Big data, big knowledge: big data for personalised healthcare

Marco Viceconti[1,2] , Peter Hunter[1,3], Keith McCormack[1,2], Adriano Henney[1,4], Stig W. Omholt[1,5], Norbert Graf[1,6], Edwin Morley-Fletcher[1,7], Liesbet Geris[1,8], and Rod Hose[2]

[1]VPH Institute for integrative biomedical research

[2] Insigneo institute for *in silico* Medicine, University of Sheffield

[3] Auckland Bioengineering Institute, University of Auckland, New Zealand

[4] Virtual Liver Network, Germany

[5] Norwegian University of Science and Technology, Norway

[6] Saarland University, Germany

[7] Lynkeus, Italy

[8] University of Liège, Belgium

Executive summary

The idea that the purely phenomenological knowledge that we can extract by analysing large amounts of data can be useful in healthcare seems to contradict the desire of VPH researchers to build detailed mechanistic models for individual patients. But in practice no model is ever entirely phenomenological or entirely mechanistic. We propose in this position paper that big data analytics can be successfully combined with VPH technologies to produce robust and effective *in silico* medicine solutions. In order to do this, big data technologies must be further developed to cope with some specific requirements that emerge from this application. Such requirements are: working with sensitive data; analytics of complex and heterogeneous data spaces, including non-textual information; distributed data management under security and performance constraints; specialised analytics to integrate bioinformatics and systems biology information with clinical observations at tissue, organ and organisms scales; and specialised analytics to define the "physiological envelope" during the daily life of each patient. These domain-specific requirements suggest a need for targeted funding, in which big data technologies for *in silico* medicine becomes the research priority.

Introduction

The birth of big data, as a concept if not as a term, is usually associated with a META Group report by Doug Laney entitled "3D Data Management: Controlling Data Volume, Velocity, and Variety" published in 2001. For a long time the development of big data technologies was inspired by business trends analysis and by big science (such as the Large Hadron Collider at CERN). But when in 2010 Google Flu, simply by analysing Google queries, predicted flu-like illness rates as accurately as the CDC's enormously complex and expensive monitoring network, some media started to shout that all problems of modern healthcare could be solved by big data.

In 2005, the term *Virtual Physiological Human* (VPH) was introduced to indicate "a framework of methods and technologies that, once established, will make possible the collaborative investigation of the human body as a single complex system". The idea was quite simple:

- To reduce the complexity of living organisms, we *decompose* them into parts (cells, tissues, organs, organ systems) and investigate one part in isolation from the others. This approach has produced, for example, the medical specialties, where the nephrologist looks only at your kidneys, and the dermatologist only at your skin; this makes it very difficult to cope with multi-organ or systemic diseases, to treat multiple diseases (so common in the ageing population), and in general to unravel systemic emergence due to genotype-phenotype interactions.
- But if we can *recompose* with computer models all the data and all the knowledge we have obtained about each part, we can use simulations to investigate how these parts interact with one another, across space and time and across organ systems.

Though this may be conceptually simple, the VPH vision contains a tremendous challenge, namely the development of mathematical models capable of accurately predicting what will happen to a biological system. To tackle this huge challenge, multifaceted research is necessary: around medical imaging and sensing technologies (to produce quantitative data about the patient's anatomy and physiology), data processing to extract from such data information that in some cases is not immediately available, biomedical modelling to capture the available knowledge into predictive simulations, and computational science and engineering to run huge *hypermodels* (orchestrations of multiple models) under the operational conditions imposed by clinical usage.

But the real challenge is the production of that mechanistic knowledge, quantitative, and defined over space, time and across multiple space-time scales, capable of being predictive with sufficient accuracy. After ten years of research this has produced a complex impact scenario in which a number of target applications, where such knowledge was already available, are now being tested clinically, while others are still attempting to produce such knowledge.

So it is perhaps not surprising that recently, especially in the area of personalised healthcare (so promising but so challenging) some people have started to advocate the use of big data technologies as an alternative approach, in order to reduce the complexity that developing a reliable, quantitative mechanistic knowledge involves.

This trend is fascinating from an epistemological point of view. The VPH was born around the need to overcome the limitations of a biology founded on the collection of a huge amount of observational data, frequently affected by considerable noise, and boxed into a radical reductionism that prevented most researchers from looking at anything bigger than a single cell. Suggesting that we revert to a phenomenological approach where a predictive model is supposed to emerge not from mechanistic theories but by only doing high-dimensional big data analysis,, may be perceived by some as a step toward that empiricism the VPH was born to overcome.

In the following we will explain why the use of big data methods and technologies could actually empower and strengthen current VPH approaches, increasing considerably its chances of clinical impact in many "difficult" targets. But in order for that to happen, it is important that big data researchers are aware that when used in the context of computational biomedicine, big data methods need to cope with a number of hurdles that are specific to the domain. Only by developing a research agenda for big data in computational biomedicine can we hope to achieve this ambitious goal.

Doctors and Engineers: joined at the hip

As engineers who have worked for many years in research hospitals we recognise that clinical and engineering researchers share a similar mind-set. Both in traditional engineering and in medicine, the research domain is defined in terms of problem solving, not of knowledge discovery. The motto common to both disciplines is "whatever works".

But there is a fundamental difference: engineers usually deal with problems related to phenomena on which there is a large body of reliable knowledge from physics and chemistry. When a good reliable mechanistic theory is not available we resort to empirical models, as far as they can solve the problem at hand. But when we do this we are left with a sense of fragility and mistrust, and we try to replace them as soon as possible with theory-based mechanistic models, which are both predictive and explanatory.

Medical researchers deal with problems for which there is a much less well established body of knowledge; in addition, this knowledge is frequently qualitative or semi-quantitative, and obtained in highly controlled experiments quite removed from clinical reality, in order to tame the complexity involved. Thus, not surprisingly, many clinical researchers consider mechanistic models "too simple to be trusted", and in general the whole idea of a mechanistic model is looked upon with suspicion.

But in the end "whatever works" remains the basic principle. In some VPH clinical target areas where we can prove convincingly that our mechanistic models can provide more accurate predictions than the epidemiology-based phenomenological models, the penetration into clinical practice is happening. On the other hand, when our mechanistic knowledge is insufficient, the predictive accuracy of our models is poor, and models based on empirical/statistical evidences are still preferred.

The true problem behind this story is the competition between two methods of modelling nature that are both effective in certain cases. Big data can help computational biomedicine to transform this competition into collaboration, significantly increasing the acceptance of VPH technologies in clinical practice.

Big data VPH: an example in osteoporosis

In order to illustrate this concept we will use as a guiding example the problem of predicting the risk of bone fracture in a woman affected by osteoporosis, a pathological reduction of her bone mineralised mass. The goal is to develop predictors that indicate whether the patient is likely to fracture over a given time (typically in the following ten years). If a fracture actually occurs in that period, this is the true value used to decide if the outcome prediction was right or wrong.

Because the primary manifestation of the disease is quite simple (reduction of the mineral density of the bone tissue) not surprisingly researchers found that when such mineral density could be accurately measured in the regions where the most disabling fractures occurred (hip and spine), such measurement was a predictor of the risk of fracture. In controlled clinical trials, where patients are recruited to exclude all confounding factors, the Bone Mineral Density (BMD) can predict hip fractures with an accuracy of 70-75%. Unfortunately, when the predictor is used on randomised populations, the accuracy drops to 60-65%. Given that fracture is a binary event, tossing a coin would give us 50%, so this is considered not good enough.

Epidemiologists run huge international, multicentre clinical trials where the fracture events are related to a number of observables; the data are then fitted with statistical models that provide phenomenological models capable of predicting the likelihood of fracture; the most famous, called FRAX, was developed by John Kanis at the University of Sheffield, UK, and is considered by the World Health Organisation the reference tool to predict risk of fractures. The predictive accuracy of FRAX is comparable to that of BMD, but it seems more robust for randomised cohorts.

In the VPHOP project, one of the flagship VPH projects funded by the Seventh Framework Program of the European Commission, we took a different approach: we developed a multiscale patient-specific model informed by medical imaging and wearable sensors, and used this model to predict the actual risk of fracture of the hip and at the spine, essentially simulating 10 years of the patient's daily life. The results of the first clinical assessment, published only few weeks ago, suggest that the VPHOP approach could increase the predictive accuracy to 80-85%; significant but not dramatic; no information is available yet on the accuracy with fully randomised cohorts, although we expect the mechanistic model to be less sensitive to biases.

Our goal in VPHOP was to replace FRAX; in doing this we took a considerable risk, common to most VPH projects, where a radically new and complex technology aims to replace an established standard of care. The difficulty arises from the need to step into the unknown, with the outcome remaining unpredictable until the work is complete. In our opinion big data technologies could change this high-risk scenario.

From data to theory: a continuum

Modern big data technologies make it possible in a short time to analyse a large collection of data from thousands of patients, identify clusters and correlations, and develop predictive models using statistical or machine-learning modelling techniques. In this new context it would be feasible to take all the data collected in all past epidemiology studies - for example, those used to develop FRAX - and continue to enrich them with new studies where not only new patients are added, but different types of information are collected. (This would create an incomplete data matrix, but methods are available to deal with this).

Another mechanism that very high-throughput technologies make viable for exploration is the normalisation of digital medical images to conventional space-time reference systems, using elastic registration methods, followed by the treatment of the quantities expressed by each voxel value in the image as independent data quanta. The voxel values of the scan then become another medical dataset, potentially to be correlated with average blood pressure, body weight, age, or any other clinical information.

Using statistical modelling or machine learning techniques we may obtain good predictors valid for the range of the data sets analysed; if a database contains outcome observables for a sub-set of patients, we will be able to compute automatically the accuracy of such a predictor. Typically the result of this process would be a potential clinical tool with known accuracy; in some cases the result would provide a predictive accuracy sufficient for clinical purposes, in others a higher accuracy might be desirable.

In some cases there is need for an explanatory theory, which answers the "how" question, and which may be used in a wider context than that a statistical model normally is. As a second step, one could use the correlation identified by the empirical modelling to elaborate possible mechanistic theories. Given that the available mechanistic knowledge is quite incomplete, in many

cases we will be able to express a mathematical model only for a part of the process to be modelled; various "grey-box" modelling methods have been developed in the last few years that allow one to combine partial mechanistic knowledge with phenomenological modelling.

The last step is where physiology, biochemistry, biomechanics, and biophysics mechanistic models are used. These models contain a large amount of validated knowledge, and require only a relatively small amount of data to be properly identified.

In many cases these mechanistic models are extremely expensive in terms of computational cost; therefore input-output sets of these models may also be stored in a data repository in order to identify reduced-order models (also referred as 'surrogate' models and 'meta-models') that accurately replace a computationally expensive model with a a cheaper/faster simulation. Experimental design methods are used to choose the input and output parameters or variables with which to run the mechanistic model in order to generate the meta-model's state space description of the input-output relations – which is often replaced with a piecewise partial least-squares regression (PLSR) approximation. Another approach is to use Nonlinear AutoRegressive Moving Average model with eXogenous inputs (NARMAX) in the framework of non-linear systems identification.

It is interesting to note that no real model is ever fully "white-box". In all cases, some phenomenological modelling is required to define the interaction of the portion of reality under investigation with the rest of the universe. If we accept that a model describes a process at a certain characteristic space-time scale, everything that happens at any scale larger or smaller than that must also be accounted for phenomenologically. Thus, it is possible to imagine a complex process being modelled as an orchestration of sub-models, each predicting a part of the process (for example at different scales), and we can expect that, while initially all sub-models will be phenomenological, more and more will progressively include some mechanistic knowledge.

The idea of a progressive increase of the explanatory content of a hypermodel is not fundamentally new, other domains of science already pursue the approach described here; but in the context of computational biomedicine this is an approach used only incidentally, and not as a systematic strategy for the progressive refinement of clinical predictors.

Big data for computational biomedicine: requirements

In a complex scenario such as the one described above, are the currently available technologies sufficient to cope with this application context?

The brief answer is no. A number of shortcomings that need to be addressed before big data technologies can be effectively and extensively used in computational biomedicine. Here we list five of the most important.

Confidential data

The majority of big data applications deal with data that do not refer to an individual person. This does not exclude the possibility that their aggregated information content might not be socially sensitive, but very rarely is it possible to reconnect such content to the identity of an individual.

In the cases where sensitive data are involved, it is usually possible to collect and analyse the data at a single location; so this becomes a problem of computer security; within the secure box, the treatment of the data is identical to that of non-sensitive data.

Healthcare poses some peculiar problems in this area. First, all medical data are highly sensitive, and in most developed countries are legally considered owned by the patient, and the healthcare provider is required to respect patient confidentiality. The European parliament is currently involved in a complex debate about data protection legislation, where the need for individual confidentiality can be in conflict with the needs of society.

Secondly, in order to be useful for diagnosis, prognosis or treatment planning purposes the data analytics results must in most cases be re-linked to the identity of the patient. This implies that the clinical data cannot be fully and irreversibly anonymised before leaving the hospital, but requires complex pseudo-anonymisation procedures. Normally the clinical data are pseudo-anonymised so as to ensure a certain k-anonymity, which is considered legally and ethically acceptable. But when the data are, as part of big data mash-ups, relinked to other data, for example from social networks or other public sources, there is a risk that the k-anonymity of the mash-up can be drastically reduced. Specific algorithms need to be developed that prevent such data aggregation when the k-anonymity could drop below the required level.

Big data: big size or big complexity?

Consider two data collections:

- In one case we have 500 TB of log data from a popular web site: a huge list of text strings, typically encoding 7-10 pieces of information for transaction.
- In the other, we have a full VPHOP dataset for 100 patients, a total of 1TB; for each patient we have 122 textual information items that encode the clinical data, three medical imaging datasets of different types, 100 signal files from wearable sensors, a neuromuscular dynamics output database, an organ-level finite element model with the results, and a microFE model with the predictions of bone remodelling over 10 years. This is a typical VPH data folder; some applications require even more complex data spaces.

Which one of these two data collections should be considered big data? We suggest that the idea held by some funding agencies, that the only worthwhile applications are those targeting data collections over a certain size, trivialises the problem of big data analytics. While the legacy role of big data analysis is the examination of large amounts of scarcely complex data, the future lies in the analysis of complex data, eventually even in smaller amounts.

Integrating bioinformatics, systems biology, and phenomics data

Genomics and post-genomics technologies produce very large amounts of raw data about the complex biochemical processes that regulate each living organism; nowadays a single deep-sequencing dataset can exceed 1TB. More recently we started to see the generation of "deep phenotyping" data, where biochemical, imaging, and sensing technologies are used to quantify complex phenotypical traits and link them to the genetic information. These data are processed with specialised big data analytics techniques, which come from bioinformatics, but recently there is growing interest in building mechanistic models of how the many species present inside a cell interact along complex biochemical pathways. Because of the complexity and the redundancy

involved, linking this very large body of mechanistic knowledge to the higher-order cell-cell and cell-tissue interactions remains very difficult, primarily for the data analytics problems it involves. But when this is possible, genomics research results finally link to clinically relevant pathological signs, observed at tissue, organ, and organism scales, opening the door to a true systems medicine.

Where are the data?

In big data research, the data are usually stored and organised in order to maximise the efficiency of the data analytics process. In the scenario described here however, it is possible that parts of the simulation workflow require special hardware, or can be run only on certain computers because of licence limitations. Thus one ends up trading the needs of the data analytics part with those of the VPH simulation part, always ending up with a sub-optimal solution.

In such complex simulation scenarios, data management becomes part of the simulation process; clever methods must be developed to replicate/store certain portions of the data within organisations and at locations that maximise the performance of the overall simulation.

The physiological envelope and the predictive avatar

In the last decade there has been a great deal of interest in the generation and analysis of patient-specific models. Enormous progress has been made in the integration of image processing and engineering analysis, with many applications in healthcare across the spectrum from orthopaedics to cardiovascular systems and often multiscale models of disease processes, including cancer, are included in these analyses. Very efficient methods, and associated workflows, have been developed that support the generation of patient-specific anatomical models based on exquisite three and four-dimensional medical images. The major challenge now is to use these models to predict acute and longer-term physiological and biological changes that will occur under the progression of disease and under candidate interventions, whether pharmacological or surgical. There is a wealth of data in the clinical record that could support this, but its transformation into relevant information is enormously difficult.

All engineering models of human organ systems, whether focused on structural or fluid flow applications, require not only the geometry (the anatomy) but also constitutive equations and boundary conditions. The biomedical engineering community is only beginning to learn how to perform truly personalised analysis, in which these parameters are all based on individual physiology. There are many challenges around the interpretation of the data that is collected in the course of routine clinical investigation, or indeed assembled in the Electronic Health Record or Personal Health Records. Is it possible to predict the threat or challenge conditions (e.g. limits of blood pressure, flow waveforms, joint loads), and their frequency or duration, from the data that is collected? How can the physiological envelope of the individual be described and characterised? How many analyses need to be done to characterise the effect of the physiological envelope on the progression of disease or on the effectiveness of treatment? How are these analyses best formulated and executed computationally? How is information on disease interpreted in terms of physiology? As an example, how (quantitatively) should we adapt a patient-specific multiscale model of coronary artery disease to reflect the likelihood that a diabetic patient has impaired coronary microcirculation? At a more generic level, how can the priors (in terms of physical relationships) that are available from engineering analysis be integrated into machine learning operations in the context of digital healthcare, or alternatively how can machine learning be used to characterise

the physiological envelope to support meaningful diagnostic and prognostic patient-specific analyses? For a simple example, consider material properties: arteries stiffen as an individual ages, but diseases such as *moyamoya* syndrome can also dramatically affect arterial stiffness; how should models be modified to take into account such incidental data entries in the patient record?

Conclusions

Although sometimes overhyped, big data technologies do have great potential in the domain of computational biomedicine, but their development should take place in combination with other modelling strategies, and not in competition. This will minimise the risk of research investments, and will ensure a constant improvement of *in silico* medicine, favouring its clinical adoption.

For many years the high-performance computing world was afflicted by a one-size-fits-all mentality that prevented many research domains from fully exploiting the potential of these technologies; more recently the promotion of centres of excellence, etc., targeting specific application domains, demonstrates that the original strategy was a mistake, and that technological research must be conducted at least in part in the context of each application domain.

It is very important that the big data research community does not repeat the same mistake. While there is clearly an important research space examining the fundamental methods and technologies for big data analytics, it is vital to acknowledge that it is also necessary to fund domain-targeted research that allows specialised solutions to be developed for specific applications. Healthcare in general, and computational biomedicine in particular, seems a natural candidate for this.