MATHEMATICS
MATTERS

# Big Data in Healthcare

Healthcare is increasingly reliant on data both to evidence practice and to personalise health management, not just with medicinal drugs but also through better prevention and early diagnosis. It is the analytical tools of mathematics that are helping tease useful information from this wealth of medical information.

There was a time when medical records sat gathering dust in dark rooms and were only pulled out when the doctor required them in relation to an individual patient's medical needs. Today medical records are often electronic and sometimes in a standard format, making them easier to analyse. In addition, routinely collected large-scale records such as Hospital Episode Statistics scale up to big data.  Further, the relatively new science of bioinformatics generates output at an accelerating rate through high throughput and other studies. The medical field is now generating information every day in unprecedented amounts. Some estimates suggest that the amount of healthcare data currently in existence worldwide is roughly thirty times larger than every word spoken in human history. We truly live in the 'Big Data' era.

> "But how do you make sense of the burgeoning repositories of health data? The answer, of course, is mathematics."

The lure of this wealth of information isn't hard to see – it provides strength in numbers. Clinicians and researchers look to find which treatments worked previously in certain situations, not just under the controlled conditions of major trials but also with co-morbidities and all of the other vagaries of routine clinical practice. They will be able to see whether a particular treatment caused complications if the patient also had a secondary condition. They will also be able to see earlier signs that indicate the threat of disease progression. Such information can save lives. It can also protect increasingly scarce resources by moving steadily towards better prevention. This could also provide insights into new treatments and new medicines, to better identify and mitigate adverse effects, however rare, and will also open new opportunities for the re-use of existing medicines to treat other conditions.
But how do you make sense of these burgeoning repositories of health data? The answer, of course, is mathematics. There are many ways to analyse 'Big Data' mathematically. One is known as topological data analysis (TDA). In healthcare, it is often the case that data is already split into groups. It could be a group with a particular disease, or a group that has been treated with a particular drug. The mathematical branch of topology is useful because it looks to find connections between groups. Conventionally, groups are depicted as circles called "nodes" and the connections between them drawn as lines, called "edges". Suddenly, rather than having a flood of numbers, this method generates a 'map' of the data which begins to show its shape and structure. By representing the data geometrically in this way, established techniques for analysing shapes can be used to extract patterns from the data. In 2011, a version of this technique was used to identify

Institute of
mathematics
& its applications

a sub-group of breast cancer patients who expressed high levels of a particular protein which coincided with a 100% survival rate. That's clearly a useful insight.

But what if the data aren't in clear groups to start with? Mathematics can still help, through a technique known as clustering analysis. The goal is to find the optimal set of groups where there is the most similarity within the group and the least similarity between the groups. In a 2013 study, researchers applied clustering analysis to data on almost 1500 cases of fibromyalgia – a condition that sees sufferers experience pain all over their body. Researchers deploying clustering analysis were able to collect these patients into three sub-groups based on the severity of their condition. The authors of the study suggest these three groups could represent different forms of the disease. Future research will look to probe this idea further, but the original insight was provided mathematically. Medical big data is still a field in its infancy and its potential is huge, but it isn't without its challenges. There are, of course, the issues surrounding the confidentiality of the information. There is also the problem of how to store and manage the sheer amount of data. Storing the complete genetic data of just one person would take up a staggering 150 trillion gigabytes. Yet those running our health service, along with the medical charities who support medical funding and innovation, are crying out for modern tools to grapple with the challenges we face. Twenty first century technology for twenty first century healthcare. Only mathematics will make that possible.

# TECHNICAL SUPPLEMENT

## Topological data analysis (TDA)

Topology has its roots with the work of 18th century Swiss mathematician Leonhard Euler. He famously used topology to prove that it was impossible to cross the seven bridges on Königsberg without retracing your steps.

Using topology to extract information about patterns and shapes within large data sets is based on three key assumptions. The first is *co-ordinate invariance*: that the analysis is independent of the co-ordinate system used. This can be key in healthcare where different data sets can come from different platforms. The second is *deformation invariance* – that a shape remains the same even when stretched or deformed. A good example of this is that letters of the alphabet are still recognisable by their shape, even when written in different fonts. This makes topology a good tool for big data analysis because it is less sensitive to noise and can pick out groups even if on the face of it the members of that group look like they don't belong. The third assumption is *compressed representations*; that a shape can often be represented by a simplification. Take a circle with its infinite number of sides. An icosagon with its finite 20 sides is a very good approximation but often easier to work with.

## Clustering analysis

The term clustering analysis was first used as far back as 1939. There are many different techniques for organising data into groups in this way. They include Joining (Tree Clustering), Two-Way Joining (Block Clustering) and k-means clustering. There are also different ways of discerning how far "apart" the groups (or members of the group) are. These can include Euclidean distance, Squared Euclidean distance, Manhattan distance, Chebychev distance, power distance and percentage disagreement.

## Expert

Professor Paulo Lisboa, Liverpool John Moores University

## References

Nicolau, M. et al. (2011). Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. Proceedings of the National Academy of Sciences. 108 (17), p7265-7270.

Lum, P.Y. et al. (2012). Extracting insights from the shape of complex data using topology. Scientific Reports. 3 (1236)

Docampo E. et al. (2013) Cluster Analysis of Clinical Data Identifies Fibromyalgia Subgroups. PLoS ONE 8(9): e74873

*The IMA would like to thank Paulo Lisboa, Liverpool John Moores University, for his help in the preparation of this document.*