Big Data in healthcare is a complex, rapidly evolving and exciting area. It is an important source of Real World Evidence (RWE) - data which has been collected outside of a conventional clinical trial.

In Part III of the RWE Series we distil this wide-ranging and sometimes intimidating topic down to its key features to understand how we can take advantage of this data and maximise its potential.

We pay special attention to the visualisation methods for interpreting and presenting Big Data.

This is the third part of a regular series available at svmpharma.com
The capabilities and potential of Big Data have become a major focus of attention in recent years across many industries. There is no shortage of astounding facts about Big Data, for example, 90% of the data in the world today has been created in the last 2 years alone.\(^1\)\(^2\) Furthermore, there have been frequent predictions that Big Data will solve persistent issues or lead to previously unimaginable cost reductions.\(^4\)

However, it can be difficult to relate to and contextualise these impressive facts and possibilities. First we need to gain a better understanding of the core principles of healthcare analytics and Big Data.

Big Data has been defined as ‘high volume, high velocity, and high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimisation’. This influential ‘3Vs’ model of Big Data was first envisioned in 2001 and has been referred to in the majority of articles since that time.\(^5\)\(^6\) More recently a fourth ‘V’ has been incorporated which emphasises the need for the data to be accurate, this is known as data veracity.\(^7\)\(^8\)

Big Data & Healthcare Analytics

This model provides an excellent starting point for exploring the topic, and as you will see, healthcare data analytics provides the most powerful examples of the characteristics of Big Data.

In this article we propose for the first time, a fifth ‘V’: visualisation.

We will highlight the visualisation tools and techniques which are needed to drive meaningful outcomes from Big Data.
Volume
The sheer quantity of data that is captured, stored and processed is the eponymous characteristic of Big Data. Compared to other industries, the healthcare services have always generated vast quantities of data. This is through meticulous record keeping, compliance/ regulatory requirements, and comprehensive patient care notes. As we move towards universal electronic health records in primary care and increasing digitisation of hospital data, we now have a remarkable volume of data. Recent technological advancements in data storage capacity is now prepared for this, whilst networking capabilities can handle the movement of large quantities of data. In addition to clinical data, there has been a notable increase in the collection of consumer driven lifestyle data which can provide valuable insights into general health and wellbeing.

Velocity
New health data is generated and captured continuously and at a rapid pace. There is a constant flow of new data. The health service is moving away from traditional static data such as x-rays films and scripts to more dynamic systems involving multi-dimensional and regular measurements. Real-time data monitoring is associated with critical care units and the operating theatre but can increasingly be applied to care at home and the community. Technology allows for immediate data analysis in real time from any location and any number of patients. The general consensus is that real-time analytics of high volume data has the potential to revolutionise healthcare.

Variety
Healthcare data exists in many forms including written notes and prescriptions, medical imaging, laboratory, pharmacy, insurance and administrative data in addition to sensor data. Increasingly health information can be found within dynamic web content including social media. Big Data can be coded and stored in a structured form, for example, in the large datasets, Hospital Episode Statistics (HES) and the Clinical Practice Research Datalink (CPRD). However a significant amount of data exists in unstructured or semi-structured formats and the processing, storage and analysis of data is adapting to accommodate this.

Veracity
Veracity of data is its accuracy and conformity to facts. Veracity is an essential part of any discussion on data and is particularly important in healthcare, where the implications of inaccurate data can be severe. To attain accurate analyses in Big Data, there needs to be a scaling up in granularity and performance of the architectures, platforms, algorithms, methodologies and tools. Big Data analysis has been compared to financial data in this regard. As Big Data in healthcare is measuring real-world outcomes, it can offer a more accurate representation of real world practice than controlled trials.

Visualisation
Due to its volume, variety and velocity, Big Data analysis and presentation can be difficult. Traditional written, tabular and graphical formats can be inadequate in conveying the findings, and key details can be lost in summarising and aggregating the data. There a number of visualisation tools and solutions which have been proven to be valuable in identifying relationships and conveying information. The two most commonly used for Big Data in healthcare are Cytoscape and Gephi. Cytoscape initially began in bioinformatics (gene processing and molecular networks) whilst Gephi has often been used for consumer and social research. Both of these tools can be adapted to enhance healthcare data analytics as our examples will show. Network graphs consists of nodes (a set of objects representing entities in a dataset) and edges (connections between nodes that provide visual cues as to the degree of connectedness).
Visualisation in Cystic Fibrosis

SVMPharma carried out a data visualisation exercise using Gephi to show the multi-disciplinary nature of Cystic Fibrosis (CF) by highlighting its co-morbidities and associated health problems.

15,000 Cystic Fibrosis patients who visited hospital from 2010-2014 were selected by their ICD-10 code (diagnosis code) from the Hospital Episodes Statistics (HES) dataset.

This provided 3,800 ICD-10 codes and over 20,000 different connections between them. The data was entered into Gephi, with each ICD-10 code represented by a node, and each different connection between them via a line. The radial layout was used in Gephi to create the image on the right (Image 5) which includes the entirety of the data.

To take a closer look at the data, the top 100 different connections leading from the CF nodes were identified. This is represented below (Image 8) with both the node size and connecting line representing the strength and volume of the connection, and the nodes positioned and colour-coded according to disease group.

The final visual conveys at a glance, the main co-morbidities of CF patients and associated health issues. This shows the strength of the relationships between CF and its secondary complications and directions the disease is likely to progress.

8. A Closer Look

© 2015 SVMPharma Ltd. All rights reserved.
The Technology

To cope with Big Data, a number of hardware and software solutions are used and these have replaced the traditional Business Intelligence (BI) tools. The first step involves massively parallel computing and the processing of large datasets across clusters using a Hadoop based program. Specialist tools are available for rapid processing and querying including Spark for structured and semi-structured data, whilst unstructured data can be stored in the document database software MongoDB. Saiku allows users to create, visualize and analyse information in graphical & pivot table formats. These wholesale changes in data storage and processing architecture has been crucial and sets the platform for future innovation and expansion.

Big Data analytics in healthcare is almost ready to fulfil its potential: this is due to the rapid digitisation of information, the adaptations in technological architecture and the increasing capabilities for real-time analysis. However, to maximise its potential we need to fully understand the key characteristics of Big Data: volume, variety and velocity. A consideration of the veracity of the data and visualisation techniques make up our ‘5Vs’ of Big Data in healthcare and provide a useful framework to explore the key issues.