Data Science Challenges in Computational Psychiatry and Psychiatric Research

Daniel Stahl
Department of Biostatistics and Health Informatics,
Institute of Psychiatry, Psychology and Neuroscience
King's College London
London, UK
daniel.r.stahl@kcl.ac.uk

Daniel Stamate
Data Science & Soft Computing Lab, and
Department of Computing
Goldsmiths, University of London
London, UK
d.stamate@gold.ac.uk

***Abstract*— The special session "Data Science in Computational Psychiatry and Psychiatric Research" at the 5th IEEE International Conference on Data Science and Advanced Analytics in Turin, Italy, 2018, presents papers specifically addressing psychiatric research and computational approaches in psychiatry. In this overview, we describe the challenges of psychiatric research and computational psychiatry and demonstrate how the presented papers approach some of the problems of interest.**

Keywords—Data science, Psychiatry, Precision medicine, Prediction modelling, Machine learning, Big data

Psychiatric research entered the age of big data with patient databases now available with thousands of clinical, demographical, social, environmental, neuroimaging, genomic, proteomic and other -omic measures. The analysis of big data is essential for precision medicine, "an emerging approach for treatment and prevention that takes into account each person's variability in genes, environment, and lifestyle" [1]. Examples of precision medicine include using targeted therapies to specific genetic changes which cause individual cancer to grow and progress, such as the HER2+ gene cells. However, unlike in many other medical and clinical areas, psychiatry has not yet benefited from the possibilities of big data and modern computer and data science technologies to improve diagnostic, monitoring, prognostic and predictive technologies which underpin precision medicine [2].

Evidence-based psychiatric medicine still focuses on randomized controlled trials to establish the best treatment for the average patient. Clinical trial provides the highest level of evidence by preventing selection bias through random allocation of patients in either treatment or control group. However, this approach ignores statistical heterogeneity and a new treatment is applied to all patients. Precision medicine, on the other hand, accepts that reality is not homogenous and treatments are tailored to the individual. Precision medicine is made possible by the provision of "new" data (such as Imaging, Omics, patient records, sensors in smartphones, wearables and internet and smartphone usage) and high-performance computing technology that effectively collects, processes, stores and analyses "big data". It is becoming increasingly synonymous with "analytical approach of prediction modelling" [3].

Data-driven approaches apply machine and statistical learning methods to high-dimensional data to improve classification of disease, to predict the likelihood of development of a disease, make a prognosis about likely course of a clinical condition, predict treatment outcomes or improve treatment selection. The use of big data is usually discussed in the context of improving medical care but preventing disease became increasingly important in recent years [4].

Psychiatry has not yet benefited from the availability of big data and clinical prediction modelling application [2, 5]. There are only a few clinical prediction models available and even these need to be considered preliminary. Why is this the case? The analysis of such data is often more challenging than in other medical research areas because:

i. Psychiatrists study traits which are not easily measurable; they need to be measured indirectly e.g. by questionnaires (Measurement error).

ii. Although mental disorders typically show a strong heritability, genetic variants for most traits account for far less than 1% of the variability.

iii. Susceptibility genes are often common variants rather than mutations which does not allow to target specific genetic changes as in cancer.

iv. The definition of a mental disease is often very broad and often includes distinct but unknown subcategories, different mental health problems have similar phenotypes and many mental (and physical) health problems occur together (comorbidity).

v. Patient recruitment is difficult and there is a high proportion of drop-out in many studies and patients often do not adhere to the treatment.

vi. Studies using non-random samples with restrictive inclusion criteria and self-selection bias or case-healthy control studies are often used in psychiatric research. Models with good internal validity typically show poor external validity and do not generalize to the clinical population.

vii. Treatment interventions often have several interacting components and it is often difficult to measure the active ingredients of an intervention (complex interventions).

viii. Important data such as Electronic Health records of mental health patients are typically narrative text or other unstructured data types (which would require non-conventional data analysis technologies such as NLP to take full advantage of such data).

A mental health disorder, like depression, results from a range of different biological and environmental causes, and the same cause may lead to different disorders in different people [6]. Concentrating on a gene-centred and symptom-based classification approach was, therefore, not successful

Daniel Stahl was partly funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. Daniel Stamate was partly funded by Alzheimer's Research UK (ARUK) - Predicting risk of dementia using routine primary care records grant, and Goldsmiths, University of London. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, ARUK, or the Department of Health.

and there is a shift towards machine learning and big data analysis which aim to identify data-driven types that are coherent across symptoms, brain, genes and behaviour, and relevant to a clinical outcome such as the risk of relapse or response to treatment [7]. This requires a multimodal approach that combines cognitive and clinical measures, demographics, environmental, social, genetics, brain imaging, omics, social media use, real-time physiological data or real-time symptoms survey data. The development of pipelines to acquire, process and merge data from multiple sources into a database in ways that allows prompt and reliable exploration and analyses of data and subsequent recommendations is a challenge, which can be only achieved by the joint effort of computer scientists, clinicians, statisticians, patients, carers and other stakeholders.

Computational Psychiatry will play the leading role in the development and application of clinical prediction models in mental health. However, data-driven approaches are typically using "black box" machine learning methods, which do not help to understand the development of a mental illness in the brain [6,8]. Computational psychiatry also needs to apply theory-driven approaches that attempt to model mental processes using abstract algorithms. This will provide a better understanding between psychiatric symptoms and the underlying neurobiological processes and may provide in the future new ways of developing treatments or prevention programmes. Computational psychiatry has the potential to change psychiatry and move it towards a truly personalized medicine approach, and our special session should be part of this exciting movement.

The aim of our special session is to promote researchers from both academia and industry to participate in this workshop to present, discuss, and share the latest findings in the field, and exchange ideas that address real-world problems with real-world solutions, as well as to discuss future research directions and applications, and to identify a set of recommendations for future research activities.

In this special session, papers address a variety of the described problems. Data analytics is not only concerned about prediction but also to gain an understanding about the mechanisms that cause structural and covariation in data. Keynote speaker Fionn Murtagh describes how correspondence analyses can be used to contextualized activity data for general health, depression and demographics. Free text is a common problem in mental health research and Yuelin Lin and Thomas Atkinson introduce a method to summarize such qualitative data automatically using latent Dirichlet allocation. Personalized medicine aims to assign the best treatment to a patient. Xuan Zhou, Yuanjia Wang and Donglin Zeng propose a novel machine learning technique to generalize outcome-weighted learning for binary treatment to multi-treatment settings via sequential weighted support vector machines. High-dimensional datasets with multiple categorical variables and missing data are a common challenge in psychiatric survey data sets. Constantin Ahlmann-Eltze and Christopher Yau propose a new scalable Bayesian clustering algorithm to identify low-dimensional structures embedded within the high-dimensional dataset of categorical data. Comorbidities are a challenge in psychiatric research. Maxim Sharaev and his colleagues describe the development of an efficient pipeline to investigate neuroimaging data as candidate biomarkers for depression, and depression and epilepsy comorbidity.

In conclusion, our special session provides an exciting overview of new data science applications in psychiatric research. It provides a vision of the near future for the field of "precision psychiatry" with the ultimate goal to provide better lives for those suffering from mental illness [3].

ACKNOWLEDGMENT

We thank our Programme Committee members Danielle Belgrave, Erik J. Linstead, Yuelin Li, Cedric Ginestat, Matthias Pierce, Fionn Murtagh, Sinan Guloksuz, Evan Kontopantelis, David Reeves, Taposhri Ganguly, Alexander Zamyatin and Raquel Iniesta for their careful reviews of the submitted manuscripts.

REFERENCES

[1] Genetics Home Reference (2018) "What Is Precision Medicine?" Genetics Home Reference. Accessed May 8, 2018. https://ghr.nlm.nih.gov/primer/precisionmedicine/ definition.

[2] Fernandes, B. S., Williams, L. M., Steiner, J., Leboyer, M., Carvalho, A. F., & Berk, M. (2017). The new field of "precision psychiatry". BMC Medicine, 15, 80.

[3] Checkroud, A.M. and Koutsouleris, N. (2018) The perilous path from publication to practice. Molecular Psychiatry, 23, 24–25.

[4] Barrett, M.A., Humblet, O, Hiatt, R.A., and Adler, N.E. (2013) Big Data and Disease Prevention: From Quantified Self to Quantified Communities. Big Data, 1(3).

[5] Stewart, R., & Davis, K. (2016). "Big data" in mental health research: current status and emerging possibilities. Social Psychiatry and Psychiatric Epidemiology, 51, 1055–1072.

[6] Huy, Q.M. , Maia, T. V. and Frank, M. J. (2016) Computational psychiatry as a bridge from neuroscience to clinical applications. Nature Neuroscience volume19, 404–413.

[7] Marquand, A. F.,Wolfers, T., Mennes, M. Buitelaar, J, and Beckmann,C.F. (2016) Beyond Lumping and Splitting: A Review of Computational Approaches for Stratifying Psychiatric Disorders. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 1(5),433-447.

[8] Makin, S. (2016) Can Big Data Help Psychiatry Unravel the Complexity of Mental Illness? Scientific American. Retrieved from https://www.scientificamerican.com/article/can-big-data-help-psychiatry-unravel-the-complexity-of-mental-illness/?WT.mc_id=send-to-friend