

A Year in Statistics – the view from the trenches

EBM Methods Verdicts

Last year we launched the EBM Methods Verdicts as part of an attempt to keep on top of developments in Statistics that will impact on the way we do research.^{1,2} We focused on Statistical Methods and in particular on the output from five different journals: Journal of Clinical Epidemiology, Statistics in Medicine, Statistical Methods in Medical Research, BMC Medical Research Methodology, and Biostatistics.

The idea is simple: identify papers that we believe are likely to change the way we analyse or report evidence syntheses, randomised trials, or studies related to diagnosis, prognosis or monitoring. Our editorial team typically focuses on one of these per month and aims to write a short commentary with a ‘verdict’ explaining how this will change how we do things. We are also interested in people [submitting](#) proposals for similar short pieces.

Machine Learning – more weapons in the armoury

Bayesian neural networks were the rage in the early ’90s. Work in this area seemed to go quiet for most of the ’00s but has exploded back into our consciousness in the last few years. This is clear by the number of papers published on the topic (**Figure 1**) with nearly half of the total appearing in the last two years.

Statisticians by their nature are a sceptical bunch, and this trait is particularly evident for artificial intelligence (AI) and machine learning. Reservations that are well placed following the high-profile break-down of the ‘Google flu trends’ algorithm, and the allegations that surfaced after IBM’s AI tool ‘Watson for Oncology’ made “unsafe and incorrect treatment recommendations”.³ A 2019 systematic review of clinical prediction models appeared to confirm what many statisticians have suspected for some time - logistic regression works just as well as machine learning for clinical prediction. Predictably, in studies where machine learning did better than logistic regression, the validation procedures were often sub-standard or poorly reported.⁴

Is the case closed or should statisticians remain vigilant? After all, both Garry Kasparov and Lee Sedol, in chess and go respectively, believed they could not be beaten by AI and were humbled in the process. Some recent examples should make statisticians sit up and take notice. DeepMind’s future acute kidney injury algorithm⁵, a recurrent neural network model, outperformed a logistic regression at predicting any AKI by a potentially clinically relevant margin (AUC 92% vs 86%). A machine learning technique also beat logistic regression in an experiment based on 243 real datasets⁶ and in a Kaggle competition of lung cancer detection, the top three performing models for lung cancer detection all utilized convolution neural networks.⁷

There is no doubt that there is a lot of hype around AI; few algorithms are validated prospectively, and even fewer are actually integrated into clinical practice. What is clear is that researchers interested in prediction and classification should perhaps keep one eye over their shoulder or fear being left behind!

p-values on the firing line

The year has also seen a resumption of the ongoing quarrels around the use of null hypothesis testing, statistical significance and p-value reporting, as highlighted by the paper by Aguinis *et al.* in this journal.⁸ This fans the flames of a debate that has persisted for some time,

summarised by an enormous – and extremely influential – special issue of *The American Statistician*, which runs to more than 400 pages and drew contrasting opinions from many prominent statisticians.⁹

Concerns that p-values are at least partly to blame for poor scientific quality have led some journals to enforce a blanket ban on the use of p-values entirely. Another highly cited recent polemic called for an end to the use of “statistical significance”, if not quite the demise of the p-value itself.¹⁰ The Royal Statistical Society’s outreach journal, *Significance*, then took up the baton with its issue titled ‘The S Word’. One article suggested that the associated language be recast in terms of “outliers”, a familiar concept even for many inexperienced data analysts, which may well help in improving initial understanding.¹¹

If any consensus can be reached in a subject that clearly continues to provoke strong opinions, it is surely that p-values presented in isolation are seldom sufficient to answer a well-formed research question. That effect size estimates, together with uncertainty measures such as confidence intervals, are usually more appropriate is well-established and has long been reflected in guidelines such as CONSORT.¹²

Of the criticisms that have been levelled at p-values, two of the most damaging are misinterpreting the p-value as the probability that the null hypothesis is true, and that the quest to obtain a p-value less than 0.05 is a trenchant and praiseworthy scientific endeavour *per se*. Yet both of these misconceptions reflect a lack of understanding of the scientific and statistical method, rather than that the p-value itself is inherently at fault. Are they simply associated with poor research practice, rather than its primary cause?

Our view is that attempts to impose generalised rules about which statistical measures are suitable to use, or not to use, in a way that makes them applicable in all circumstances, is unhelpful and potentially counterproductive. Perhaps, as Aguinis *et al.* note, “a focus on education and reform may be more helpful than the abandonment of statistical significance testing”.⁸

We look forward to 2020: to new methods, old debates, and plenty of data to play with.

References:

1. Fanshawe TR, Perera R. Automatic extraction of quantitative data from ClinicalTrials.gov to conduct meta-analyses. *BMJ Evidence-Based Medicine* Published Online First: 17 July 2019. doi: 10.1136/bmjebm-2019-111206
2. Fanshawe TR, Perera R Conducting one-stage IPD meta-analysis: which approach should I choose? *BMJ Evidence-Based Medicine* 2019;24:190.
3. Ross C. IBM’s Watson supercomputer recommended ‘unsafe and incorrect’ cancer treatments, internal documents show 2018/19 [cited 2019 04/12/2019]. Available from: <https://www.statnews.com/wp-content/uploads/2018/09/IBMs-Watson-recommended-unsafe-and-incorrect-cancer-treatments-STAT.pdf>.
4. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol.* 2019;110:12-22.
5. Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature.* 2019;572(7767):116-9.
6. Couronné, R., Probst, P. & Boulesteix, A. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics* 19, 270 (2018) doi:10.1186/s12859-018-2264-5

7. Kadir T, Gleeson F. Lung cancer prediction using machine learning and advanced imaging techniques. *Transl Lung Cancer R*. 2018;7(3):304-12.
8. Aguinis H, Vassar M, Wayant C. On reporting and interpreting statistical significance and p values in medical research. *BMJ Evidence-Based Medicine* Published Online First: 15 November 2019. doi: 10.1136/bmjebm-2019-111264
9. Wasserstein RL, Schirm AL, Lazar NA (2019) Moving to a World Beyond “ $p < 0.05$ ”, *The American Statistician* 2019; 73:sup1, 1-19
10. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature* 2019; 567:305–7
11. Sheldon N. What does it all mean? *Significance* 2019; 16(4):15-7
12. Schulz KF, Altman DG, Moher D. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010; 340:c332

PubMed Publications identified using 'Machine Learning'

