the diagnosis of Parkinson's disease and the prognosis of disease progression in amyotrophic lateral sclerosis (Dinov *et al.*, 2016; Küffner *et al.*, 2015). For MS, machine learning algorithms can correctly classify disease course in about 70% of cases of both clinically definite MS and of clinically isolated syndrome (Fiorini *et al.*, 2015; Wottschel *et al.*, 2014; Zhao *et al.*, 2017), a good result that still requires improvement to become of clinical value.

Through collective reasoning, or collective intelligence, groups of lay people may perform as well as experts. In principle, the larger the group, the higher the prediction accuracy (see for review Ponsonby & Mattingly, 2015), which led to the development of several crowdsourcing initiatives. Possibly, the forerunner was FOLDIT study on protein folding (Cooper *et al.*, 2010), but crowdsourcing has been exploited also for diagnostic purposes in pathologies, such as breast cancer (Candido dos Reis *et al.*, 2015), skin cancer (King *et al.*, 2013) or ophtalmology (Wang *et al.*, 2016). However, when expert people are involved, even small groups can outperform the best among them, at least when a yes/no answer to well-defined diagnostic questions is requested based on radiographic/histological images, (Kurvers *et al.*, 2016; Sonabend *et al.*, 2017; Wolf *et al.*, 2015). Studies with medical students show that working in pairs, either interacting while responding (Hautz *et al.*, 2015) or aggregating responses *ex post* (Kämmer *et al.*, 2017), ameliorates diagnostic ability, with further improvements when group size increases (Hautz *et al.*, 2015; Kämmer *et al.*, 2017), in line with the core idea of Collective intelligence. Similar results have been obtained also for prognoses on critically ill patients (Poses *et al.*, 1990)

Combination of human and machine predictions into hybrid forecasts exploits human intuitive reasoning and computer classification capabilities, potentially boosting both. Indeed, at least in the case of predicting the course of actions in American football games within the frame of prediction markets, hybrid groups performed better than either humans or computers. (Nagar & Malone, 2011). In this paper, we report the promising results of a preliminary study on the combination of predictions made by humans with those of a machine learning algorithm on the progression of multiple sclerosis in a set of patients. Both agents (humans and computers) considered clinical data typically available to neurologists during routine visits. Magnetic Resonance Imaging (MRI) data were not included as more clinical than radiological exams are routinely performed (on average 3 visits per year *vs.* 1 MRI). Moreover, images, acquired and analysed at specialized centres can improve the algorithm performance (Zhao *et al.*, 2017), but in real world imaging data usually lack the standardization required for analysis, for instance in term of head position reproducibility (Weinstock-Guttman *et al.*, 2018), and research-grade image analysis is not routinely performed. Conversely, clinical data have recently been shown to have good predictive value (Goodin *et al.*, 2018). Machine learning and collective intelligence performed almost equally well, but their combination yielded a small, yet statistically significant, improvement in the reliability of the forecasts on disease evolution over different time periods.

These results indicate that it is worth deepening the study of human and machine clinical predictions, as well as the potentiality of hybrid predictions, for which we propose a crowdsourcing approach on a platform specifically designed for this analysis (*DiagnoShare*).

Methods and results

Dataset structure

Our dataset is composed by clinical records gathered during 527 visits of 84 outpatients followed at the Multiple Sclerosis service of Sant'Andrea hospital in Rome. All patients had clinically definite MS in the RR stage at the time of the visit(s) included in the database and transitioned to the SP phase at some time point. Parameters evaluated during each neurological visit are listed in Supplementary Table 1. Numerical values were provided for each parameter, referring to age, time to complete a task, clinical score or presence/absence of each symptom. For each visit, we noted if the patient was in RR or SP stage after 180, 360 and 720 days, so that predictions could be compared with the *true* progression of disease in each patient, reported in Supplementary File: TrueOutcomes.xlsx, where 0 means "still in RR phase", 1 indicates "transitioned to SP phase". Notice that several patients reached the SP MS stage after the last visit included in the database, so that the number (percentage of entries) of "1" records is 65 (12.3%), 125 (23.7%), and 211 (40.0%) at 180, 360, and 720 days, respectively. Data potentially revealing the identity of the patients was removed from the shared database.

Ethics

Use of database for research purposes was authorized by the Ethical committee of Sapienza University (Authorization n. 4254_ 2016, dated November 2, 2016).

Classification with machine learning

Having a correctly labelled dataset (Supplementary Table 1), in which each entry is associated to the outcome, we used the Random forest supervised approach to classification (Breiman, 2001; Liaw & Weiner, 2002), using the *Scikit-learn* toolbox version 0.16.1.

To benchmark the performance of the trained models, we used a modified *leave-one-out approach*. Since data was limited (a set of 527 records), and not independent, as it had been obtained from 84 patients, with a simple random *leave-one-out* the training set would be composed of many correlated same-patient data. Even worse, some of the data from patients present in the training set would be used to validate the model in the benchmarking stage. As a consequence, the model would overfit the training data, misleadingly showing very good performance. Being presented with many data from the same patient, the model optimizes its ability in recognizing patients themselves, through their highly correlated clinical variables.

To avoid these problems, we used a modified *leave-one-out* approach, training the algorithm with the following rules:

1. We excluded all visits from one patient from the dataset
2. We built 50 training sets, each composed by 83 records, taking care to include only one clinical record (randomly chosen) for every remaining patient
3. We trained 50 Random Forest models, one for each training set.
4. We computed the probability of the transition from RR to SP by averaging the predictions of the 50 models on all the visits of the excluded patient. Predictions consisted in scores from 0 (Extremely unlikely) to 1 (Highly probable).

We repeated the procedure for the 84 patients, obtaining an estimation of the probability of the RR to SP transition for each of the 527 clinical records. Three different prediction delays were considered, namely 180, 360 and 720 days. Results obtained are presented in Supplementary File: RF_Predictions.xlsx. The performance of the model was estimated by the Area Under the "*Receiver Operating Characteristic*" (ROC) Curve (AUC) computed on all the 527 examples. The AUC values obtained are shown in Table 1.

Human predictions

Forty-two medical students in the final two years of their course (Sapienza University, Rome Italy, based within Sant'Andrea hospital), volunteered to participate in the task. All were familiar with clinical records in general, and were instructed on the meaning of each entry present in the medical records of MS patients. This part of the study was approved by the Ethical Committee of the Department of Physiology and Pharmacology, Sapienza University on July 13, 2017.

For adequate comparison with computer predictions, students evaluated 50 medical records, collected in a questionnaire, randomly extracted from the same dataset used for machine learning and estimated the probability that the patient would progress to the SP phase within 180, 360 and 720 days. Scores were from 0 (Extremely unlikely) to 5 (Highly probable). Predictions (see Supplementary file Student_Predictions.xlsx) were analysed, using the AUC.

On average, each clinical record was evaluated by 4 of the 42 students.

Table 1. Predictions on disease course by different agents.

Agent	180 days	360 days	720 days
Random Forest	0.710	0.670	0.679
Singles (n=42)	*0.57 ± 0.15*	*0.57 ± 0.11*	*0.57 ± 0.10*
Pairs	*0.68*	*0.65*	*0.65*
Group	*0.703*	*0.667*	*0.666*
Hybrid predictions	0.725*	0.694*	0.696*

For each clinical record, the indicated agents evaluated the probability that disease evolved from the RR to the SP phase after 180, 360 or 720 days. Data represent the AUC values obtained for each method. *: P<0.001 when compared to *Group* or Random Forest values at the same time points.

Predictions were less accurate than those proposed by machine learning (Table 1). Standard deviation was larger for the 180 day time point, indicating that opinions on the long-term evolution of the disease are more widely shared, although they are not more precise. To evaluate the impact of collective intelligence, we measured the performance of *Pairs*, considering all visits evaluated by at least two individual students, randomly selecting only 2 scores when more were available. The prognoses were averaged before computing the AUC, which showed a marked increase (Table 1). Aggregation of all singles (*Group*) yielded a further small increase in the performance of the forecasting (Table 1), which almost equalled that of random forest algorithm.

Hybrid predictions

We next integrated human and computer predictions into a hybrid prediction, which combines human clinical reasoning with the classification approach of machine learning algorithms. These different "ways of reasoning" possibly lead to quite divergent predictions on individual cases, a complementarity that should be exploited taking the difference into account when creating hybrid predictions.

The simplest approach to aggregate forecast is performing a linear or weighted average of the predictions released by humans or computer. For each clinical record at a given time point (180, 360 and 720 days), the final forecast by either agent is the average of "unitary predictions" given by several individuals or decision trees. If "unitary predictions" of one agent are highly concordant, it means that the prediction is quite obvious for the agent, suggesting that it is probably correct. We therefore ranked forecasts on clinical records in order of concordance of "unitary predictions", for the two agents separately. Then, a normalized ranking was assigned, ranging from 1 for the most consistent predictions to 0 for the most scattered and ranks were squared to emphasize the contribution of the most consistent agent. The hybrid prediction score for each clinical record was then obtained by summing the two squared ranks.

Note that a linear combination of rankings resulted in a worse performance of hybrid predictions, as the information about the most consistent prediction between the two agents was be lost. A similarly degraded performance was observed when predictions were not ranked.

Since our dataset is relatively small, as is the number of students that evaluated the clinical records, we used a bootstrap procedure to evaluate the statistical significance of the improvement. The bootstrap (Efron & Tibshirani, 1994; Felsenstein, 1985) consists in random sampling of the dataset that allows the estimation of confidence intervals.

As shown in Table 1 and Figure 1, hybrid predictions yielded a small but statistically significant (P<0.001) improvement in the prediction of disease course in time. Significance was evaluated from confidence limits using standard methods (Altman & Bland, 2011).

Dataset 1. True outcome of patients, indexed as clinical records

http://dx.doi.org/10.5256/f1000research.13114.d188355

More than one clinical record is pertinent to each patient. T_180, T_360, T_720: clinical conditions of the patient 180, 360 and 720 days after the visit in which clinical record was obtained. 0: still in RR phase; 1: transitioned to SP phase.

Dataset 2. Predictions on individual clinical records made by medical students

http://dx.doi.org/10.5256/f1000research.13114.d188356

Each student worked on a questionnaire (lines labelled "questionnaire", column B.) listing 50 clinical reports (lines labelled "Clinical report N", columns B to AY) and made a prediction on the probability of RR –to–SP transition within 180, 360 and 720 days (lines labelled Prediction @ 180, 360, 720, columns B to AY)

The numbering of Clinical reports is the same used in Dataset 1.

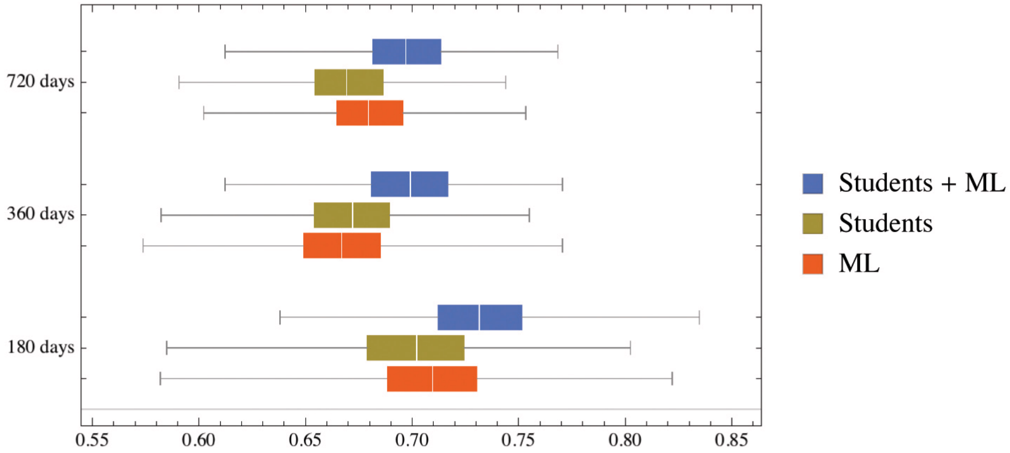

Figure 1. Hybrid Students – Machine Learning predictions outperform both human group and computer alone. The box plot shows the distribution of the AUC obtained from the bootstrap. In particular, the colored boxes correspond to quartiles, while the lines show the full range of the generated AUCs.

Dataset 3. Predictions on individual clinical records made by a Random Forest algorithm

http://dx.doi.org/10.5256/f1000research.13114.d188357

Score_180, Score_360, Score_720: Probability that the patient will transition to SP phase within 180, 360 and 720 days after the visit in which clinical record was obtained. The numbering of Clinical reports is the same used in Dataset 1.

Discussion

A number of studies have investigated the possibility to increase the appropriateness of clinical decisions through collective intelligence of human groups (for instance, Kurvers *et al.*, 2016; Sonabend *et al.*, 2017; Wolf *et al.*, 2015) or machine learning algorithms. The latter approach has been used in a great variety of tasks, and its value in the medical realm is possibly overstated (Chen & Asch, 2017). However, machine learning methods performed well for prognostic predictions (Küffner *et al.*, 2015; Zhao *et al.*, 2017). In particular, the Random forest approach provided good predictions on ALS course (Küffner *et al.*, 2015).

In this work we present proof-of-principle that human-machine hybrid predictions attain prognostic ability above that of machine learning algorithms and groups of humans alone.

The duration of the RR phase before its shift into progression has always been difficult to predict, and possibly the random occurrence of relapses (Bordi *et al.*, 2013) contributes to the lack of univocal indicators. No approach, no matter how good, can yield certainty when cause-effect relations are unknown. Thus, our aim has been to obtain predictions on the probability that MS patients in the RR phase will convert to a SP form within a certain time frame. Predictions on the course of real patients were provided by medical students and a random forest algorithm. A significant improvement of predictive ability was obtained when predictions were combined in a non-linear manner, with a weight that depends on the consistence of human (or algorithm) forecasts on a given clinical record.

This result can be considered in agreement with several studies on different medical issues showing that predictor's confidence correlates very well with the correctness of the prediction (Detsky *et al.*, 2017; Hautz *et al.*, 2015; Kämmer *et al.*, 2017; Kurvers *et al.*, 2016). Indeed, the concordance of different members of a given group (students or runs of the random forest model) can be taken as indicating that the agent is "sure" of the forecast. Further work investigating the best ways to combine predictions of different agents is ongoing.

In spite of the relatively basic machine learning technique used, the small number of students involved and their limited clinical knowledge, this work suggests that hybrid predictions can be useful to improve the prognosis of MS course. A deeper study is therefore of interest, to evaluate how general this conclusion is. To recruit more and more skilled humans, we propose a crowdsourcing initiative called *DiagnoShare* that is being advertised among physicians.

A reliable tool to predict MS progression can be of aid to clinicians to tailor therapy to each patient, but also in clinical trials, to evaluate whether drugs modify the estimated outcome of each enrolled patient, as proposed for ALS (Küffner *et al.*, 2015).

In the long run, it is possible that further developments in our ability to combine collective reasoning and machine predictions will have a profound impact also on the organization and management of medical care, particularly in hospital settings.

Data availability

Dataset 1: *True outcome of patients, indexed as clinical records*. More than one clinical record is pertinent to each patient. T_180, T_360, T_720: clinical conditions of the patient 180, 360 and 720 days after the visit in which clinical record was obtained. 0: still in RR phase; 1: transitioned to SP phase. DOI: 10.5256/f1000research.13114.d188355 (Tacchella *et al.*, 2017a)

Dataset 2: *Predictions on individual clinical records made by medical students.* Each student worked on a questionnaire (lines labelled "questionnaire", column B.) listing 50 clinical reports (lines labelled "Clinical report N", columns B to AY) and made a prediction on the probability of RR –to–SP transition within 180, 360 and 720 days (lines labelled Prediction @ 180, 360, 720, columns B to AY)

The numbering of Clinical reports is the same used in Dataset 1. DOI: 10.5256/f1000research.13114.d188356 (Tacchella *et al.*, 2017b)

Dataset 3: *Predictions on individual clinical records made by a Random Forest algorithm.* Score_180, Score_360, Score_720: Probability that the patient will transition to SP phase within 180, 360 and 720 days after the visit in which clinical record was obtained. The numbering of Clinical reports is the same used in Dataset 1. DOI: 10.5256/f1000research.13114.d188357 (Tacchella *et al.*, 2017c)

Competing interests

No competing interests were disclosed.

Grant information

CENTERS is a special project of, and is supported by, Fondazione Italiana Sclerosi Multipla. AT and AZ acknowledge funding from the "CNR Progetto di Interesse CRISIS LAB".

All funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

We thank all the students that participated in the project.

Supplementary material

Supplementary Table 1: Parameters evaluated for each patient and included in clinical records.

Click here to access the data.

References

Altman DG, Bland JM: **How to obtain the P value from a confidence interval.** *BMJ.* 2011; **343**: d2304.
PubMed Abstract | Publisher Full Text

Bordi I, Umeton R, Ricigliano VA, *et al.*: **A mechanistic, stochastic model helps understand multiple sclerosis course and pathogenesis.** *Int J Genomics.* 2013; **2013**: 910321.
PubMed Abstract | Publisher Full Text | Free Full Text

Breiman L: **Random Forests.** *Mach Learn.* 2001; **45**(1): 5–32.
Publisher Full Text

Candido Dos Reis FJ, Lynn S, Ali HR, *et al.*: **Crowdsourcing the General Public for Large Scale Molecular Pathology Studies in Cancer.** *EBioMedicine.* 2015; **2**(7): 681–689.
PubMed Abstract | Publisher Full Text | Free Full Text

Chen JH, Asch SM: **Machine Learning and Prediction in Medicine - Beyond the Peak of Inflated Expectations.** *N Engl J Med.* 2017; **376**(26): 2507–2509.
PubMed Abstract | Publisher Full Text | Free Full Text

Cooper S, Khatib F, Treuille A, *et al.*: **Predicting protein structures with a multiplayer online game.** *Nature.* 2010; **466**(7307): 756–60.
PubMed Abstract | Publisher Full Text | Free Full Text

Detsky ME, Harhay MO, Bayard DF, *et al.*: **Discriminative Accuracy of Physician and Nurse Predictions for Survival and Functional Outcomes 6 Months After an ICU Admission.** *JAMA.* 2017; **317**(21): 2187–2195.
PubMed Abstract | Publisher Full Text | Free Full Text

Dinov ID, Heavner B, Tang M, *et al.*: **Predictive Big Data Analytics: A Study of Parkinson's Disease Using Large, Complex, Heterogeneous, Incongruent, Multi-Source and Incomplete Observations.** *PLoS One.* 2016; **11**(8): e0157077.
PubMed Abstract | Publisher Full Text | Free Full Text

Efron B, Tibshirani RJ: **An introduction to the bootstrap.** CRC press, 1994.
Reference Source

Felsenstein J: **CONFIDENCE LIMITS ON PHYLOGENIES: AN APPROACH USING THE BOOTSTRAP.** *Evolution.* 1985; **39**(4): 783–791.
PubMed Abstract | Publisher Full Text

Fiorini S, Verri A, Tacchino A, *et al.*: **A machine learning pipeline for multiple sclerosis course detection from clinical scales and patient reported outcomes.** *Conf Proc IEEE Eng Med Biol Soc.* 2015; **2015**: 4443–6.
PubMed Abstract | Publisher Full Text

Goodin DS, Reder AT, Traboulsee AL, *et al.*: **Predictive validity of NEDA in the 16- and 21-year follow-up from the pivotal trial of interferon beta-1b.** *Mult Scler.* 2018; 1352458518773511.
PubMed Abstract | Publisher Full Text

Hautz WE, Kämmer JE, Schauber SK, *et al.*: **Diagnostic performance by medical students working individually or in teams.** *JAMA.* 2015; **313**(3): 303–304.
PubMed Abstract | Publisher Full Text

Kämmer JE, Hautz WE, Herzog SM, *et al.*: **The Potential of Collective Intelligence in Emergency Medicine: Pooling Medical Students' Independent Decisions Improves Diagnostic Performance.** *Med Decis Making.* 2017; **37**(6): 715–724.
PubMed Abstract | Publisher Full Text

King AJ, Gehl RW, Grossman D, *et al.*: **Skin self-examinations and visual identification of atypical nevi: comparing individual and crowdsourcing approaches.** *Cancer Epidemiol.* 2013; 37(6): 979–84.
PubMed Abstract | Publisher Full Text | Free Full Text

Kurvers RH, Herzog SM, Hertwig R, *et al.*: **Boosting medical diagnostics by pooling independent judgments.** *Proc Natl Acad Sci U S A.* 2016; **113**(31): 8777–8782.
PubMed Abstract | Publisher Full Text | Free Full Text

Küffner R, Zach N, Norel R, *et al.*: **Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression.** *Nat Biotechnol.* 2015; **33**(1): 51–57.
PubMed Abstract | Publisher Full Text

Liaw A, Wiener M: **Classification and regression by random Forest.** *R News.* 2002; **2**: 18–22.
Reference Source

Nagar Y, Malone TW: **Making Business Predictions by Combining Human and Machine Intelligence in Prediction Markets.** *Proceedings of the International Conference on Information Systems ICIS 2011.* Shanghai, China. 2011.
Reference Source

Ponsonby AL, Mattingly K: **Evaluating New Ways of Working Collectively in Science with a Focus on Crowdsourcing.** *EBioMedicine.* 2015; **2**(7): 627–8.
PubMed Abstract | Publisher Full Text | Free Full Text

Poses RM, Bekes C, Winkler RL, *et al.*: **Are two (inexperienced) heads better than one (experienced) head? Averaging house officers' prognostic judgments for critically ill patients.** *Arch Intern Med.* 1990; **150**(9): 1874–8.
PubMed Abstract | Publisher Full Text

Sonabend AM, Zacharia BE, Cloney MB, *et al.*: **Defining Glioblastoma Resectability Through the Wisdom of the Crowd: A Proof-of-Principle Study.** *Neurosurgery.* 2017; **80**(4): 590–601.
PubMed Abstract | Publisher Full Text

Tacchella A, Romano S, Ferraldeschi M, *et al.*: **Dataset 1 in: Collaboration between a human group and artificial intelligence can improve prediction of multiple sclerosis course: a proof-of-principle study.** *F1000Research.* 2017a.
http://www.doi.org/10.5256/f1000research.13114.d188355

Tacchella A, Romano S, Ferraldeschi M, *et al.*: **Dataset 2 in: Collaboration between a human group and artificial intelligence can improve prediction of multiple sclerosis course: a proof-of-principle study.** *F1000Research.* 2017b.
http://www.doi.org/10.5256/f1000research.13114.d188356

Tacchella A, Romano S, Ferraldeschi M, *et al.*: **Dataset 3 in: Collaboration between a human group and artificial intelligence can improve prediction of multiple sclerosis course: a proof-of-principle study.** *F1000Research.* 2017c.
http://www.doi.org/10.5256/f1000research.13114.d188357

Wang X, Mudie L, Brady CJ: **Crowdsourcing: an overview and applications to ophthalmology.** *Curr Opin Ophthalmol.* 2016; **27**(3): 256–61.
PubMed Abstract | Publisher Full Text | Free Full Text

Weinstock-Guttman B, Medin J, Khan N, *et al.*: **Assessing 'No Evidence of Disease Activity' Status in Patients with Relapsing-Remitting Multiple Sclerosis Receiving Fingolimod in Routine Clinical Practice: A Retrospective Analysis of the Multiple Sclerosis Clinical and Magnetic Resonance Imaging Outcomes in the USA (MS-MRIUS) Study.** *CNS Drugs.* 2018; **32**(1): 75–84.
PubMed Abstract | Publisher Full Text | Free Full Text

Wolf M, Krause J, Carney PA, *et al.*: **Collective intelligence meets medical decision-making: the collective outperforms the best radiologist.** *PLoS One.* 2015; **10**(8): e0134269.
PubMed Abstract | Publisher Full Text | Free Full Text

Wottschel V, Alexander DC, Kwok PP, *et al.*: **Predicting outcome in clinically isolated syndrome using machine learning.** *Neuroimage Clin.* 2014; **7**: 281–7.
PubMed Abstract | Publisher Full Text | Free Full Text

Zhao Y, Healy BC, Rotstein D, *et al.*: **Exploration of machine learning techniques in predicting multiple sclerosis disease course.** *PLoS One.* 2017; **12**(4): e0174866.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Referee Status: ✓ ? ?

Version 1

Referee Report 06 April 2018

doi:10.5256/f1000research.14226.r31371

?

Bjoern Menze
Department of Informatics, Technische Universität München, Munich, Germany

General:

I think exploring how to fuse multiple expert opinions is a very interesting line of research in computer aided diagnostics. Here, the authors test how to make use of lay persons, and I would agree that there are many tasks when a (briefly trained) lay person or non-expert can contribute significantly to an analytical task.

In the application here, I would be rather critical about this approach, though. For example, the authors write "through collective reasoning, or collective intelligence, groups of lay people may perform as well as experts." I would not agree, by any means. How would a lay person without training be able to distinguish, for example, a stroke related white matter hyper-intensity from an MS lesion? Or even a large MR artifact? Averaging will reduce variance in prediction, but the individual prediction itself has to be unbiased. In other words: the layman predictor has to be correct on average. But how would they possibly be in case they have no idea about how to read these data? Moreover, the authors point out that "studies with medical students show that working in pairs ameliorates diagnostic ability". Is this because of a better discussion of the decision? With two subjects it cannot be the power of large numbers that this study relies on.

Instead of exploring how to fuse layman's decisions, I would recommend the authors to explore how to fuse decisions of different algorithms, or from neurologists of different training/seniority level, or decisions based on different sources.

Technical:

Experimental setup and evaluation: The authors describe a "leave-one patient-out" cross-validation as an innovation of their study. While this is a good approach, it is not new.

Algorithm and training: There are different classes - what is the distribution of those classes for the 84 patients? What is in the reports? Numbers? Free text? What features are input to the random forest algorithm? How many features at all? How did you train the algorithm (parameters "mtry", why 50 trees?) Without this information it is difficult to assess whether the performance of your random forest is bad (i.e., close to layman's predictions) because of an suboptimal training procedure, or because this is a hard problem indeed.

Fusion rule: (Described in the section "To compare the two sets... of the most consistent agent.") I don't understand what you do. How does summing a squared ranking lead to a prediction score? I assume you are using the normalized (and squared) ranking as a sort of weight? Why do you square the rankings? What happens when you use other non-linear transformations? Is there any way you illustrate the distributions so that we can follow your reasoning? How about presenting simple rules like averaging, or majority voting at least as a baseline we can compare against?

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
No

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
No

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 26 Jul 2018

Francesca Grassi, Sapienza University of Rome, Italy

First of all, thank you for taking the time to read our work and to give useful comments. We hope that our responses will clear your doubts. We list below the changes introduced in the new version, prompted by your observations. We hope that you agree with us that it is improved

In the application here, I would be rather critical about this approach, though. For example, the authors write "through collective reasoning, or collective intelligence, groups of lay people may perform as well as experts." I would not agree, by any means. How would a lay person without training be able to distinguish, for example, a stroke related white matter hyper-intensity from an MS lesion? Or even a large MR artifact? Averaging will reduce variance in prediction, but the individual prediction itself has to be unbiased. In other words: the layman predictor has to be correct on average. But how would they possibly be in case they have no idea about how to read these data?
ANSWER

Although your point of view is quite understandable, there is a large body of literature on the topic of collective intelligence. In the hope to overcome your skepticism on this point, we added some more references to published work on diagnostic crowdsourcing initiatives.

Moreover, the authors point out that "studies with medical students show that working in pairs ameliorates diagnostic ability". Is this because of a better discussion of the decision? With two subjects it cannot be the power of large numbers that this study relies on.
ANSWER
It is now better explained that the two quoted studies use different methods: real pairs in one, aggregated opinions in the other, yet both obtain better performances. Authors do not discuss the underlying processes, so we cannot indicate the real reason of a better performance.

Instead of exploring how to fuse layman's decisions, I would recommend the authors to explore how to fuse decisions of different algorithms, or from neurologists of different training/seniority level, or decisions based on different sources.
ANSWER
Thank you for your suggestion. Understanding that we have to deepen out study (now repeatedly stated throughout the paper) we have developed DiagnoShare to obtain the predictions of clinicians of different expertise and we are investigating the performance of algorithm combinations.

Experimental setup and evaluation: The authors describe a "leave-one patient-out" cross-validation as an innovation of their study. While this is a good approach, it is not new.
ANSWER
Thank you for your observation. Indeed, it is better to define our approach as a modified leave-one-out method. It is modified, because we not only left one patient out, we also included only one record for each of the remaining patients.

Algorithm and training: There are different classes - what is the distribution of those classes for the 84 patients? What is in the reports? Numbers? Free text? What features are input to the random forest algorithm? How many features at all? How did you train the algorithm (parameters "mtry", why 50 trees?) Without this information it is difficult to assess whether the performance of your random forest is bad (i.e., close to layman's predictions) because of an suboptimal training procedure, or because this is a hard problem indeed
ANSWER
Thank you for pointing out that this part of the paper required clarification. As now better emphasized in the text, in this proof-of-concept work we considered only patients that actually transitioned to the SP phase, so there is a unique class of patients. Features input to the RF algorithm are listed in Supplementary Table 1. We added a statement to declare what types of numerical values we used in the work. The results presented show that the RF algorithm performs better than layman, as its performance is however better than that of individual medical students, that are not quite laymen, although not experts as well. In any case, the focus of the paper is not on the goodness of the algorithm, but on the value of combining different approaches to the prediction problem, which indeed has been resisting solution for many years of medical analysis.

Fusion rule: (Described in the section "To compare the two sets... of the most consistent agent.") I don't understand what you do. How does summing a squared ranking lead to a prediction score? I assume you are using the normalized (and squared) ranking as a sort of weight? Why do you square the rankings? What happens when you use other non-linear transformations? Is there any

way you illustrate the distributions so that we can follow your reasoning? How about presenting simple rules like averaging, or majority voting at least as a baseline we can compare against?
ANSWER
We agree with you that, indeed, this point is complex and we try a different explanation, hoping that it is clearer. First of all, ranking is inherent to building a ROC curve. Since we have only two agents (humans and RF algorithm), we cannot use a majority rule, we can only perform an average (linear or weighted) of the scores. For any clinical record, the final forecast is the average of "unitary predictions" by multiple individuals or decision trees. If "unitary predictions" of one agent are highly concordant, it means that the prediction is quite obvious for the agent, suggesting that is more probably correct than others. We ranked forecasts on clinical records in order of concordance of "unitary predictions" and emphasized the value of agreement by squaring the ranks.
In line with other pieces of research, this weighted average performed better than linear averaging, as stated in the paper

Competing Interests: No competing interest

Referee Report 21 March 2018

doi:10.5256/f1000research.14226.r31369

?

Roger Tam

Department of Radiology and MS/MRI Research Group, University of British Columbia, Vancouver, BC, Canada

This is an interesting and clearly written article on using a machine learning method (random forests) and medical students to form "hybrid" predictions of disease progression in MS, specifically the conversion from RRMS to SPMS. The article claims that the results are a proof-of-principle that combining machine learning and human predictions is better than either approach alone.

The main strengths of the article are its clear writing, the reproducbility of the experiments, the clinical importance of the application, and topical nature of the subject, as machine learning for clinical prediction is such a hot topic that integration with the clinical workflow is a critical area of study.

The main limitations of the article are that only clinical parameters were used to perform the predictions, and the longitudinal nature of the data was not used to its full benefit. To realize the potential of machine learning for MS prediction, imaging parameters should be included (there is good literature on MS prediction using imaging), and examining changes over time is important for both machine (eg, using recurrent networks) and human raters (examining clinical changes over multiple time points). The article places some importance on having the computer and humans using the same set of input parameters, but I do not feel that this is warranted; the data should be selected to be most appropriate for each approach.

Given the above limitations, it is difficult to generalize the findings to say that hybrid predictions are better than either machine learning or humans. This could be true, and the article provides some support for that, but more work needs to be done to provide strong evidence.

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 26 Jul 2018

Francesca Grassi, Sapienza University of Rome, Italy

First of all, thank you for taking the time to read our work and to give useful comments. We hope that our responses will clear your doubts. We list below the changes introduced in the new version, prompted by your observations. We hope that you agree with us that it is improved.

The main limitations of the article are that only clinical parameters were used to perform the predictions, and the longitudinal nature of the data was not used to its full benefit. To realize the potential of machine learning for MS prediction, imaging parameters should be included (there is good literature on MS prediction using imaging), and examining changes over time is important for both machine (eg, using recurrent networks) and human raters (examining clinical changes over multiple time points). The article places some importance on having the computer and humans using the same set of input parameters, but I do not feel that this is warranted; the data should be selected to be most appropriate for each approach.
ANSWER
We agree with you that many other approaches could be used. As now stated in the text, we chose to explore predictions done using only clinical data, available to all neurologists, which have recently been independently demonstrated to have good predictive value (Goodin et al., 2018; reference added to the paper). Imaging data performed in real-world clinical settings do not have the standardization required for predictions either by experts or algorithms. However, future studies aimed at confirming this proof-of-principle, initial work will surely consider different options.

Given the above limitations, it is difficult to generalize the findings to say that hybrid predictions are better than either machine learning or humans. This could be true, and the article provides some support for that, but more work needs to be done to provide strong evidence.

ANSWER
We completely agree with you that this is a preliminary, proof-of-concept work. We state it more clearly in the Discussion

Competing Interests: Nothing to disclose

Referee Report 26 February 2018

doi:10.5256/f1000research.14226.r30349

✓ **Bruno Bonetti**
USD Stroke Unit, DAI di Neuroscienze, Azienda Ospedaliera Universitaria Integrata, Verona, Italy

The manuscript is interesting and intriguing, since it opens new possibilities in MS prognosis combining human expertise and "artificial intelligence". I do not understand why medical students have been chosen instead of (young) neurologists who may have additional skills in the specific task. Apart from this aspect, the manuscript is well written and easy to follow. Deserves publication.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
I cannot comment. A qualified statistician is required.

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 01 Mar 2018
Francesca Grassi, Sapienza University of Rome, Italy

Thank you very much for reviewing our paper.

In this proof-of-concept study, we chose to work with medical students instead of neurologists because we wanted to test if even a group of relatively uneducated people can enhance the predictive ability of machine learning algorithms, which is now well established.

We agree with you that the next step is to obtain predictions by neurologists and other medical doctors, and in fact we set up the platform DiagnoShare (http://www.phys.uniroma1.it/diagnoshare) to extend the study.

Hopefully, we can soon extend this work with a final study

Competing Interests: None