Eye (2020) 34:795–796
https://doi.org/10.1038/s41433-019-0682-x

The ROYAL COLLEGE of OPHTHALMOLOGISTS

EDITORIAL

A reality check on big data

Tom H. Williamson[1,2]

Received: 15 October 2019 / Accepted: 17 October 2019 / Published online: 13 November 2019

Big Data, Tom H Williamson, 14 October 2019
Big data can be defined as "an accumulation of data that is too large and complex for processing by traditional database management tools". The three Vs of big data are:

Volume—large low-density unstructured datasets.

Velocity—received very rapidly, e.g., accumulation may be real time.

Variety—inputs are highly variable e.g., images, audio, text, web clicks.

If we accept this definition the label is almost always erroneously applied in ophthalmology where our datasets are of a size which allow conventional or artificial intelligence analysis (another term used loosely and usually indicating artificial neural networks or deep learning). The first appearance was with the use of "big data people" by Charles Tilly in 1980 referring to historians using data in cliometrics (applying methods used in statistics and economics to history). When it morphed into its current usage is uncertain.

The phrase is now common in the public vernacular and therefore the definition varies as the use of the label develops. It might be expanded to mean the datasets of a few thousand to those of millions of records along with the analytics to study these datasets. Whatever its meaning, the belief that the sizes of the datasets are so big as to iron out potential biases in data collection is certainly wrong. Big data analyses have the same potential problems as observational studies in medicine, including missing values, confounding, lack of generalisability and bias control to name a few [1, 2]. Over-simplistic application has led to error before [3].

✉ Tom H. Williamson
tom@retinasurgery.co.uk

[1] Department of Ophthalmology, St Thomas Hospital, London SE1 7EH, UK

[2] Department of Engineering and Biological Sciences, University of Surrey, Guildford GU2 7XH, UK

An evangelical belief that big data are infallible because of their size does us no good, and will lead us down the wrong road if we do not exercise our usual scientific scepticism and criticism. For example, as datasets increase in size spurious correlations also increase. If a simulated set of 200 variables is given only 1000 data points each, then a correlation would be expected to be found even though the variables are meaningless. By definition the use of big data enables the researcher to fit a hypothesis to the findings and not the other way around, introducing potential for biases. The data accumulation allows a researcher to reach a finding and stop analysis rather than to force analysis at a predetermined time point. Any findings are associations which will then need experimental testing [4].

Some areas where big data have been used well is in media and entertainment (e.g. predicting viewing patterns on Netflix), weather forecasting, transport, government (security and welfare) and banking. In healthcare Apple has introduced Healthkit, Carekit and Researchkit to allow the recording of data but we will need to wait to see how these work. It is however a far cry from the extremely structured data collection and validation that we expect in clinical research. Whereas we are intolerant of incomplete or inaccurate data collection in research, there is an expectation that somehow big data, because of its size, will miraculously iron out such errors. It is possible however that instead it is amplifying such error or at least giving it false credibility. When good data and bad data are combined the weakest link dictates. It may be that higher statistical thresholds will be required [5].

The input behaviours of the users of a digital technology cannot be altered post hoc by analysis. A good example is the use of EMR system only for surgical input and not as a full clinical record. If used along with a written clinical record, there is the possibility of recording for medicolegal purposes a surgical complication in the written notes and not in the EMR. Hey presto your digital record shows a low complication rate to compare yourself with other surgeons. I have seen examples of this when studying records for medicolegal expert opinions. The ability to game the

complexity of the clinical presentation can also be used to make yourself look good. Many clinical features in medicine are subjective and therefore the physician can consciously or subconsciously increase their complexity assessment of the patient. The effect of these errors will not disappear in analysis.

Using data to monitor surgical performance without any data validation will lead to the wrong kind of change in behaviour, especially if livelihoods are at stake. If cardiac surgeons are going to be stopped from operating because of mortality rates, they can reduce their rates by avoiding complex cases. Can this be overcome by complexity adjustment, not if the grading of complexity is subjective and done by the surgeons themselves. Perhaps also to reduce mortality the surgeon might operate very conservatively potentially increasing post-operative morbidity but reducing mortality. Some of these adjustments in practice can be conscious but some may be subconscious. Such behaviour change is potentially detrimental to the patient.

We are rightly sceptical when a single surgeon presents us their results. We also need to be sceptical when a group of single surgeons present us their results albeit through a multiuser collection system especially when they are put into a competitive environment with each other. Without data validation, as seen in prospectively structured multicentre trials, we are uncertain of the veracity of the data even in full EMR usage. Time constraints and convenience are likely to affect data entry. The threat of "big brother" influences our behaviour and data collection.

The looseness of the data collection is detrimental to scientific analysis. There are many biases. We can use the results with the hope to influence behaviour in a positive way, but the data then becomes useful for social or political reasons and less so scientifically. If the data are used as a tool to influence surgeons, it will become biased. If the data are only used for scientific enquiry it is more likely to remain unbiased. In simple terms if the surgeons remain anonymous, they are encouraged to be truthful if they are named they may be tempted to "game" the system. There are many examples of how crude targets in society have changed behaviour to allow meeting the target. For scientific progress we need the truth good or bad but undoctored.

There is no doubt big data will provide insights and potentials for improvement in medical care from predictive modelling, clinical decision support, disease monitoring and safety surveillance, allocation of resources and public health. Data mining will allow classification, clustering and regression to be applied. It will however not replace structured prospective clinical research as suggested by some.

Compliance with ethical standards

Conflict of interest The author declares that he has no conflict of interest.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Tanaka S, Tanaka S, Kawakami K. Methodological issues in observational studies and non-randomized controlled trials in oncology in the era of big data. Jpn J Clin Oncol. 2015;45:323–7.
2. Sinha A, Hripcsak G, Markatou M. Large datasets in biomedicine: a discussion of salient analytic issues. J Am Med Inf Assoc. 2009;16:759–67.
3. Butler D. When Google got flu wrong. Nature. 2013;494:155–6.
4. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA. 2013;309:1351–2.
5. Ioannidis JPA, Khoury MJ. Evidence-based medicine and big genomic data. Hum Mol Genet. 2018;27:R2–R7.