Clinical Biochemistry xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Clinical Biochemistry

journal homepage: www.elsevier.com/locate/clinbiochem

Review

Clinical chemistry in higher dimensions: Machine-learning and enhanced prediction from routine clinical chemistry data

Alice Richardson [a], Ben M. Signor [b], Brett A. Lidbury [b], Tony Badrick [b,c,*]

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

ARTICLE INFO

Article history:
Received 26 May 2016
Received in revised form 19 July 2016
Accepted 20 July 2016
Available online xxxx

Keywords:
Anaemia
Big data
Bilirubin
Biomarkers
Hepatitis
Liver function tests
Misconceptions
Predictive modelling
Statistics

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.

Contents

* Corresponding author.
E-mail address: tony.badrick@rcpaqap.com.au (T. Badrick).

http://dx.doi.org/10.1016/j.clinbiochem.2016.07.013
0009-9120/

Please cite this article as: A. Richardson, et al., Clinical chemistry in higher dimensions: Machine-learning and enhanced prediction from routine clinical chemistry data, Clin Biochem (2016), http://dx.doi.org/10.1016/j.clinbiochem.2016.07.013

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, >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

Please cite this article as: A. Richardson, et al., Clinical chemistry in higher dimensions: Machine-learning and enhanced prediction from routine clinical chemistry data, Clin Biochem (2016), http://dx.doi.org/10.1016/j.clinbiochem.2016.07.013