All about love and big data

Alice Richardson[1], Tony Badrick[2]

[1]National Centre for Epidemiology and Population Health, Australian National University, Canberra, Australia; [2]Biomedical Science, Bond University, Queensland and RCPA Quality Assurance Programs, Sydney, Australia

Correspondence to: Tony Badrick. Biomedical Science, Bond University, Queensland and RCPA Quality Assurance Programs, Sydney, Australia. Email: Tony.badrick@rcpaqap.com.au.

Received: 30 August 2018; Accepted: 12 September 2018; Published: 26 September 2018.
doi: 10.21037/jlpm.2018.09.03
View this article at: http://dx.doi.org/10.21037/jlpm.2018.09.03

Introduction

"Love means never having to say you're sorry"—according to Erich Segal in his novel "*Love Story*" which was popularised in the 1970 movie adaptation starring Ali McGraw and Ryan O'Neal.

With the advent of electronic databases, pathology laboratories the world over find themselves in possession of large amounts of data. It is tempting to think that mining these databases for their clinical/medical nuggets are going to be pleasant as being in love. In this column we will be urging a measured approach, because it is hardly ever true that big data means never having to do a variety of tasks before, during and after the data mining process itself.

We'll address four "Nevers" in this column: never having to say what your research question is, never having to say what your model is, never having to distinguish between association and causation, and never using a classical statistical method again (1).

Saying what your research question is

Clarity of thought around the research question remains at the core of effective use of clinical databases in laboratory medicine research. We do acknowledge that hypothesis generation is a legitimate research endeavour, which often takes a highly operationalised (or statistical) view of a hypothesis. In this statistical view, a hypothesis is typically of the form "parameter in a model equals a number". The number is often zero. On the other hand, a less operationalised view of a hypothesis would be "there is a relationship between x and y". Either way, unless the research question is clear you risk committing a type III error: the right answer to the wrong question (2).

Saying what your model is

The concept of a statistical model is one of the most powerful constructs of scientific research. The term "model" itself is an ambiguous term that carries many meanings outside of science, ranging from the model who glides down the catwalk in the latest designer clothes, to the model train installed in the spare room. However, both of these contain elements of the statistical model—they are idealised representations of reality, with just enough complexity to capture the essence of reality but not too much. Such models are powerful tools for exploring relationships and should not be discarded just because a data set is big.

On a related note, let's make it clear that a nonparametric method might be free of the parameters typically associated with models, but it is not necessarily free of assumptions, or for that matter of models! The usual assumptions of independent observations and random sampling from a given population are very rarely dropped, even in the typical scenarios for the use of nonparametric methods when the sample size is small or normal distributions cannot be assumed.

Distinguishing between association and causation

The distinction between association and causation is one of the very first statistical principles drummed into students and researchers. The importance of the difference is drawn particularly strongly in epidemiology where the identification of a risk factor plays an important part in public health messaging (3) but the ability of observational data sets to contribute to the causation argument is still questioned.

One of the recent spectacular failures of big data to provide accurate predictions is related to distinguishing between association and causation. Remember Google Flu Trends (4)? This project was launched in 2008 and attempted to predict flu outbreaks by tracking instances of Google searches on "influenza" and related terms. The data is therefore very unlike the usual kind of quantitative measures that clinical biochemists might encounter on a day-to-day basis. It is also possible that despite the fact that Google had millions of counts of search terms with location, time and so on to support it, key variables that are much more strongly associated with flu diagnosis were unavailable. By 2013 the predictions were becoming quite inaccurate and Google eventually gave up maintaining the site though the data is still there for research purposes.

On another related note, it is useful to remember that classification is an exercise in prediction, not necessarily a separate activity. Given the value of a set of biomarkers, for instance, an individual can be classified (predicted to be in a group) such as diseased/healthy, type A/type B; if the outcome is known through clinical notes or subsequent biomarker analysis, then the classification can be checked for its accuracy and rates of sensitivity, specificity, positive predictive value and so on calculated across a large number of individuals.

Using a classical statistical method again

Some classical statistical methods do scale up to the terabytes of data that pour into clinical databases; for example, the logistic regressions used in the analysis of data from wearable medical sensors (5). Others have undergone refinement to handle the large volumes of testing required; for example, the use of the false discovery rate (6) to determine significance of thousands of *t*-tests to replace assessment of thousands of P values as done by Si and Liu (7). Some classical statistical methods are designed to compress large datasets to make them amenable to application of further classical methods e.g., principal components analysis for dimension reduction. Examples include the nine biomarkers in a PCA on 262 individuals in a study of chronic rhinosinusitis (8); twenty-six food groups in a PCA on 4,316 individuals in a study of diabetes (9); and 80 genes in a PCA on 20 individuals in a study of micro-RNA in urine (10).

Simple data mining methods e.g., decision trees also still have a place in laboratory medicine research. The simple rules generated by a decision tree can be very compelling. However, researchers do still need to take care that a training and test data set are employed, to guard against overfitting of the training data.

Conclusions

Big data is one of the best things to have happened to lab medicine research at least in terms of sample size leading to increased power. But researchers need to be up front about their research question (and whether it is associative or causative) and about their model, and open to the continued use of classical methods where appropriate. It is very rarely going to be the case that a large quantity of data frees you from having to do any of these things. Only then will the intersection between statistics, computer science, and laboratory medicine avoid being a battleground or a barren wasteland and become the rich and harmonious space where evidence for action can be generated.

Acknowledgements

None.

Footnote

Conflicts of Interest: The authors have no conflicts of interest to declare.

References

1. Richardson A, Signor BM, Lidbury BA, et al. Clinical chemistry in higher dimensions: Machine-learning and enhanced prediction from routine clinical chemistry data. Clin Biochem 2016;49:1213-20.
2. Kimball AW. Errors of the third kind in statistical consulting. J Am Stat Assoc 1957;52:133-42.
3. Rothman KJ, Greenland S. Causation and causal inference in epidemiology. Am J Public Health 2005;95 Suppl 1:S144-50.
4. Google Flu Trends. Available online: https://en.wikipedia.org/wiki/Google_Flu_Trends
5. Manogaran G, Lopez D. Health data analytics using scalable logistic regression with stochastic gradient descent. Inter J Adv Intell Paradigms 2018;10:118-32.
6. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc (Ser B Methodological) 1995;57:289-300.
7. Si Y, Liu P. An optimal test with maximum average power while controlling FDR with application to RNA-Seq data.

Biometrics 2013;69:594-605.

8. Tomasssen P, Vanelplaset G, Van Zele T, et al. Inflammatory endotypes of chronic rhinosinusitis based on cluster analysis of biomarkers. J Allergy Clin Immunol; Tomassen P 2016;137:1449-56.
9. Batis C, Mendez MA, Gordon-Larsen P, et al. Using both principal component analysis and reduced rank regression to study dietary patterns and diabetes in Chinese adults. Public Health Nutr 2016;19;195-203.
10. Ben-Dov IZ, Whalen VM, Goilav B, et al. Cell and Microvesicle urine microRNA deep sequencing profiles from healthy individuals: observations with potential impact on biomarker studies. PLoS One 2016;11:e0147249.

doi: 10.21037/jlpm.2018.09.03
Cite this article as: Richardson A, Badrick T. All about love and big data. J Lab Precis Med 2018;3:78.