BJPsych | The British Journal of Psychiatry (2018) 213, 509–510. doi: 10.1192/bjp.2018.105

Editorial

Rise of the machines? Machine learning approaches and mental health: opportunities and challenges

Paul A. Tiffin and Lewis W. Paton

Summary

Machine learning methods are being increasingly applied to physical healthcare. In this article we describe some of the potential benefits, challenges and limitations of this approach in a mental health context. We provide a number of examples where machine learning could add value beyond conventional statistical modelling.

Declaration of interest

None.

Keywords

Machine learning; mental health services; personlised medicine; diagnosis; prediction.

Copyright and usage

Paul A. Tiffin (pictured) is a Reader (Associate Professor) in Psychometric Epidemiology at the University of York and a National Institute for Health Research Career Development Fellow. His main research interest is the development and application of statistical and psychometric modelling approaches. **Lewis W. Paton** is a Research Fellow (Assistant Professor) at the University of York. He is a statistician interested in developing novel methodological approaches.

Background

Stephen Hawking is quoted as stating that 'the development of full artificial intelligence could spell the end of the human race' and Elon Musk that artificial intelligence is a 'fundamental existential risk for human civilisation'.[1,2] However, if such risks can be averted it is clear that there is a huge potential for artificial intelligence to enhance the lives of humanity. Machine learning is the basis of artificial intelligence and there has been a growing interest in the application of the approach to address diagnostic and prediction issues in healthcare. Machine learning occurs when a system is able to learn from new information in order to complete a specified task. Learning in this sense can be 'supervised' or 'unsupervised'. In supervised learning the machine is shown a series of examples so it learns how to link predictors ('features') to a target outcome, for example converting handwritten characters to postcodes. In unsupervised tasks the system learns how to cluster or categorise variables in the absence of training data.

In contrast to conventional statistical modelling, which usually facilitates explanation, machine learning tends to be focused solely on prediction. Thus, in medicine, the technique has been used primarily to generate automated approaches to both diagnosis and treatment. There have been a number of well-publicised 'proof of concepts' in physical healthcare, such as those provided by DeepMind,[3] formerly part of Google. DeepMind have been successful in generating machine learning algorithms that can predict impending mortality in cardiac patients and renal failure.

Unlike physical health, machine learning approaches to mental health have mainly been applied in research contexts rather than being implemented in real-world settings. For example, machine learning algorithms have been evaluated in relation to the identification of patients affected by obsessive–compulsive disorder via functional brain images,[4] the detection of markers of depression using Instagram photos[5] and, combined with automated speech analysis, the prediction of psychosis onset in clinically at-risk individuals.[6] Artificial intelligence is also being used to power 'Tess', a mental health support messaging service.[7]

The strengths and limitations of machine learning in mental health

Differing machine learning approaches vary in relation to their mathematical underpinnings. Nevertheless, they often share the same relative strengths and limitations of conventional statistical methods. However, machine learning is firmly focused on optimising performance on a prediction task. As such, learning algorithms will pursue this task with all the relentlessness of a T-101 terminator pursuing Sarah Connor through a Los Angeles police station. Unlike traditional approaches to statistics, where linear (or log-linear) relationships between predictor variables and an outcome are often assumed, machine learning takes a more flexible approach. As such linear or curvilinear relationships between the input variables (features) can be postulated, sometimes in numerous permutations, in order to maximise the predictive performance of an algorithm. Moreover, 'ensemble' techniques use many (sometimes thousands) of different models simultaneously in order to link features and outcomes most effectively under differing conditions. Thus, machine learning is able to harness the 'brute force' of modern computing power to bear on prediction tasks.

Inevitably, these approaches bring with them the risk of 'overfitting' – that is, deriving a well-fitting model to the training data that has little or no predictive validity in a separate data-set. This is generally a consequence of the model being fitted to noise rather than the signal of interest. Machine learning experts have thus spent considerable time and effort in developing measures that counter this risk.

Machine learning models are sometimes described as 'black boxes' in that they are usually not interpretable. This means little is learned about the relationship between the predictors and the outcome of interest. However, this also means that models cannot be easily accommodated to changing trends. For example, models based on 'web scraping' approaches that harvest images or text from the internet in order to identify individuals who may be at

https://doi.org/10.1192/bjp.2018.105 Published online by Cambridge University Press

risk of depression or suicide may be invalidated by changing trends in the use of language, or even camera settings. This might not be immediately apparent to end-users of the algorithms.

Potential sources of data

It is clear that machine learning has the potential to address important diagnostic and clinical decision-making issues in mental health. However, in order to realise this routinely available data of acceptable quality must exist – machine learning is no 'magic bullet' for 'dirty data'. Three potential sources of such data are worth mentioning.

First, the implementation of electronic records in health services provides both coded and free text that can be exploited by machine learning. An example of this has been the CRIS and D-CRIS systems,[8] which have enabled a variety of research projects, including the use of natural language processing in order to describe symptom profiles in severe mental illness.[9] The possibility of obtaining automated diagnoses that are at least as accurate as those given by experienced clinicians could offer cost-savings as resources become ever more straitened. Nevertheless, there will inevitably be human factors and public and professional perceptions of technology that may affect its implementation. Certainly, the experience to date in the use of mathematically driven clinician-decision support tools is that they are frequently overridden by practitioners, who may feel that they have more to lose than the machine if the diagnosis is wrong, even by some very small chance.[10] However, if automated systems effectively communicate their own degree of uncertainty in relation to a prediction then practitioners can take this into account when making their final clinical decision. Thus, in practice, as with the use of autopilots in aeroplanes, machine learning is likely to complement rather than replace human clinicians in the near future.

The roll-out of Improving Access to Psychological Therapies services in England has also led to large, routinely arising datasets that include repeated brief symptom measures. It is possible that, combined with other social demographic background information as well as symptom profile data, patients who are unlikely to respond to brief interventions could be identified early on and more intense therapies offered. This is in keeping with the 'stratified/personalised medicine movement' where management is tailored to individuals or subgroups of patients. Moreover, such psychosocial data could be combined with biological markers (including genetic or immunological profile) or neuroimaging to improve the predictive accuracy of models.

Furthermore, there is the possibility that previous existing data from large-scale trials could be reanalysed in order to understand which patients are most likely to benefit from certain interventions, as well as answering questions related to prognosis. Ethically, it would be important to provide alternative treatments for those identified as being highly likely to benefit from a particular therapeutic approach. Moreover, in practice, the 'negative predictive value' of algorithms will not be 100%. Consequently, it may be that such automated predictions would be best employed to guide early cessation of a treatment approach that was not appearing to be effective, rather than prevent access to such a therapy altogether.

Another important source of routinely arising data is from the 'digital footprint' left by internet-based activity. Machine learning can detect when an individual's online footprint, generated from apps, social media and other personal monitoring devices may indicate an impending risk of illness or self-harm or even violence.[11] The advent of affordable wearable technology, combined with machine learning techniques, could lead to an era of automated risk monitoring, although ethical considerations must be taken into account.

In theory, there could come a time when machine learning approaches can deliver aspects of mental healthcare remotely and independently of clinical input. Such an achievement has the potential to increase the reach of mental healthcare, particularly to those individuals reticent to access mental health services currently. There have been previous attempts at 'remote care', such as computerised cognitive–behavioural therapy and telemedicine. Such approaches have had some success in delivering care remotely, hinting at the potential for machine learning approaches. Nevertheless, machine learning approaches must learn the lessons of other remote care attempts, such as higher drop-out rates compared with treatment as usual.[12]

To conclude, machine learning should not be considered a magic solution to predictive problems in mental healthcare, but as one tool in the quantitative box of approaches that may be useful in both research and, eventually, practice. Like a novel class of psychotropic drug, machine learning will undoubtedly eventually find its correct place within mental health services.

Paul A. Tiffin, Reader (Associate Professor) in Psychometric Epidemiology, Department of Health Sciences, University of York, UK; **Lewis W. Paton**, Research Fellow (Assistant Professor), Department of Health Sciences, University of York, UK

Correspondence: Paul A. Tiffin, Department of Health Sciences, Seebohm Rowntree Building, University of York, Heslington, YO10 5DD. Email: paul.tiffin@york.ac.uk

First received 23 Feb 2018, final revision 29 Apr 2018, accepted 11 May 2018

Funding

PAT is supported in his research by an NIHR-funded Career Development Fellowship. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

References

1 Cellan-Jones R. Stephen Hawking warns artificial intelligence could end mankind. *BBC News* 2014; 2 December (http://www.bbc.co.uk/news/technology-30290540).

2 Sulleyman A. Elon Musk: AI is a 'fundamental existential risk for human civilisation' and creators must slow down. *Independent* 2017; 17 July (https://www.independent.co.uk/life-style/gadgets-and-tech/news/elon-musk-ai-human-civilisation-existential-risk-artificial-intelligence-creator-slow-down-tesla-a7845491.html)

3 DeepMind. *Solve Intelligence. Use it to Make the World a Better Place*. DeepMind, no date (https://deepmind.com/).

4 Takagi Y, Sakai Y, Lisi G, Yahata N, Abe Y, Nishida S, et al. A neural marker of obsessive-compulsive disorder from whole-brain functional connectivity. *Sci Rep* 2017; **7**: 7538.

5 Reece AG, Danforth CM. Instagram photos reveal predictive markers of depression. *EPJ Data Sci* 2017; **6**: 15.

6 Bedi G, Carrillo F, Cecchi GA, Slezak DF, Sigman M, Mota NB, et al. Automated analysis of free speech predicts psychosis onset in high-risk youths. *NPJ Schizophr* 2015; **1**: 15030.

7 X2AI. *Affordable, On-Demand, and Quality Mental Healthcare for Everyone Using Psychological Artificial Intelligence*. X2AI, no date (https://x2.ai/assets/X2AI_Executive_Summary.pdf).

8 South London and Maudsley NHS Foundation Trust. *CRIS*. South London and Maudsley NHS Foundation Trust, no date (https://www.maudsleybrc.nihr.ac.uk/facilities/clinical-record-interactive-search-cris/).

9 Jackson RG, Patel R, Jayatilleke N, Kolliakou A, Ball M, Gorrell G, et al. Natural language processing to extract symptoms of severe mental illness from clinical text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project. *BMJ Open* 2017; **7**: e012012.

10 Roshanov PS, Fernandes N, Wilczynski JM, Hemens BJ, You JJ, Handler SM, et al. Features of effective computerised clinical decision support systems: meta-regression of 162 randomised trials. *BMJ* 2013; **346**: f657.

11 Walsh CG, Ribeiro JD, Franklin JC. Predicting risk of suicide attempts over time through machine learning. *Clin Psychol Sci* 2017; **5**: 457–69.

12 Waller R, Gilbody S. Barriers to the uptake of computerized cognitive behavioural therapy: a systematic review of the quantitative and qualitative evidence. *Psychol Med* 2009; **39**: 705–12.

https://doi.org/10.1192/bjp.2018.105 Published online by Cambridge University Press