Clinical Chemistry in Higher Dimensions: Machine-Learning and Enhanced Prediction from Routine Clinical Chemistry Data

Richardson Alice¹, Signor Ben M², Lidbury Brett A², Badrick Tony^{2,3*}.

1. National Centre for Epidemiology & Population Health, Australian National University, 62 Mills Rd, Acton, ACT 2601, Australia
2. Pattern Recognition and Pathology, John Curtin School of Medical Research, Australian National University, 62 Mills Rd, Acton ACT 2601, Australia
3. RCPAQAP, Suite 201/8 Herbert Street, St Leonards, NSW 2065, Australia

*corresponding author: tony.badrick@rcpaqap.com.au

Highlights:

- There are a number of “myths” surrounding the power of Big Data to answer research questions in a variety of fields including clinical biochemistry.
- Astute clarification of these “myths” reveals a number of applications for Big Data in clinical biochemistry, by judicious use of modern data mining methods, along with classical statistical thinking.
- There have been genuine advances in clinical biochemistry through the use of machine learning techniques.

Keywords: anaemia; big data; bilirubin; biomarkers; hepatitis; liver function tests; misconceptions; predictive modelling; statistics

Acknowledgements: The authors wish to thank the editors for the opportunity to contribute this article, and the referees for their suggestions that improved the quality of it.

Funding: This work was supported by the *Quality Use of Pathology Programme* (QUPP), The Commonwealth Department of Health, Canberra, ACT 2601, Australia.

Abstract: Big Data is having an impact on many areas of research, not the least of which is biomedical science. In this review paper, big data and machine learning are defined in terms accessible to the clinical chemistry community. Seven myths associated with machine learning and big data are then presented, with the aim of managing expectation of machine learning amongst clinical chemists. The myths are illustrated with four examples investigating the relationship between biomarkers in liver function tests, enhanced laboratory prediction of hepatitis virus infection, the relationship between bilirubin and white cell count, and the relationship between red cell distribution width and laboratory prediction of anaemia.

1. Introduction

The arrival of the new millennium inspired many professional groups to reflect on where statistics research had come in the last 100 years or so, and where it was heading. Breslow [1] noted that "... the statistics of the twenty-first century will be heavily influenced by the revolutionary developments in technology, particularly in the information and biomedical sciences, and by the availability of vast new repositories of geographic and molecular data". Leaving aside the geographic component, the first 15 years of the millennium have proven to be as Breslow thought. "Big Data" is having a major impact in many areas of biomedical science, particularly in pathology where the most obvious of these impacts has been the reclassification of malignancy [2]. The clinical biochemistry community has similarly been reflecting on the direction of its research effort into the new millennium, and quantitative methods are bound to play a part. Universal reference intervals and validation of formulae such as the estimated Glomerular Filtration Rate are examples of the use of Big Data to provide better interpretation of clinical biochemistry data. Foster et al [3] bring the issues of tuning quantitative methods to the attention of the biomedical engineering community.

The interpretation of pathology tests, particularly those involved with screening, is complex and relies on an understanding of the diagnostic sensitivity and specificity of the test and the prevalence in the community of diseases that the test can predict. Tests become more predictive if the pre-test probability for disease can be improved. This could be achieved by better history, more specific tests or by a better understanding of the interrelationships, if they exist, between the routine tests that may be used at the screening episode. Some of these interrelationships are well known, such as urea and creatinine, or aspartate transaminase (AST) and alanine transaminase (ALT). These tests together reinforce the presence of a possible disease state whilst each can provide some additional information about that disease. Finding more subtle relationships amongst routine tests requires more intricate techniques that could include some of the powerful new techniques from machine learning such as decision trees and support vector machines. Recursive partitioning-based decision models have been applied to medical knowledge domains [4, 5], and learning from decision trees provides advantages of applicability to both Gaussian and non-Gaussian data, as well as options for multiple decision boundaries [6]. Support vector machines (SVMs) provide a very powerful classification and regression pattern recognition tool through the analysis of images between data points in high dimensional space (kernels), without high computational cost [7]. A combined tree and SVM method was successfully applied to a study of assay redundancy for liver function test (LFT)

profiles, combining the advantages of each to recommend two LFT markers as sufficient for screening community patients [8]. Machine learning is therefore a data analysis initiative that can be used by the clinical biochemistry community.

1.1. Structure of this paper

There is a very large amount of routine diagnostic pathology testing performed every year in western countries. According to The Royal College of Pathologists of Australia, more than 11 million Australians have at least one pathology test a year for a variety of reasons (<https://www.rcpa.edu.au/getattachment/4501e94c-251e-4f17-91e8-fa695a7d6139/FctSht2-Why-Path-Test.aspx>), which are collected on both healthy and diseased subjects. These data represent a significant data mine, and subject to ethical approval, make available an enormous pool of information often serial in a subject over many years, and often with significant other physiological and demographic data. The use of this information can offer a way to make efficient use of existing data and offers a way to extract information from high-dimensional data sets. On the other hand, as with all methodological innovations, there are areas of active research, unanswered questions and traps for the novice user.

With that in mind, this review begins with a brief history and definitions of higher dimensions and machine learning. Then, the key messages about machine learning and enhanced prediction from routine clinical chemistry data are conveyed in the form of seven “myths” about machine learning. The practical use of machine learning for enhanced prediction in clinical biochemistry is illustrated using data obtained from routine pathology testing performed in Australia.

2. Definitions

2.1. Higher dimensions: what is Big Data?

The original terms that encompass the definition of Big Data are the three “V”s of Volume, Variety and Velocity (<http://www.ibmbigdatahub.com/infographic/four-vs-big-data>). Size, diversity, and the speed with which the data arrives are still at the core of the definition of Big Data, but other concepts are now often included in broader definitions, such as the way a problem is approached (technologies) and the uses to which it is put (decisions and solutions).

2.2 What is machine learning?

Machine learning refers to a set of tools and techniques, ranging from artificial neural networks and support vector machines to random forests and decision trees. It consists of many of the tools used in the activity known variously as “data mining”, “knowledge discovery in databases” (KDD) and “pattern recognition”. Peter Bruce (www.statistics.com) notes that many use the term “machine learning” synonymously with “predictive modelling”; in reality, some of the workhorses in predictive modelling and classification are statistical methods that date back as many as 80 years. This will become apparent in the examples used to illustrate this paper in the next section.

The machine part of the name comes about because machine learning, KDD and data mining all have a concept of automated analysis associated with them – the idea [9] being that there is too much data for statisticians to analyse without machine assistance. The speed and functionality of software for carrying out this automated process has moved ahead exponentially in the last two decades. Increasingly, analysts are able to access packages to do their machine learning on their desktops, and do not have to engage a software engineer to write a special-purpose application to do the work. Most statisticians would be familiar with R, the software environment for statistical computing and graphics. Products such as SPSS, SAS and even Excel have add-on modules that carry out machine learning algorithms.

The learning part of the name refers to the activity of assigning a predicted outcome (which can be categorical e.g. positive or negative; or continuous e.g. a real number) to a new observation on the basis of a database of individuals with known outcomes. The term pattern recognition is used interchangeably, where the pattern being recognised maps to the outcome being learnt. As a term, knowledge discovery is broader as it allows for discovery of associations, not just the stating of predictions.

2.2 Key concepts of machine learning

The definition of machine learning above can be supplemented with two important contrasts that help place machine learning and classical statistics in relation to each other. These contrasts are “algorithms and models”, and “train-test and fit-diagnose”.

2.2.1 Algorithms and models

The statistics community offers to researchers a variety of models for many research questions: the linear regression model, the logistic regression model, the factor analysis model etc. The machine learning community prefers to present algorithms, for example, for a decision tree, whereby a series of splits in the data are identified which best divide the observations into homogeneous sets.

2.2.2 Train-test and fit-diagnose

“Train-test” is the concept of training a machine learning algorithm on most of the data and then testing its efficacy on a small subset of data saved for the purpose. It came to prominence with the advent of Big Data because the quantity of data available meant analysts could afford to leave some data to one side for validation of the model later. But the train-test concept also works on small datasets through Harrell’s [10] approach.

The fit-diagnose approach of classical statistics depends very heavily on the model that is being used, whether it be a linear regression, logistic regression, factor analysis or any other model. An entire data set is used to fit a model. Then, any problems with the fit are diagnosed using simple numbers such as R, graphical methods such as residual plots, or formal tests such as the Hosmer-Lemeshow goodness-of-fit test for logistic regression.

2.2.3 Description and prediction

Another dichotomy where both parts are acknowledged by the machine learning community is that of description and prediction. Together, these two terms capture the majority of analysis activities in both machine learning and classical statistics.

The description of a data set includes visualisations and statistics such as means, standard deviations and correlation coefficients from the classical side, and more innovative visualisations along with techniques such as association rules from the machine learning side.

Prediction includes linear regression, logistic regression and discriminant analysis from classical statistics, and decision trees, random forests, support vector machines and neural networks from machine learning methodology.

3. Seven myths of machine learning

In this section, we summarise four studies recently published by the authors of this review and their associates, identifying both machine learning and classical analyses used in each. We follow by using these studies to illustrate and present seven “myths” of machine learning. We choose to emphasise our own work here, as we feel this better captures our experiences to convey an understanding of the concepts presented. The example studies highlighted below are: Richardson & Lidbury (2013) [11], Badrick, Richardson & Lidbury (2016) [12], Lidbury, Richardson & Badrick (2015) [8] and Badrick, Richardson, Arnott & Lidbury (2015) [13].

3.1 Example 1: Enhanced understanding of the relationship between GGT and other components of the routine LFT

In this first example, the authors showed how machine learning can be used to discover relationships. Lidbury, Richardson & Badrick [8] demonstrated the interrelationship of gamma glutamyl transferase (GGT) with other components of a routine liver function test (LFT) consisting of alkaline phosphatase (ALP), alanine transaminase (ALT), aspartate aminotransferase (AST), serum albumin, total serum bilirubin, and to an extent, lactate dehydrogenase (LD). The response variable for all analyses was the concentration of serum GGT (U/L). The analyses used both classical statistical methods (i.e. logistic regression) and modern machine learning methods (single decision trees and support vector machines).

Machine-learning Analysis: All decision tree and SVM analyses were performed using the R packages *rpart* [14] and *e1071* [15] respectively. Single decision tree rules based on calculated predictor variable thresholds (e.g. ALT < 55 U/L) at decision nodes were determined for each GGT response category (0 or 1). SVM predictions were based on 10-fold cross comparison of training and testing data to arrive at an overall mean percentage of class (0 or 1) prediction. Prior to SVM analysis, the combination of the response variable (GGT category) and predictor variables of interest were subject to tuning to calculate the best cost and gamma coefficients for the model [7]. Decision trees containing all predictor variables were developed to identify the most likely four to six predictors for further SVM analysis. Only these predictor variables were tested in the SVM models, with only three required for the final SVM prediction model.

Classical analysis: Logistic regression was performed for the same GGT response categories and associated pathology data using SPSS 20. Model fit was assessed by the Hosmer-Lemeshow (H-L) statistic. Variance inflation factors (VIFs) and correlation coefficients were assessed collinearity, and co-linear variables removed (where $VIF \geq 10$). Predictors were

deemed significant at $p \leq 0.001$. Using this LFT data to predict GGT category (elevated or within the laboratory reference interval) proved challenging, with acceptable model fit difficult to achieve, as measured by the H-L statistics.

Example 2: Enhanced laboratory prediction of hepatitis B (HBV) and C (HCV)

Richardson & Lidbury [7] described an empirical investigation of immunoassay results (HBV or HCV) and associated routine pathology data, by constructing single decision trees and ensembles [16], and using different data pre-processing techniques on the routine pathology data. The aim of the study was to use the resulting trees for the enhanced laboratory diagnosis of hepatitis virus infection, by exploiting the range of multi-variable pathology laboratory data associated with direct virus immunoassay testing.

The immunoassay marker examined for HBV infection was hepatitis B surface antigen (HBsAg), and for HCV, a polyclonal anti-HCV antibody response (HepC). As well as the specific immunoassay data, routine diagnostic pathology variables were included in the machine learning analyses.

Machine learning analysis: Decision tree analyses were performed using the R packages *rpart* [14]. Single decision tree rules based on calculated predictor variable thresholds (e.g. ALT < 55 U/L) at decision nodes were determined for each HBsAg response category (0 or 1 for negative or positive).

Classical analysis: No classical analysis was reported in this paper.

Example 3: Enhanced understanding of the relationship between bilirubin and WCC

The letter by Badrick, Richardson & Lidbury [12] was written to support and extend the observations of Tsai et al [17] who described a significant inverse association between total serum bilirubin and total white (blood) cell count (WCC) in a cohort of healthy Taiwanese adults. Badrick et al. investigated similar relationships in two distinct cohorts of Australian patients who had routine pathology performed, including assessments of inflammatory markers.

The first cohort consisted of community patients ($n = 980$), and included 510 cases testing positive to the presence of hepatitis B surface antigen (HBsAg). The second cohort consisted of intensive care unit (ICU) cases ($n = 955$). In order to reflect the Tsai et al. [17] study, cases with

elevated serum bilirubin ($>20 \text{ umol/L}$) and WCC ($>12 \times 10^9/\text{L}$) were removed from the community cohort prior to analysis.

Machine learning analysis: No machine learning analysis was reported in this letter.

Classical analysis: Bivariate Pearson correlations were calculated to assess the association in each cohort between total serum bilirubin and total WCC.

Example 4: Enhanced understanding of the longitudinal effects of RDW on laboratory diagnosis of anaemia

The final example demonstrates the use of cross-sectional and longitudinal data in the investigation of serum and blood markers of anaemia. The analyses presented by Lidbury, Richardson & Badrick [8] comprised a red cell survival model that predicted a previously reported longitudinal relationship between increasing red cell distribution width (RDW) and decreasing haemoglobin [18], but in cross-sectional data representing hundreds of community patients.

Machine learning analysis: No machine learning analysis was reported in this paper.

Classical analysis: The capacity of routine RDW measurements to predict the importance of second tier anaemia markers (e.g. ferritin, vitamin B₁₂), and hence likely anaemia aetiology, was the subject of analysis of covariance (ANCOVA) due to the “broken stick” distribution in relation to serum ferritin.

3.2 The seven myths

In considering the above examples, and the prospects of solving problems via access to Big Data and sophisticated machine-learning algorithms, a range of potential pitfalls must be outlined before embarking on these data-intensive investigations. We have chosen to approach these considerations by presenting “seven myths of big data” that we have identified in our own investigations.

Myth 1. Big data is universally big

As highlighted above, Big Data is not usually defined in absolute terms of size, but rather in terms of various “V”s. The original three “V”s were Volume, Variety and Velocity

(<http://www.ibmbigdatahub.com/infographic/four-vs-big-data>). The fourth one, added more recently, is Veracity.

Volume refers to the absolute size of the dataset. The absolute bigness of a data set depends on its intended use. It is also worth remembering that a data set has two dimensions - rows (or depth) and columns (or width). It is generally assumed that Big Data refers to deep data with many individuals contributing. However, wide data with many columns or variables measured, is just as much a big data set with its own statistical issues to be negotiated. Most classical statistical theory is built on the notion of data sets getting deeper, and new theory is being developed for situations where data sets are getting wider [19].

Variety refers to the different forms of data, ranging from numbers stored in traditional databases to videos and text such as social media posts. Pathology data such as those used in the illustrative papers of this article, are a standard numerical format, but free text in the form of clinical notes is also sometimes available.

Velocity refers to the speed with which the data accumulates. With approximately 500 million tests in Australia per year, that's well over 1 million tests per day. In business applications, the amount of data requiring storage can be many terabytes, and petabytes of data in research areas such as astronomy are not uncommon.

Veracity is more of a negative than a positive "V", and refers to the fact that the data may not be trustworthy, but be subject to errors, missing values or other inadequacies. In pathology, gross errors are not so much of an issue, as the quality standards of laboratories are rigorously maintained. On the other hand, missing values due to tests not being requested, and censored data such as levels that fall below detection limits, are regular occurrences in the output from laboratory machinery.

A fifth "V" sometimes added is Volatility. The volatility of sources for big data is considered more of a concern in the social sciences where individuals and entities are ever changing, but the concern also applies in the biomedical field. Assays, detectable limits, and fashions in testing are all prone to rapid change. Other "V"s have been introduced e.g. Visualisation and Value (<http://dataconomy.com/seven-vs-big-data/>) but they do not refer to characteristics of the data itself, so they are not considered here.

Our work has not ventured into the realm of truly Big Data. Although the databases we work with are potentially many millions of rows, by the time an appropriate subset of rows is identified, it is often in the thousands. Ultimately, the quality as opposed to the size of the data set is the prevalent issue. We need to know that the data is fit for the purpose for which it is to be used.

Myth 2. Big data means never having to say what your research question is

The risk in investigating a data set without any idea of what a research question might be is that a machine learning technique (or indeed any statistical technique) will typically provide an answer, but it may not be to the questions that you should have asked. Indeed, using big data will not necessarily remove the need to have a proper process for hypothesis generation. In fact, it may increase its importance.

Furthermore, different approaches, whether classical or machine learning, may be needed for different research questions. Different models may answer the research question better for different subsets of data (and when the sample size is big enough, the subsets are likely to be big enough to analyse sensibly too). “Most predictive models are statistical models” [20], which suggests that the border between machine learning and classical statistics is not completely impermeable.

Our practice is to clearly delineate research questions for each piece of work we do. Each time we have a hypothesis we wish to test and begin with a well-defined set of cases. The set consists of subjects that have a disease process which is classified by an abnormal pathology test. Then, using patient groups that are either positive or not for this disease, we look for relationships between the diagnostic tests. These have often revealed both known and unknown relationships within the disease. Once confirmed, the interpretation of these findings requires an understanding of the underlying physiology and may suggest new hypotheses that can be tested.

Myth 3. Big data means never having to say what your model is

The model we have in mind here is a statistical model, rather than, say, a mouse model or a cell culture model. So when “saying what your model is”, you are saying what basic structure holds true in your data, and then using the data to estimate the key parameters of the model to make

it specific to the data at hand. The Big Data myth is that if there is enough data, there is no need to say what that basic structure looks like, and it is sufficient to let the “black box” of a machine learning algorithm figure it out for itself.

The phrase “black box” is a very emotive one and Johnson [20] suggests that “predictive modelling” could be seen as a more neutral term. On the other hand, a black box takes input and produces output, as does a predictive model. But we would like to suggest that a predictive model could be as transparent as a simple linear regression. A predictive model doesn’t have to be opaque in the middle.

Some scientists think that approaches based on a model are limited in some way, and that it is always better to go for a more general approach that makes no model assumptions. Yet every approach makes assumptions. These may not be about the parameters of a distribution, but there is always something that has to be assumed, for example, independence. Durrant (personal communication) puts it this way: “‘no assumptions’ is formally equivalent to ‘no guarantees’”.

Breiman [21] says “the goal [of a model, either data or algorithmic] is not interpretability but accurate information”. This is a sound statement, especially bearing in mind that information could be description or prediction. Is the goal ever allowed to be interpretability? Perhaps not, if we are unlikely ever to have the “correct” predictors.

The classical methods in the examples all had specific models associated with them, be they logistic regression, ANCOVA or the bivariate Normal model that leads to a Pearson correlation. The machine learning methods we used sometimes built on these classical methods, such as following up a logistic regression with a random forest using the most important variables identified in the regression. It would also be possible to reverse the process and carry out the machine learning first. Either way, the statement of models is an important part of our process.

Myth 4. Big Data means never having to consider sampling theory, a standard error, or a p-value

The inspiration for a number of the techniques of machine learning was often the size of the data set requiring analysis, and that size was Big. But a Big Data set doesn’t always consist of data from an entire population. More often than not, it will be a very big sample from an even

bigger population. More often than not, there is a way of looking at the data that identifies the hyper-population from which the data came, allowing then for statistical inference through such tools as standard errors and p-values.

A data set might consist of millions of records of FBC, but if the records are all males 20 – 40 years old and the aim of the study is early diagnosis of late-onset Alzheimer's, then generalising predictions to the Australian population is likely to be invalid. And however large a sample is, if it is not representative of the population from which it is taken, no amount of statistics will make it so.

As Harford [22] suggests, it can be tempting to assume that because a data set is very large, it must capture an entire population. Occasionally it does, for example, all the votes in an Australian election. However, to imagine that, for example, inferences made on all users of a social media site (and there are millions of them) can be generalised to the Australian or world populations as a whole, clearly does not make sense.

This notion of the continued importance of statistical inference, even as sample size increases, is not well understood. Bezzina & Saunders [23] report on a very interesting study in which they presented a series of statements to business research academics. One statement was "When the whole population is used, no inferential statistics are required since the statistical summary of the data represents a parameter rather than a statistic." The authors regarded this statement as false, and only 22% of respondents agreed (33% said it was true and 25% didn't know). This is despite the fact that the authors were expecting respondents to disagree. This is clearly a misconception that needs to be addressed.

Gelman [24] points out that the question of how to approach statistical inference for population data presents itself in many contexts. As a conclusion, he presents the idea that if researchers think seriously about replications (for the goal of getting the right standard error), they might well get a better understanding of what they're trying to do with a model, and whether it is appropriate to declare that the standard error is zero, or that at some other level in the study, there is replication and a sensible measure of variability.

Myth 5. Big data means more valuable information

The results of a machine learning exercise on a Big Data set often have a lot at stake. Sometimes the stakes are more populist, such as predicting the next Oscar winner. Sometimes the stakes are much higher and more vital to business, industry or government, such as predicting “who will be admitted to hospital within the next year, using historical claims data” (www.heritagehealthprize.com/c/chp). The notion that one should gather all the relevant data together is a worthwhile one. However, data from different sources, collected in different ways and with different original purposes in mind can vary enormously in terms of its quality and relevance. A small dataset collected with care may easily yield more valuable information than a large one drawn together from mixed sources. Clear sample design, however big the sample will be, is an important part of study design. In fact, a Big Data set should often be scaled *down* through sub-setting, sampling, or some other inclusion criteria to ensure a better analysis will result.

It is certainly true that classical statistics theory suggests that the larger the sample size, the more powerful hypothesis tests will be and the more accurate sample statistics will be. But if the sample design is weak, or unknown/unknowable, then the results cannot be generalised to any sensible population, because the population from which the sample came cannot be clearly identified.

More variables do not necessarily help either. Feeding irrelevant variables into a model does not help improve its performance – rather, it leaves you open to the risk of overfitting. Overfitting means that a model doesn’t just follow the signal in the data concerning the relationship between variables – it also follows the noise. Overfitting is best solved by having training and testing data sets (see section 2.2.2), an approach which is easily implemented when the sample size is large enough.

In our illustrations, the data sets have limited capability to get ‘bigger’, as the subset of rows that fit the inclusion criteria may not be large. We have concentrated on single databases from pathology laboratories so that issues of varying quality data from varying sources are minimised.

Myth 6. Big data means observational data can be used to measure causal relationships

This myth is disproved in any first year statistics class (the phrase “correlation does not imply causation” is found in most first year textbooks and conveyed by most first-year statistics

lecturers). This taught version of the myth however is rather simplified. The more advanced version of the myth is more important for discussion because of the extra complexity of the algorithms used to measure causation in Big Data. Most exercises involving big data can quickly calculate a correlation, but the answer as to what is causing the observed relationship remains as elusive as ever.

It is also important to note that while data sets may be increasing in size, their fundamental nature is remaining the same. Data collection typically takes one of two forms - observational and experimental. Many of the biggest data sets are actually still observational [25] as opposed to experimental, and so are prone to all the downsides of observational data, most specifically the inability to assess cause-effect relationships over associations (see Myth 3).

The examples from our research are all classic cases of observational data. Plenty of models have been fitted and plenty of relationships proposed, but we recognise that we too are keen to undertake more research to identify causal pathways that underpin observed correlations such as the ones between RDW and anaemia [13].

Myth 7. Classical statistical methods are inadequate to deal with big data

Amongst the machine learning community, the methods of statistics are often dismissed as having many shortcomings. Maindonald (pers. comm.) lists a number of alleged shortcomings, including the following:

- A. Focus on central tendency and spread
- B. Assumptions don't match reality
- C. Don't handle missing values

The focus on central tendency and spread is typical in a first-year statistics course, and it is possible that many researchers in the biomedical sciences think that what they learnt in their first-year statistics course is all there is to learn. Nothing could be further from the truth.

Statistical techniques that have flowered since the 1990s include the bootstrap [26], wavelets [27] and the false discovery rate [28]. The quality of research has benefitted enormously from the discovery of these methods, even though none of them usually make it into a first-year statistics course.

The allegation around assumptions falls away further when the difference between statistical modelling and algorithmic modelling is taken into account. Statistical models typically come with restrictions, but have known distributional properties that make them very powerful. Algorithmic models developed in the data mining community have fewer assumptions. Cross-validation (i.e. the train-test of section 2.2.2) is widely used to study the power of the implementation empirically, but much less is known about their large-sample behaviour.

Dealing with missing values and other data pre-processing is a major issue for both Big and small data. Here, the data mining literature runs true. There is scope for trading off time spent in cleaning data against time spent in analysis. Pre-processing of Big Data can be very time-consuming for a range of reasons. The structure of data may vary subtly from a variety of sources. There may be missing values. Individuals may appear in multiple data sets but with multiple identifiers, and linking data from multiple sources is an essential field in computer science and statistical research.

Standard statistical methods often aren't sufficient for analysis of Big Data, in that they may be too slow or provide an insufficient summary. The p-values of thousands of t tests are far better replaced with statistics such as q-values [29, 5] and the false discovery rate of Benjamini & Hochberg [27].

Some of the perceived deficiencies in statistics come about because critics are considering statistics to consist only of "lesser" statistics when "greater" statistics has a lot to offer. "Greater" and "lesser" statistics were introduced by Chambers [30]. It would be useful for biochemists to embrace what "greater" statistics has to offer. Essentially, statistics is about decision-making in the presence of uncertainty. Horton [31] puts it this way: "The appropriate use of statistics ensures that variability and bias are addressed, suitable analytic methods are undertaken, interpretations are rational and defensible, and decisions are made to account for uncertainty."

The four examples already discussed show that it is possible to combine classical and machine learning methods on the same data set, and that both approaches to analysis can provide valuable insight on the same research question. For the size of data set we have experience with, there is not a huge time consideration to be made for using classical methods, although we are aware that for larger data sets, time to carry out analysis can be an issue. That said,

developments such as parallel computing will assist in “divide-and-conquer” type approaches to speeding up analysis of very large data sets.

4. Enhanced prediction in clinical biochemistry

The main purpose of each of the papers used for illustration here has been the enhanced prediction of the title of this paper. Typically the prediction is of laboratory diagnosis of a condition, such as anaemia or hepatitis B. In this section, we will describe in detail how the machine learning methods applied to our large data sets resulted in enhanced prediction, and details of what those predictions were will be provided.

Example 1: Enhanced understanding of the relationship between GGT and other components of the routine LFT

Logistic regression modelling of elevated GGT response categories (0, 1) by the full range of predictor variables in the abnormal LFT group showed expected predictors, such as ALP, but also serum cholesterol and urate as significant. The prediction of GGT response category was good at 80%, but model fit was poor (H-L: $p < 0.05$). Log_{10} transformation of all liver enzyme predictor variables (e.g. ALT) improved the model fit ($p > 0.05$). Single decision tree analysis successfully predicted a category 1 GGT response (181 – 2613 U/L) to 90% accuracy, although this result was based on only 20 cases from a total cohort of 347 cases, determined via two ALP branches, and decisions for cholesterol and urate. A category 0 GGT response (40 – 180 U/L) was more robustly predicted from 198 cases at 81%, and only required two ALP branch thresholds for this decision ($\text{ALP} < 172.5 \text{ U/L} + < 89.5 \text{ U/L}$).

Overall prediction accuracy was 74.5%, with the model predicting 464 cases specifically as category 0 (GGT 40 - 180 U/L) and 227 cases as Category 1 (GGT 181 - 2613 U/L). The model, therefore, overestimated category 0 cases ($n = 344$), while underestimating category 1 ($n = 347$).

Example 2: Enhanced laboratory prediction of hepatitis B and C

Single decision trees were extremely effective (98-99% accuracy) for negative HBsAg and negative HepC predictions derived from routine pathology test explanatory variables. This is associated with the large number of patients who tested negative to HBV or HCV. For positive HBsAg and HepC cases, where patient sample size was limited ($n = 212$ and $n = 641$ respectively), the basic single approach was very poor at predicting a positive

HBsAg (< 5.0%), but improved for predicting a positive HepC result – most likely a result of the 3-fold larger sample size. For both positive HBsAg and HepC cases, the maximum accuracy rates of approximately 65% were achieved through applying the “matched single” approach of data pre-processing.

Example 3: Enhanced understanding of relationships between bilirubin and WCC

A significant inverse correlation was detected between bilirubin and WCC ($r = 0.195$, $p < 0.001$, $n = 833$). Significant inverse correlations were also found separately for HBsAg positive cases ($n = 340$, $r = 0.223$, $p < 0.001$), and cases not tested for HBsAg ($n = 402$, $r = 0.155$, $p < 0.005$), but not the HBsAg negative cases ($n = 91$, $r = 0.190$, $p > 0.07$). The same analyses were performed on the total cohort (including cases with elevated bilirubin and/or WCC), although no significant correlations were observed.

By contrast, in the ICU cohort, with all cases included, a significant positive correlation was found between total serum bilirubin and WCC. Whereas after the removal of elevated bilirubin and WCC cases, no significant association was found.

The opposite was found for high-dependency patients requiring intensive care. In this cohort, acute inflammation and multi-system disease processes resulted in a significant positive correlation between bilirubin and WCC across the range of responses. Consistent with this explanation, elevated inflammatory markers were seen in the ICU cohort, with ICU median CRP, ESR and WCC significantly higher compared with the community cohort (30 mg/mL versus 2 mg/mL, 18.5 mm/h versus 6 mm/h, and $11.1 \times 10^9/L$ versus $6.8 \times 10^9/L$ respectively: all $p < 0.001$, Mann-Whitney U test).

Example 4: Enhanced prediction of anaemia

As a predictor of RDW, ferritin was only statistically significant for younger (15 - 50 years) women and not for older women (≥ 55 years), nor men of either age range. Serum iron and TIBC/transferrin were consistent RDW predictors between the second tier tests (but TIBC/transferrin was not significant for ≥ 55 years women). The strong correlation between these iron sufficiency markers and RDW is well known [32]. However the low correlation with TIBC/Transferrin saturation in the older group has not been reported previously. In general, RCF and vitamin B₁₂ were not significant RDW predictors and are not associated with RDW variation in contrast to the literature where RDW has been shown to have some sensitivity to

detecting vitamin B₁₂ deficiency [33], but surprisingly was a strongly significant predictor of RDW in men of 55 years of age or older. This sex difference with the sensitivity of vitamin B₁₂ deficiency warrants further investigation and may reflect the severity of the deficiency in this age group.

Elevated WCC is associated with infection, but also lifestyle including obesity and smoking. In the younger cohorts, RDW is probably associated with iron status, as evidenced by our finding that serum ferritin was an important RDW predictor for young women. The significance of RDW in the older cohorts may be more closely associated with the immune function, disordered haematopoiesis or physiological stress. The RDW relationship with lactate dehydrogenase and bilirubin found during this study supports disordered haematopoiesis.

As previously observed by Dugdale [18], RDW has a significant negative correlation with haemoglobin that provides for the early detection of iron-deficient anaemia (IDA) well in advance of clinically significant decreases in longitudinal haemoglobin measurements, but often is not presented routinely on laboratory reports. In further exploring the utility of RDW for early detection of anaemia, ANCOVA models were conducted on cross-sectional community patient data to assess whether RDW also could give early indications of anaemia aetiology, and perhaps reduce the need to perform additional second tier tests. As the standard second tier test for IDA, serum ferritin was included in all ANCOVA with separate models conducted for routine and second tier test covariates, in data stratified by age and sex. Interestingly, ferritin was not universally significant, with some variations in RDW predictor patterns found for different age and sex. In evaluating RDW as an early marker for anaemia and its cause, several other health conditions that alter RDW need to be considered as well.

5. Conclusion

Clearly, the characteristics of laboratory data, the choice of pre-processing method and tuning of the machine learning algorithms can improve the sensitivity and specificity of diagnostic laboratory decision-making (beyond the sensitivity and specificity of the assays themselves). In a broader conceptual sense, the four studies used to illustrate this paper demonstrated the value of looking at routine laboratory results in total, rather than just those that are traditionally considered to be directly diagnostic for a disease. Subtle changes in results, particularly changes within a reference interval may be useful in understanding pathological processes.

Results from studies like the ones referred to here may not be immediately useful to individual patient diagnosis, but are essential to maximising the capacity for diagnostic pathology tests to accurately and quickly assist laboratory diagnosis through revealing new marker relationship patterns. Such analyses of the subtle interactions between routine and second tier diagnostic markers will also be crucial to deeper understandings of disease processes directly in the human context.

The seven myths exposed in this paper also should encourage pathology laboratories to make use of the data amassing in their workplaces, and conduct similar experiments on appropriate data. These techniques applied to patient data in the future may reduce unnecessary testing, identify potential non-symptomatic disease, and provide insight into new mechanisms of disease.

Glossary

Algorithmic modelling: Modelling that is based on a defined formula

ANCOVA: An analysis of variance between means whereby covariates are controlled for using linear regression.

Artificial neural network: Computer-intensive model used to describe and predict measures of samples

Bivariate Pearson correlation: Measure of correlation between two variables, whereby the coefficient is between -1 (for maximal negative correlation) and +1 (for maximal positive correlation), with a coefficient of 0 indicative of no correlation

Bootstrap: Random sampling of a sample that allows for assignment of accuracy measures to sample characteristic estimates

Broken stick model: A linear model consisting of two or more linear pieces, each connected like a stick that has been broken in several places

Classical statistics: Traditional methods of using models and fit-diagnoses to describe and predict measures of samples

Co-linear variable: A predictor variable that is highly correlated to one or more other predictor variables

Discriminant analysis: A statistical technique used to find a linear combination of features that characterises or separates two or more classes of objects

Factor analysis model: Model used to estimate unobserved variables (factors) in terms of correlated observed variables

False discovery rate: A way of expressing the rate of incorrect rejections of the null hypothesis in multiple comparisons

Fit-diagnose: The process of fitting a model to a dataset and then diagnosing any problems with the fit

Gamma coefficient: A measure of correlation between quantified rankings of different variables or the same variable

Gaussian/non-Gaussian data: Data that is Normally/not Normally distributed

Overfitting: When a model describes random error as opposed to the fundamental relationship

Q-value: The minimum false discovery rate at which a test may be deemed significant

Random forest: Use of many small decision trees in an ensemble to classify members of a population into sub-populations

Recursive partitioning-based decision model: Use of a decision tree to classify members of a population into sub-populations based on independent variables

Support vector machine: A data mining method that identifies the largest gap between features of individuals in order to classify them into sub-populations

Train-test: To train a machine learning algorithm on the majority of a data set and then test the efficacy of the algorithm on the remaining subset of data

Variance inflation factor: Value that quantifies the extent of multicollinearity in a regression analysis

Wavelets: A set of basic functions used to smooth a scatter plot

References

- [1] Breslow, N.E. (2000). Statistics in the life and medical sciences. *Journal of the American Statistical Association* 95, 281 - 282.
- [2] Gu, J., & Taylor, C.R. (2014). Practicing pathology in the era of big data and personalized medicine. *Appl Immunohistochem Mol Morphol*. 2014 Jan;22(1):1-9. doi: 10.1097/PAI.0000000000000022.
- [3] Foster, K.R., Koprowski, R. & Skufca, J.D. (2014). Machine learning medical diagnosis and biomedical engineering research - commentary. *BioMedical Engineering OnLine* 13, 94.
- [4] Crowley, S., Tognarini, D., Desmond, P., Lees, M., Saal, G., 2002. Introduction of lamivudine for the treatment of chronic hepatitis B: expected clinical and economic outcomes based on 4-year clinical trial data. *J Gastroenterol Hepatol* 17, 153-164.
- [5] Shang et al. (2013). Predicting the Presence of Hepatitis B Virus Surface Antigen in Chinese Patients by Pathology Data Mining. *Journal of Medical Virology* 85, 1334 - 1339.
- [6] Kingsford, C., Salzberg, S.L., 2008. What are decision trees? *Nature Biotechnology* 26, 1011-1013.
- [7] Karatzoglou, A., Meyer, D., Hornik, K., 2006. Support Vector Machines in R. *Journal of Statistical Software* 15, 1 - 29.
- [8] Lidbury, B.A., Richardson, A.M., and Badrick, T. (2015). Assessment of machine-learning techniques on large pathology data sets to address assay redundancy in routine liver function test profiles. *Diagnosis* 2, 41 - 51.
- [9] Maindonald, J.H. (n.d.) "Data mining from a statistical perspective". (<http://maths-people.anu.edu.au/~johnm/dm/dmpaper.html>).

- [10] Harrell, F.E. (2001). *Regression Modelling Strategies*. New York: Springer.
- [11] Richardson, A.M. & Lidbury, B.A. (2013). Infection status outcome, machine learning method and virus type interact to affect the optimised prediction of hepatitis virus immunoassay results from routine pathology laboratory assays in unbalanced data. *BMC Bioinformatics* 14, 206 - 213.
- [12] Badrick, T., Richardson, A.M. & Lidbury, B.A. (2016). Response to article: serum total bilirubin concentrations are inversely associated with total white blood cell counts in an adult population. *Annals of Clinical Biochemistry* **53**, 412 – 414.
doi:10.1177/0004563215608654
- [13] Badrick, T., Richardson, A.M., Arnott, A. & Lidbury, B.A. (2015). The early detection of anaemia and aetiology prediction through the modelling of red cell distribution width (RDW) in cross-sectional community patient data. *Diagnosis* **2**, 171 - 179.
- [14] Therneau, T.M. and Atkinson B., 2012. R port by Brian Ripley; rpart: Recursive Partitioning. R package version 3.1-53. <http://CRAN.R-project.org/package=rpart>
- [15] Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D. and Weingessel, A., 2011. e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6.
<http://CRAN.R-project.org/package=e1071>
- [16] Zhou, Z. and Tang, W. (2003). Selective ensemble of decision trees. *Lecture Notes in Artificial Intelligence* 2369, 476 - 483.
- [17] Tsai, W.-N., Wang, Y.-Y., Liang, J.-T., Lin, S.-Y. Sheu, W. H.-H. & Chang, W.-D. (2015). Serum total bilirubin concentrations are inversely associated with total white blood cell counts in an adult population. *Annals of Clinical Biochemistry* 52, 251 - 258.
- [18] Dugdale, A.E. (2011). Diagnosis and management of iron deficiency anaemia: a clinical update. *Med J Aust* 194:429.

- [19] Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*. New York: Springer.
- [20] Johnson, K. (2014). Who's afraid of the big black box? *Significance* 11(3), 35 - 37.
- [21] Breiman, L. (2001). Statistical modelling: the two cultures. *Statistical Science* 16, 199 - 231.
- [22] Harford, T. (2014). Big data: are we making a big mistake? *Significance* 11(5), 14 - 19.
- [23] Bezzina, F. & Saunders, M.N.K. (2014). The pervasiveness and implications of statistical misconceptions among academics with a special interest in business research methods. *Electronic Journal of Business Research Methods* 2, 29-42.
- [24] Gelman, A. (2011). <http://andrewgelman.com/2011/10/25/how-do-you-interpret-standard-errors-from-a-regression-fit-to-the-entire-population/>
- [25] Kettenring, J.R., Koehler, K.J. & McKenzie Jr., J.D. (2015). Challenges and opportunities for statistics in the next 25 years. *The American Statistician* 69, 86 - 90.
- [26] Efron, B. and Tibshirani, R.J. (1994). *An Introduction to the Bootstrap*. Boca Raton FL: Chapman & Hall / CRC Press.
- [27] Walker, J.W. (2008). *A primer on wavelets and their scientific applications*, 2nd edition. Boca Raton FL: Chapman & Hall. CRC Press.
- [28] Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* 57, 289 - 300.
- [29] Storey, J.D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, 64:479 - 498.
- [30] Chambers (1993). Greater or lesser statistics: a choice for future research. *Statistics and Computing* 3, 182 - 184.

- [31] Horton, N.J. (2015). Challenges and opportunities for statistics and statistical education: looking back, looking forward. *The American Statistician* 69, 138 - 145.

- [32] Wians FH, Urban JE, Keffer JH, Kroft SH. Discriminating between iron deficiency anaemia and anaemia of chronic disease using traditional indices of iron status versus transferrin receptor concentration. *AJCP* 2001; 115: 112-118.

- [33] Thompson WG, Cassino C, Meola T, Berman R, Lipkin M, Freedman M. Hypersegmented neutrophils and vitamin B12 deficiency. *Acta Haematol* 1989; 81: 186-191