BJPsych | The British Journal of Psychiatry (2022)
220, 307–308. doi: 10.1192/bjp.2022.45

Kaleidoscope

Derek K. Tracy, Dan W. Joyce,
Dawn N. Albertson, Sukhwinder S. Shergill

Randomised controlled trials (RCTs): the top of the research pyramid, but we need to remain mindful of their limitations. Taipale et al[1] considered studies on individuals with schizophrenia, taking typical RCT eligibility criteria in relapse prevention work and applying them to broad, clinically representative Scandinavian cohorts covering over 25 000 individuals with a history of the condition. About 80% would have been ineligible for standard research trials; the most common reasons were serious somatic comorbidities, the use of antidepressants and mood stabilisers, substance use and perceived risk of suicide. This raises the obvious concern about how well this gold standard of research maps on to the broader real-life population. The authors looked at some high-level outcome variables and found that the 'ineligible' cohort had worse outcomes, being more likely to be admitted to hospital despite being on maintenance treatment, have a refractory condition and have greater number of suicide attempts. The 'C' in RCT is of course the culprit, as researchers understandably try to minimise and control for confounders, but if it means only a fifth of an illness cohort can be studied, and their outcomes differ from those of the rest, well, some new ideas are needed. We need more work on more real-world populations: as the authors note, 'RCT outcomes (efficacy) may differ from the utility of interventions in routine clinical practice (effectiveness)'. Mining of electronic patient records (EPR) has been proposed as one way around this, but see later in the column for potential pushback against that.

In depression, trials are often predicated on short intervention periods in individuals with more benign illness histories. Rush et al[2] talk through the challenges in the recently proposed heuristic 'difficult-to-treat depression' (DTD). First, what's in a name? The more common concept of treatment-resistant depression (TRD), while 'simple' in terms of easily defining failure to improve, misses much nuance: response or remission, the 'adequacy' of a trial duration, how to measure inability to tolerate a medication, and timing issues such as improvement that lapses or the question of which 'past failures' count. This heterogeneity challenges the face validity of TRD, or certainly that of extrapolating findings from one such group to another. DTD sidesteps this and, the authors propose, may better facilitate timely identification of personalised interventions. Back to methodology, and three fundamental challenges are now held up for future work to tackle: participant selection, outcome assessment and study design. DTD subpopulations need delineating, whether through clinical features, perpetuating factors or temporal evolution of illness. It has been noted that depression lacks a key prognostic marker, such as HbA1C in diabetes, although it has been proposed that composite multidimensional primary and secondary outcomes might ultimately act as a proxy for such. Finally, resonating with the schizophrenia paper, future work needs to be of longer duration and consider sample sourcing, trial execution and intervention study designs that preserve causal inference.

We're all familiar with the carrot and stick approach, and we all prefer the carrot, right? Pike et al[3] explored reinforcement learning in individuals with mood and anxiety disorders. Utilising data from 27 relevant studies ($N = 3085$), participant-level parameters were extracted. In a rather clever design, on top of a 'regular' meta-analysis, a novel computational simulation allowed the generation of a proposed model of underlying mechanisms outside the specific study from which they were derived. The authors took the originally reported model from each study and used its reported parameters to simulate choice behaviours on five new tasks. A selection of reinforcement learning models was then applied to the whole large data-set, and parameters were extracted using Bayesian model averaging. This circumvented the inevitable 'apples and oranges' issue when comparing studies of varying design, even when looking at the same principled psychological construct. In the simulation meta-analysis, those with depression or anxiety showed significantly greater punishment learning rates and lowered reward learning rates; such differences were not seen in the original 'standard' meta-analysis.

What does this tell us? First of all, from a methodological perspective, it would appear that there are greater gains in signal from the simulation approach, which has wider potential application. Second, individuals with anxiety and depressive disorders update learned values more after getting a punishment and less after gaining a reward. This differentiates what is happening from an alternative hypothesis for negative affective bias, namely that it is due to a greater subjective valuation of negative outcomes, which was not seen here. In other words, there were no differences in how much participants *disliked* the outcomes; rather, it is variation in *learning* that may be perpetuating negative affective bias symptoms in the clinical conditions. Punishment learning, not punishment sensitivity, appears key. The authors argue that large-scale work such as this allows computational psychiatry approaches to quantify underpinning neuro(patho)physiology, such as dopaminergic neurons' phasic firing to reward. This gives us better mechanistic hypotheses for illness behaviours that can be tested experimentally. Moreover, this might open a path to novel intervention targets. The example given is that rather than getting patients to pay less emotional importance to negative outcomes and tolerate the distress (as is typical of, for example, dialectical behavioural therapy), interventions might better focus on directly modifying a responsive behaviour.

'Can you read my mind?' – every psychiatrist has been asked it at a party. But a psychiatric stethoscope: how about that as an idea? Though we can hear a heartbeat by putting our ear to a chest, there is no question that a stethoscope, with its power to amplify, meaningfully aids the task. Psychiatrists are trained not only to hear what is being spoken but also to extract the subtle and wordless expressions contained around and within speech. As is the case with anything that relies on humans, results can be inconsistent between individuals and are prone to error and bias. Rezaii et al[4] suggest that harnessing the potential of computer-based natural language processing (NLP) could help to illuminate what the human ear easily misses and potentially aid diagnosis across mental health conditions. Ubiquitously used, this branch of computer science creates artificial neural networks to analyse, categorise and generate text. In daily life, it looks like the helpful word suggestions that come up as we write messages, captions at the bottom of a video or the tone flagging seen in some email software, but it is used in more ways within areas as diverse as marketing and forensic psychology. Within psychiatry, it has shown some promise already with identifying incoherent language and implicit references associated with psychosis, as well as in predicting suicidal behaviour. As NLP training data-sets grow and become more sophisticated, we are likely to see advances in non-verbal communication analysis, as well as the ability to suggest diagnostic categories based on patient sample text and more objectively define new categories and boundaries among current diagnoses. Of course, there are drawbacks as well. Like all things made by humans, we must acknowledge that bias will be baked in, and there are significant ethical considerations around patient privacy and potential exploitation of the technology.

https://doi.org/10.1192/bjp.2022.45 Published online by Cambridge University Press

It also threatens to diminish the clinical relationship, directly by reducing the amount of time a clinician spends or indirectly by depersonalising the patient to some extent. However, if the objective augmentation of clinical accuracy significantly aids therapeutic success, it is hard to ignore the potential. Only time will tell whether NLP can be a psychiatric stethoscope that helps us better hear what is already there. Which leads us nicely on to a cautionary tale in our next piece.

EPRs: we all love their easy intuitive interfaces and how time-saving they are. Apologies, our sarcastic language betrays our biases. Well, perhaps surprisingly, no one had used large-scale quantitative content analyses to ascertain the magnitude of bias represented in language actually used in EPRs and how that might affect care. And then two come along at once. First, Himmelstein et al[5] examine how frequently stigmatising language is used. For example, describing a patient as displaying 'drug-seeking behaviour' or as 'non-compliant' or 'malingering'. They included three clinical areas: people with diabetes, substance use disorder and chronic pain, and included patient-level illness severity, ethnicity, age and gender as potential explanatory variables. The outcome was the occurrence of stigmatising words or word stems (as well as pairs of words such as 'substance' and 'abuser' juxtaposed) in a patient's EPR record. The lists of stigmatising words were derived from national guidelines for appropriate language use in each clinical area. A total of 29 783 patients were included, totalling 48 651 admission notes written by 1932 clinicians. Around 6.9%, 3.4% and 0.7% of diabetes, substance use disorder and chronic pain patients' admission records, respectively, contained one or more stigmatising words or phrases. In multivariable linear probability modelling, demographic variables showed no effect on the probability of an admission note containing stigmatising language, with the single exception of having Black ethnicity, which increased the probability compared with all others. Clinician profession, age and gender, similarly, had no effect.

Sun et al[6] explored racial and ethnic biases in EPR data of patients requiring COVID-19 testing at a large hospital in Chicago. This included 18 459 patients with a broad range of ICD-10 diagnoses, including mental health conditions, but here a machine learning model was trained to classify any sentence in an EPR note as either negative, positive or 'out of context'. After training, the algorithm could achieve reasonable performance on unseen notes and was used to annotate the entire data-set for inferential analyses. To uncover patterns of stigmatising language use, the authors used multilevel logistic regression to model the probability of a record containing negative language, with independent predictors of age, sex, insurance provider, marital status, primary language and type of hospital visit, alongside ethnicity and race. Once again, being Black was more likely to be associated with negative use of language in notes, as was having federal-provided medical cover (versus employer or private insurance). They also found that those over 45 years old, unmarried, and encountered as out-patients (versus in-patient or emergency department visits) were more likely to have negative language use in their EPR. We may say we practice without bias, but ghostly typed fingerprints across electronic records tell a different tale. As well as the unpleasantness of the frank prejudice, it would seem reasonable to infer that this affects clinicians' decision-making. Further, there could be future implications as we move to an era where narrative EPR data are to be used in applications that assist (or worse, automate or replace) clinical decision-making.

Finally, drug-driving is clearly illegal, but we have fewer data quantifying how cannabis impairs performance behind the wheel than with alcohol. Adverse effects of cannabis on cognitive functioning have been well documented, with some specific findings related to maintaining lane position, but fewer data are available on other aspects such as crash risk and the duration of vulnerability following drug consumption. The literature has also been limited by attempts to precisely quantify tetrahydrocannabinol (THC): typically, low levels have been administered in non-real-world dosing regimens to small sample sizes. Marcotte et al[7] recruited 191 regular cannabis consumers, who were randomised to receive either placebo or 5.9% or 13.4% THC cannabis, which they smoked *ad libitum*. Participants went on a driving simulator at various time points, while their performance was monitored objectively using the composite driving score (CDS) and subjectively according to their perceptions of how well they were doing. Perhaps unsurprisingly, smoking cannabis resulted in a non-trivial poorer performance, though these participants did not actually crash more frequently. However, more interestingly, there were no differences between the two strengths of cannabis or among participants with different THC blood concentrations, and CDS was not affected by the quantity or frequency the participant had smoked in the prior 6 months. Driving performance typically returned to normal about 4.5 h after cannabis consumption, but participants' perceptions were that this occurred by about 90 min, indicating false confidence in their actual abilities. Many people will understandably highlight that one shouldn't drive at any point in the period after consuming drugs (or alcohol), but given that some people do, it is helpful to be able to evidence that self-perception is a poor guide and that a minimum of 4.5 h between consumption and driving is a necessity.

References

1 Taipale H, Schneider-Thoma J, Pinzon-Espinosa J, Radua J, Efthimiou O, Vinkers CH, et al. Representation and outcomes of individuals with schizophrenia seen in everyday practice who are ineligible for randomized clinical trials. *JAMA Psychiatry* 2022; **79**(3): 210–8.

2 Rush AJ, Sackeim HA, Conway CR, Bunker MT, Hollon SD, Demyttenaere K, et al. Clinical research challenges posed by difficult-to-treat depression. *Psychol Med* 2022; **52**(3): 419–32.

3 Pike AC, Robinson OJ. Reinforcement learning in patients with mood and anxiety disorders vs control individuals: a systematic review and meta-analysis. *JAMA Psychiatry* [Epub ahead of print] 2 Mar 2022. Available from: https://doi.org/10.1001/jamapsychiatry.2022.0051.

4 Rezaii N, Wolff P, Price BH. Natural language processing in psychiatry: the promises and perils of a transformative approach. *Br J Psychiatry* [Epub ahead of print] 7 Jan 2022. Available from: https://doi.org/10.1192/bjp.2021.188.

5 Himmelstein G, Bates D, Zhou L. Examination of stigmatizing language in the electronic health record. *JAMA Netw Open* 2022; **5**(1): e2144967.

6 Sun M, Oliwa T, Peek ME, Tung EL. Negative patient descriptors: documenting racial bias in the electronic health record. *Health Aff* 2022; **41**(2): 203–11.

7 Marcotte TD, Umlauf A, Grelotti DJ, Sones EG, Sobolesky PM, Smith BE, et al. Driving performance and cannabis users' perception of safety: a randomized clinical trial. *JAMA Psychiatry* 2022; **79**(3): 201–9.

https://doi.org/10.1192/bjp.2022.45 Published online by Cambridge University Press