Route map for machine learning in psychiatry: absence of bias, reproducibility, and utility

Joaquim Radua[1,2,3,*] and Andre F. Carvalho[4]

[1] Imaging of Mood- and Anxiety-Related Disorders (IMARD) group, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), CIBERSAM, Barcelona, Spain

[2] Early Psychosis: Interventions and Clinical-detection (EPIC) lab, Institute of Psychiatry, Psychology, and Neuroscience, King's College London, London, UK

[3] Department of Clinical Neuroscience, Stockholm Health Care Services, Stockholm County Council, Karolinska Institutet, Stockholm, Sweden.

[4] IMPACT (Innovation in Mental and Physical Health and Clinical Treatment) Strategic Research Centre, School of Medicine, Barwon Health, Deakin University, Geelong, VIC, Australia.

[*]Corresponding Author

Joaquim Radua, MD PhD

Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain.

Email: radua@clinic.cat

Total word count: 1013

Number of references: 11

Short title: Machine learning route map

Acknowledgments

JR is supported by a grant from the Instituto de Salud Carlos III and co-funded by European Union (ERDF/ESF, "Investing in your future"): Miguel Servet Research Contract CPII19/00009.

TEXT

In the past decade, several groups reported incredible achievements using machine learning. For instance, Google reported a neural network that taught itself how to identify cats (Markoff, 2012). Or Facebook presented another network that recognized individuals in photographs with >97% accuracy (Taigman et al., 2014). Such successes led to considerable interest in the application of machine learning techniques to many disciplines. Psychiatry was not an exception, and we embraced this perspective enthusiastically.

There were reasons to be optimistic. For example, years ago, we used obscure stepwise regressions to find a model to predict treatment response from several baseline variables. We knew that stepwise regression led to inflated statistical significance (Mundry and Nunn, 2009), but we had fewer alternatives. Today, we have safe machine learning classifiers such as regularized regressions (e.g., lasso), random forests, or support vector machines (Salvador et al., 2017). Ultimately, these tools have the potential to predict the therapeutic response at an individual level.

However, we think that we should not give machine learning a blank cheque. The enthusiasm may have made us lower the guard in methodological rigor and preclinical/clinical utility. As we expose in the following, we believe that several hurdles may lead the community to think that machine learning is only about unbelievable predictions or useless studies. We also propose a route map to avoid these hurdles, guiding future studies so that machine learning becomes a reliable and valuable tool in psychiatry.

Absence of bias

The first hurdle refers to a permissive methodology that may lead to systematic biases. For instance, everyone involved in magnetic resonance imaging research knows that when you have data from different sites, you must very carefully control the effects of the site (Radua et al., 2020). However, in novel machine learning applications, analysts usually estimate the accuracy of the prediction model without considering these effects. Unfortunately, ignoring them may yield severely inflated accuracy. In other words, machine learning models may seem to predict very well when they do not even predict (Solanes et al., 2021).

We propose ensuring that machine learning studies meet the same methodological rigor as any other study. We know that in machine learning, any algorithm is possible. We are open-minded: we may accept that your algorithm includes astrology and tarot readings to conduct the predictions. But when it comes to estimating the accuracy of a machine learning model's predictions, the analysis must be as rigorous as in any other study. We must control confounders and any other source of bias with the same rigor as we do in standard statistics. For example, suppose my patient sample is composed mainly of men and my control sample of women. A machine-learning algorithm could seem to predict whether an individual is a patient with great accuracy exclusively based on his/her gender. However, this accuracy would be biased. When

estimating the accuracy, we should control the confounding effects of gender, either evaluating the accuracy separately for men or women or adding gender as a covariate.

Reproducibility

The second hurdle refers to data torturing and publication bias, which may make the experiments hardly reproducible. Before machine learning, we quickly suspected data torturing when a researcher compared patients and controls with a battery of statistical tests until the differences were "statistically significant." Conversely, people do not seem to worry about this threat in machine learning. Software like MATLAB allows the user to perform automated training to search for the best classification model type, including decision trees, discriminant analysis, support vector machines, and several other algorithms. It is not uncommon to try many machine learning algorithms until one seems to "predict" (Hosseini et al., 2020). We acknowledge that such practices may have justification on some occasions. Still, they may very easily lead to data torturing.

We propose ensuring that machine learning models are reproducible. Many voices have alerted us of the low reproducibility of scientific research (Fusar-Poli et al., 2014), and we believe this problem may be especially severe in machine learning. For instance, in a recent evaluation of regression models using clinical/neuropsychological data to predict the transition to psychosis in individuals at clinical high-risk, previously published models either failed to predict or only showed poor to fair accuracy (Rosen et al., 2021).

Therefore, while we welcome new studies reporting novel algorithms, we strongly urge funding bodies and journals to encourage the conduction of independent replication studies, which unbiasedly assess the accuracy of previously published machine-learning algorithms. We also suggest that machine learning studies in psychiatry should readily adhere to existing reporting guidelines for the same purpose.

Utility

The last hurdle refers to the preclinical/clinical utility of machine learning studies. Everyone would agree that statistical analyses are only a means to answer a relevant, unknown question. E.g., what are the brain abnormalities in patients with a disorder? Or, what is the response to a given treatment? The utility of these questions contrasts with the utility of machine learning publications about models that estimate whether a brain MRI is from a patient or healthy control. We fully acknowledge the value of these pioneering studies as "proofs of concept." However, a model that only predicts whether a brain MRI is from a patient or healthy control may have a dubious utility: the clinician already has this information. Pablo Picasso is believed to say once, "Computers are useless. They can only give you answers". In this line, machine learning may become useless if we, humans, fail to pose the right questions.

Therefore, we propose that machine learning studies have relevant preclinical/clinical objectives. E.g., to discover biological endophenotypes or to predict treatment response. We acknowledged earlier the value of studies testing whether an algorithm may be applicable, but we

suggest focusing on answering more relevant preclinical/clinical questions. In this regard, it may be necessary to increase the number of cohort and follow-up studies, as most clinical questions refer to the "future" (e.g., the prognosis or the response to therapy). Several loadable novel studies have already begun this path (Amoretti et al., 2021; Filippi et al., 2021).

With these points, we believe that machine learning may more likely become an inseparable ally of research in psychiatry.

REFERENCES

Amoretti, S., Verdolini, N., Mezquida, G., Rabelo-da-Ponte, F.D., Cuesta, M.J., Pina-Camacho, L., Gomez-Ramiro, M., De-la-Camara, C., Gonzalez-Pinto, A., Diaz-Caneja, C.M., Corripio, I., Vieta, E., de la Serna, E., Mane, A., Sole, B., Carvalho, A.F., Serra, M., Bernardo, M., 2021. Identifying clinical clusters with distinct trajectories in first-episode psychosis through an unsupervised machine learning technique. European neuropsychopharmacology : the journal of the European College of Neuropsychopharmacology.

Filippi, I., Galinowski, A., Lemaitre, H., Massot, C., Zille, P., Frere, P., Miranda-Marcos, R., Trichard, C., Guldner, S., Vulser, H., Paillere-Martinot, M.L., Quinlan, E.B., Desrivieres, S., Gowland, P., Bokde, A., Garavan, H., Heinz, A., Walter, H., Daedelow, L., Buchel, C., Bromberg, U., Conrod, P.J., Flor, H., Banaschewski, T., Nees, F., Heintz, S., Smolka, M., Vetter, N.C., Papadopoulos-Orfanos, D., Whelan, R., Poustka, L., Paus, T., Schumann, G., Artiges, E., Martinot, J.L., Consortium, I., 2021. Neuroimaging evidence for structural correlates in adolescents resilient to polysubstance use: A five-year follow-up study. European neuropsychopharmacology : the journal of the European College of Neuropsychopharmacology 49, 11-22.

Fusar-Poli, P., Radua, J., Frascarelli, M., Mechelli, A., Borgwardt, S., Di Fabio, F., Biondi, M., Ioannidis, J.P., David, S.P., 2014. Evidence of reporting biases in voxel-based morphometry (VBM) studies of psychiatric and neurological disorders. Hum Brain Mapp 35, 3052-3065.

Hosseini, M., Powell, M., Collins, J., Callahan-Flintoft, C., Jones, W., Bowman, H., Wyble, B., 2020. I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data. Neuroscience and biobehavioral reviews 119, 456-467.

Markoff, J., 2012. How Many Computers to Identify a Cat? 16,000, The New York Times.

Mundry, R., Nunn, C.L., 2009. Stepwise model fitting and statistical inference: turning noise into signal pollution. The American naturalist 173, 119-123.

Radua, J., Vieta, E., Shinohara, R., Kochunov, P., Quide, Y., Green, M.J., Weickert, C.S., Weickert, T., Bruggemann, J., Kircher, T., Nenadic, I., Cairns, M.J., Seal, M., Schall, U., Henskens, F., Fullerton, J.M., Mowry, B., Pantelis, C., Lenroot, R., Cropley, V., Loughland, C., Scott, R., Wolf, D., Satterthwaite, T.D., Tan, Y., Sim, K., Piras, F., Spalletta, G., Banaj, N., Pomarol-Clotet, E., Solanes, A., Albajes-Eizagirre, A., Canales-Rodriguez, E.J., Sarro, S., Di Giorgio, A., Bertolino, A., Stablein, M., Oertel, V., Knochel, C., Borgwardt, S., du Plessis, S., Yun, J.Y., Kwon, J.S., Dannlowski, U., Hahn, T., Grotegerd, D., Alloza, C., Arango, C., Janssen, J., Diaz-Caneja, C., Jiang, W., Calhoun, V., Ehrlich, S., Yang, K., Cascella, N.G., Takayanagi, Y., Sawa, A., Tomyshev, A., Lebedeva, I., Kaleda, V., Kirschner, M., Hoschl, C., Tomecek, D., Skoch, A., van Amelsvoort, T., Bakker, G., James, A., Preda, A., Weideman, A., Stein, D.J., Howells, F., Uhlmann, A., Temmingh, H., Lopez-Jaramillo, C., Diaz-Zuluaga, A., Fortea, L., Martinez-Heras, E., Solana, E., Llufriu, S., Jahanshad, N., Thompson, P., Turner, J., van Erp, T., collaborators, E.C., 2020. Increased power by harmonizing structural MRI site differences with the ComBat batch adjustment method in ENIGMA. Neuroimage 218, 116956.

Rosen, M., Betz, L.T., Schultze-Lutter, F., Chisholm, K., Haidl, T.K., Kambeitz-Ilankovic, L., Bertolino, A., Borgwardt, S., Brambilla, P., Lencer, R., Meisenzah, E., Ruhrmann, S., Salokangas, R.K.R., Upthegrove, R., Wood, S.J., Koutsouleris, N., Kambeitz, J., 2021. Towards Clinical Application of Prediction Models for Transition to Psychosis: A Systematic Review and External Validation Study in the PRONIA Sample. Neuroscience and biobehavioral reviews.

Salvador, R., Radua, J., Canales-Rodriguez, E.J., Solanes, A., Sarro, S., Goikolea, J.M., Valiente, A., Monte, G.C., Natividad, M.D.C., Guerrero-Pedraza, A., Moro, N., Fernandez-Corcuera, P., Amann, B.L., Maristany, T., Vieta, E., McKenna, P.J., Pomarol-Clotet, E., 2017. Evaluation of machine learning algorithms and structural features for optimal MRI-based diagnostic prediction in psychosis. PloS one 12, e0175683.

Solanes, A., Palau, P., Fortea, L., Salvador, R., Gonzalez-Navarro, L., Llach, C.D., Valenti, M., Vieta, E., Radua, J., 2021. Inflated accuracy in multisite machine-learning studies due to incomplete removal of the effects of the site. Submitted.
Taigman, Y., Yang, M., Ranzato, M.A., Wolf, L., 2014. DeepFace: Closing the Gap to Human-Level Performance in Face Verification, 2014 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Columbus, OH, USA.