Supervised machine learning in psychiatry

SUMMARY

As introduced in **Chapter 1**, mental disorders are widespread and burdensome, but most people suffering from them still miss to receive proper treatment or experience treatment unresponsiveness, relapses, and recurrence episodes. Psychiatry is still based on descriptive diagnostic taxonomies with limited validity, and clinical guidelines recommend interventions only for the 'average' patients suffering from a specific diagnostic class. The new paradigm of Personalized Medicine promises to improve the treatment and prevention of mental disorders by providing better individual indications and predictions. Recent technological advancements now permit to cost-effectively collect vast sets of information that may be used to achieve such personalized recommendations. However, data are not enough, and it is also necessary to develop models that can perform these personalized recommendations. Given the complexity and multifactorial nature of mental disorders, it is difficult to fully achieve a full *a priori* understanding of the phenomenon that is then used to develop rules that clinicians are advised to follow during their clinical decision-making process. A promising alternative to develop such tools is the use of Supervised Machine Learning (SML). These techniques use examples in which both the input and the desired output variables are available. From these examples, such techniques are able to automatically extract patterns and build algorithms that can provide an estimate of the output variables in new cases in which only the input variables are known. SML opens the possibility to develop PM tools that may have been impossible otherwise, only by having enough suitable examples and without the need for an explicit *a priori* understanding of the relationship between the input and output variables. Moreover, SML algorithms may help automatize some time-consuming clinical tasks, reducing the associate costs and clinicians' burden. In psychiatric scientific literature, a growing number of research articles using SML are available, and most of them present algorithms that seem to achieve very high performances. However, this amount of promising evidence appears to conflict with the general lack of SML-based tools in psychiatric clinical practice. Several challenges need to be faced to ensure a safe and effective application of these algorithms in clinical practice.

This doctoral dissertation aims to present the development and testing of some SML algorithms for the psychiatric clinical practice, with two main focuses: the ability to achieve good performance by solely using input information that facilitates or at least does not hinder its clinical adoption, and the necessity to provide preliminary evidence of the expected generalized performance of the algorithm even at early stages of its development.

In particular, **Chapters 2**, **3**, and **4** report three studies related to the development of an SML algorithm for the 3-year prediction of conversion from Mild Cognitive Impairment (MCI) to Alzheimer's Disease (AD). Currently available and emerging therapies for AD likely have the most significant impact when provided at the earliest disease stage. Thus, the possibility to early identify which subjects are at high risk of later developing AD, e.g., subjects with MCI, it is of crucial importance. However,

currently proposed machine learning algorithms seem to achieve only limited predictive accuracy, or they are based on expensive and hard-to-collect information.

The study presented in **Chapter 2** aimed to develop an initial proof-of-concept of a clinically-translatable SML algorithm for the 3-years prediction of conversion to AD in MCI and Pre-mild Cognitive Impairment (PreMCI) subjects. This algorithm is based only on non-invasive predictors that are either already routinely assessed or easily introducible in clinical practice. Specifically, baseline information regarding sociodemographic characteristics, clinical and neuropsychological test scores, cardiovascular risk indexes, and a visual rating scale for brain atrophy was used as input. Data were extracted from a longitudinal, multicentric dataset collected in Miami (Florida, US). A subset of 16 predictors was selected from all the abovementioned domains. The best model (support vector machine with radial-basis function kernel) resulted in a high leave-pair-out-cross-validation performance, with an Area Under the Receiver Operating Characteristics (AUROC) of 0.962, a balanced accuracy of 91.3%, a sensitivity of 95.6%, and a specificity of 87.1%. These results are among the best of the many algorithms available in the literature and the best achieved so far using only information easily collectible in clinical practice.
However, these preliminary results are based solely on a cross-validation approach, and not on a set of test examples that have completely been held out during the development of the algorithm. Thus, to provide a sounder estimate of its expected performance, the study in **Chapter 3** aimed to perform indirect testing of this algorithm via a transfer learning approach. The same predictors and SML technique used in the previous study were employed to retrain the algorithm to accomplish the task of discriminating between AD and Cognitively Normal individuals (CN). Data used for training were another sample of subjects with either the former or the latter condition that have being recruited during same longitudinal, multicentric dataset collected in Miami (Florida, US) used in the previous study. The new algorithm was then used to predict the three-year conversion to AD in the same sample of MCI subjects used in the previous study. In this study, the MCI sample was entirely held out during training and only used to test the algorithm. A reduced but still significant predictive performance was observed in the MCI sample (AUROC = 0.821; balanced accuracy = 77.9%; sensitivity = 85.2%; specificity = 70.6%), and these can be considered a first indirect, possibly conservative estimate of the performance of the algorithm presented in **Chapter 2** when applied to a sample of MCI subjects not used in the training process.

Chapter 4 presents an improved algorithm for the same task based on an ensemble of several SML algorithms whose individual predictions are aggregated with a weighted average rank approach. Data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) open database were employed in this study. A restricted set of information, which included sociodemographic and clinical characteristics and neuropsychological test scores, was used as predictors, while any imaging information was excluded entirely. Moreover, a peculiar site-independent stratified train/test split protocol was used to better estimate the generalized performance of the algorithm when applied in clinical centers different from those used for training and validation. The ensemble of the SML algorithms demonstrated a test AUROC of 0.88, a sensitivity of 77.7%, and a specificity of 79.9%. In addition, it demonstrated

a specificity of 40.2%/53% when the threshold was optimized to achieve a sensitivity of respectively 100% and 95%. These results show evidence of high predictive accuracy even when testing is performed with a sound train/test split protocol, exhibiting particularly good predictive performance when the algorithm was optimized as a screening tool. Thus, the algorithm may be useful to improve recruitment in clinical trials and to more selectively prescribe newly emerging early interventions to patients at high risk to convert from MCI to AD.

The work described in **Chapter 5** relates to the initial development of an SML algorithm for the prediction of 2-year OCD remission. The OCD course differs widely between OCD patients, varying from severe chronic symptoms to full remission. No tools for individual prediction of OCD remission are currently available. To facilitate clinical adoption, only predictors easily accessible in the daily clinical routine, such as anamnestic information and questionnaires, were used in the algorithm. Gradient boosted decision trees (GBDT) were used as a supervised machine learning technique. The training of the algorithm was performed with 227 features and a sample of 215 cases recruited in a single clinical center. The predictive performance of the algorithm was subsequently tested using an independent sample of 215 cases recruited in five different centers. All data were collected in a longitudinal multi-center study (NOCDA). The predictive performance of the algorithm in the five test centers resulted in an average AUROC of 0.7820, an average balanced accuracy of 72.73%, an average sensitivity of 73.42%, and an average specificity of 71.45%. However, a large between-center variation was observed (AUROC range = 0.636-0.906; balanced accuracy range = 58.0%-87.5%; sensitivity rage = 45.5%-100.0%; specificity range = 62.7%-76.9%), which evidence the challenge of achieving a stable generalized performance when applied into different clinical settings. These results highlight the necessity of testing SML algorithms in samples collected in different sites before being safely translated from the research environment into clinical practice.

Chapter 6 presents an extensive test of a computer program (MEBsleep by Medibio Limited) that performs automatic sleep staging of polysomnography by processing the signals of EEG montages through a processing pipeline of SML algorithms. Sleep staging of polysomnography is a time-consuming task, it requires significant training, and significant variability among scorers is expected. Testing has been based on the agreement of the staging performed by the program with the manual sleep staging performed by expert sleep technicians. The extensive test performed in this study aim to finally demonstrate the clinical applicability of the SML-based software. Forty polysomnography recordings of patients referred for sleep evaluation to three different sleep clinics were retrospectively collected. Three experienced technicians independently staged the recording twice, first taking into account only the electroencephalography signals, and then also the electromyography and electrooculography signals in compliance with the staging rules recommended by the American Academy of Sleep Medicine guidelines. In addition, the staging performed initially in clinical practice was also considered. Several agreement statistics of the automatic staging with the manual staging, among the different manual staging scoring, and their differences were calculated as a test of the performance achieved by the SML-based application. The automatic staging resulted for the most part comparable or significantly more in agreement with the technicians' staging than the

between-technician agreement, with the sole exception of a partial reduction in the positive percent agreement of the Wake stage. The same result was observed in the comparison between the agreement of the automatic staging with the clinical staging and the agreement of the technicians' staging with the clinical staging. Given these results, the use of this SML-based software may be granted as a supporting tool for sleep clinicians, helping to reduce the burden associated with manual sleep staging of inpatient polysomnography.

In conclusion, as discussed in **Chapter 7,** the studies presented in this doctoral thesis used SML techniques to develop and test algorithms that may support mental health professionals in their clinical practice. Not all studies included in this dissertation present an algorithm that has already reached a definitive, clinically applicable version. However, even at early or intermediate steps, the studies followed specific strategies to provide a preliminary estimate of the generalized performance of the algorithm in order to identify issues in their current versions promptly and better direct the following development steps. The results support the hypothesis that it is possible to develop algorithms that can achieve a clinically meaningful performance using only input information that would allow an easy clinical translation of these algorithms and avoiding using information that is currently expensive, invasive, or hard to introduce into clinical psychiatric practice. In addition, the studies evidence the crucial role of data, which needs to be of good quality and big-enough quantity to develop SML algorithms. As the use of either purely clinical or experimental data may result in specific issues, data collected in observational, multicentric, non-retrospective studies seem to be the most suitable for SML, and increased availability of open-source datasets with these characteristics may foster the development and test of SML-based clinical tools. Moreover, given that any SML algorithm needs to be transformed into a proper clinical tool before it can be translated into clinical practice, and that the performance that such tool can achieve in practice may be different than what observed for the sole SML algorithm, particular attention should be given also to this development phase, as well as to improve the clinicians' understanding of SML and to remove any resistance from clinicians associated with the utilization of SML-based tools in clinical practice. Only initiatives that promote a coordinated effort between several professional roles and stakeholders, including the end-users of such tools that are clinicians and patients, can finally make SML-based Personalized Medicine clinical tools a widespread reality.