Big Data for cardiology: novel discovery?

Viktor Mayer-Schönberger*

Oxford Internet Institute, University of Oxford, Oxford, UK

Published on behalf of the European Society of Cardiology. All rights reserved.

Aim

Big Data promises to change cardiology through a massive increase in the data gathered and analysed; but its impact goes beyond improving incrementally existing methods.

Methods and results

The potential of comprehensive data sets for scientific discovery is examined, and its impact on the scientific method generally and cardiology in particular is posited, together with likely consequences for research and practice.

Conclusion

Big Data in cardiology changes how new insights are being discovered. For it to flourish, significant modifications in the methods, structures, and institutions of the profession are necessary.

Keywords

Big Data • Scientific method • Discovery • Causality • Correlation • Empirical analysis

Every empirical science benefits from improvements in the ability to collect, manage, and analyse data. These improvements accelerate the rate of discovery as more data can be examined in less time. Thus, researchers have welcomed recent advancements in digital technology as continuations of a long series of technical changes that facilitate the existing process of discovery. However, an emerging alternative viewpoint is arguing that we are witnessing a paradigmatic change in how sciences operate. The proposition is that Big Data is reshaping the scientific method, and by extension scientific fields, especially those that are data rich such as cardiology.

There is no generally accepted definition of Big Data. Early attempts to capture Big Data focused on the so-called ‘three Vs’ of data—volume, velocity, and variety; many more data points are being used; collection and analysis are faster, and different data sources are being combined and analysed together.1 The three Vs encapsulate obvious features of an accelerated data handling technology. But by emphasizing the obvious, the three Vs also obfuscate a deeper and potentially more fundamental shift of the role data plays in our fields. It is this shift that is the topic of this review article.

I first take a look at the differences between the conventional way to work with data and a big data approach. I point to the technical developments that have made such an approach substantially more feasible in terms of effort and cost. I then look at how and to what extent Big Data may change the scientific method, identifying four distinct ways this may happen. This change in the scientific method is then applied to the field of medicine in general and cardiology in particular, suggesting specific consequences and likely trajectories of the field’s development.

In the context of this article, Big Data is thus less about a new technology of handling data, and more about a resulting shift in the method of employing data for sense making. At this more abstract level, Big Data is the ability to gain novel insights into reality through the collection and analysis of a vastly more comprehensive set of data points than was feasible before.2,3

Using data to understand reality—from small data to Big Data

Humans have always used data to make sense of the world, with a seemingly natural connection between observation and discovery. But they also understood that collecting, storing, and analysing data were often hard, time consuming, and costly. The difficulties in handling data prompted humans to use as little data as possible. The very methods and techniques, the structures, and institutions of discovery were designed so that the most insights could be squeezed out of the least amount of data. It was a rational response to the high cost of working with data.

Random sampling is a case in point. Widely adopted less than a century ago, random sampling helps us make statements about an entire population by collecting data from just a small subset of it.4 This preserves resources, as only a limited amount of data has to be gathered and examined. It thus enabled an empirical, data-driven approach in a wide variety of areas. But fundamentally random sampling is a shortcut, and results derived from it not only lack detail (because for drilling deeper into details the sample is too small),

---

*Corresponding author. Tel: +44 1865 287232, Email: viktor.ms@oii.ox.ac.uk

Published on behalf of the European Society of Cardiology. All rights reserved. © The Author 2015. For permissions please email: journals.permissions@oup.com.
but also data collection has to be undertaken quite carefully to prevent biases in the data. It is a great second-best solution designed for a world in which capturing and analysing data comprehensively is too costly.

Over the last decades, however, the economics of data collection, storage, and analysis have changed dramatically. Data processing performance has (as envisioned by Gordon Moore decades ago) roughly doubled every 2 years at constant cost (or cost halved at constant performance). As a result, a single Google search today takes as much computing as ‘all the computing done for the entire Apollo [moon landing] program.’ Data storage cost and performance have more or less tracked processing improvements. For instance, digital storage density increased by almost a billion fold since the 1950s, bringing storage cost of a megabyte of data down from over 100,000 US dollars to little more than a thousandth of a cent for the same amount today. At the same time, data management and analytics tools were developed that greatly improved the speed and ease of data storage, retrieval, and statistical analysis. In addition, advancement in sensor technology meant that more aspects of reality could reliably be translated into data in small packages, with little energy and at low cost. Taken together, today utilizing data is far less costly than ever before.

As the resource constraints have eroded, methods predicated on the limited availability of data are rethought. To the extent that data can be collected and analysed comprehensively and at scale, there is no fundamental need any more to work only with small samples.

An important point here is not only that our methods will change, but also that data quantity is less crucial in absolute terms than relative to the phenomenon that is studied. Therefore, Big Data does not necessarily require many millions or even billions of individual measurements to be gathered and examined. Rather, Big Data-based research aims to consider no longer just a small subset of the relevant data, but close to if not all of it.

**Reshaping the scientific method**

Today, the prevalent scientific method is to deductively ideate from one or more theories a concrete hypothesis, which then is tested against data. The results either falsify the hypothesis, which is thereby discarded, or confirm it. Since at least Karl Popper, empirical confirmation, however, is not proof of a hypothesis’ correctness. In fact, according to Popper, there can never be a positive proof of a hypothesis. Even with mounting empirical evidence confirming it, a hypothesis can only be assumed to not have been falsified yet. In part, the concept of falsification is built on our shortcomings of collecting and analysing data: for instance, as we only collected small subsets of data, the data that contradict a hypothesis may have eluded us.

Thus, scientific discovery rests on an iterative process of hypothesis generation and testing. If somebody comes up with a better hypothesis, we discard the prevalent one and accept the new one. In practice, this process is tedious, time consuming, and fraught with dead ends and detours.

**Speeding up discovery**

In contrast, researchers have recently employed a modified approach whereby the hypotheses are algorithmically generated and then tested against data. This approach works when the so-generated hypotheses are variations (such as permutations) of the original. To avoid potential sampling issues, the data need to be sufficiently comprehensive, i.e. moving from gathering relative small samples to capturing close to all relevant data. Data storage and especially analysis will routinely require significant computer resources, but the availability of so-called cloud computing makes resources accessible at close to commodity prices.

The fundamental advantage of this approach is that all possible hypotheses within a particular space are being evaluated so as to identify the one that fits best. Metaphorically speaking, if the current method is to reach into a haystack (representing hypotheses) in the hope of picking the needle (the best fitting hypothesis), this new method will search through the entire haystack in a relatively short period of time. The result is a quite dramatic acceleration of the discovery process.

A well-reported actual case is Google Flu Trends. Google engineers theorized that people affected by the seasonal flu would search online for flu information. This connection could be used to predict the spread of the flu in real time based on what people were searching for, when and where on Google. Google has the data; it receives ~5 billion search requests every single day and has stored past search queries dating back years. In the past, researchers would have guessed what they thought the most likely search terms were and perhaps tested a handful or a dozen to identify the best predictors. In contrast, Google algorithmically generated almost half a billion hypotheses based on 50 million most frequently used search terms to identify those with the best correlational fit.

Algorithmically generating and automatically testing hypotheses only works, of course, if the set of hypotheses is sufficiently defined and demarcated, and a comprehensive data set available. If not close to all relevant data is captured and analysed, the danger of sample biases persists. For instance, if close to all data but only about a particular sub-population is utilized, challenges of generalizability endure. And while it does away with human biases in picking hypotheses to test, it does not eliminate other biases, such as defining an incomplete hypotheses space, or collecting data erroneously.

But this speed up of the discovery process is only one of four ways Big Data may reshape the conventional scientific method.

**The quality–quantity trade-off**

The current scientific approach predicated on small data understandably places a strong emphasis on data quality: as relatively few data points are utilized, any variation in quality will have a negative impact on the validity of results. If the amount of data—in relative terms—is limited, the focus on data quality is justified; advances in data quality is the most efficient way to improve overall results.

In contrast, if the cost of using data plummets, the situation becomes more complex. Data quality is no longer automatically the most efficient strategy to ensure robust results. At times, a better approach may be to collect vastly more data at very low cost, even if the data gathered is of lower or varied quality. This does not mean that data quantity now trumps data quality, but that there is a trade-off between quantity and quality of data. This trade-off has always been present, but in practical terms the high cost of utilizing
data has made quality trump quantity. Thus, researchers in the future will more often have a choice whether to opt for more, but less quality data or for less, but high-quality data. The answer to this trade-off will depend on the specifics and circumstances of each case, rather than be based on a heuristic that favours quality over quantity.

In the context of cardiology, significant amounts of data can be and are collected, but the overall number of patients represented in data sets is often limited to one research undertaking, practice, or clinical organization. So far, one reason militating against combining data sets from different contexts was a concern about comparable data quality across settings. As a result, large-scale data analyses have been comparatively infrequent. As collecting data gets easier and cheaper, and thus dramatically more data become available, we will see more research based on a combining of many smaller data sets to one large set. One case in point is a research project named ‘My Heart Counts’. Run by researchers at Stanford University, the project is a smartphone app based on an open-source software framework and utilizing the sensors in Apple iPhones and other wearables to gather massive amounts of relevant human health data to study cardiovascular diseases (https://med.stanford.edu/myheartcounts.html (15 June 2015)).

Deductive and inductive approaches

Within the current scientific method, a deductive approach to discovery is favoured. As Popper argued, the alternative—an inductive approach—is saddled with methodological shortcomings.9 In part and in light of his theory of falsification, Popper was concerned that when humans generalize from specific cases, it opens the floodgates to biases and similar shortcomings. Deductive reasoning, on the other hand, is grounded in theory, designed for abstraction and generalization, and thus much less likely to be tainted by partialities, preferences, and prejudice.

However, deduction is also quite challenging for humans. Our impulse is to reason from concrete observations and thus to think inductively. In medicine in particular, the immediacy of a particular case requiring attention and decision lends itself to inductive perspectives. In addition, while a deductive approach may avoid some of the challenges of data bias (and resulting bias in the hypotheses), it does not avoid bias in generating hypotheses. Moreover, in a deductive setting, the emphasis on hypotheses grounded in theory rather than the product of data (however partial) also limits originality; human reasoning then tends to follow well-known paths rather than breaking out of them. The result may be that researchers err on the side of conservativism, often aiming for small improvements rather than more radical approaches.

Big Data may offer an alternative, based on a pragmatic view of discovery. If relatively speaking comprehensive data is being used, some of the biases when only collecting small samples disappear or are at least significantly reduced. Therefore, an inductive approach may no longer be as troubling as it is in a small data context.

The idea is to use comprehensive data sets, and look out for patterns in it that point to intriguing statistical connections, which warrant further exploration. This way, data (and its analysis) are used to stimulate the generation of novel hypotheses, and perhaps even theories by pointing researchers to the most promising direction. Big Data experts call this ‘letting data speak’.2

This approach clearly has inductive elements; it suggests that humans can be inspired by data and data analysis. But because of the comprehensive data sets used, it does not represent a comeback of old-fashioned inductivism. It also does not delegate in full the generation of hypotheses or theories to machines or data. Rather, the patterns emerging from data analysis are taken as useful hints what to investigate further. They are complementing rather than replacing human agency.

And yet, there is not mistaking that such a Big Data approach is different from how deductive discovery works, at least in theory, today. It does assign more credence and more value to data and places less of an emphasis on human intuition to deduce how the world works. In this very way, it is an intriguing combination of being more empirically grounded, and more reliant on data on the one hand, while accepting, even embracing the value of specific instances and cases on the other hand. It thus may be well suited for medical discovery in general and discovery in cardiology in particular.

Realizing the value of ‘What’

Correlation is not causality, caution statistics lecturers already in the first few sessions of their courses. As much (but not all) of Big Data analyses is based on identifying correlational connections in the data, such analysis cannot in itself offer causal insights.

Because humans are cognitively primed to understand the world through sequences of causes and effects, they have long privileged causal investigations. Understanding the ‘why’ has long been seen as the primary path to discovery. Realizing what was happening was seen as a vastly inferior insight. However, as Daniel Kahneman has shown our human preference to uncover causality has often led us to rely on incorrect causal connections. Humans come up with them in cases of ‘fast thinking’ (Kahneman) and believe in them even though thorough investigations reveal that frequently they are not true. In contrast, actually establishing causality is quite difficult in many more complex contexts, requiring significant resources and taking much time to complete.11

With Big Data, an alternative path of discovery may open up that is steeped in pragmatism. If in the future data can be utilized at relatively low cost, it may make sense to subject the data to correlational analyses. The results could highlight intriguing connections in the data, thereby already offering actionable insights under specific circumstances, even in the absence of uncovering causality. Secondly, such intriguing correlations could also act as valuable filters to help researchers select best candidates for causal investigations. Given the high resource investment necessary in the search for causality, identifying the most promising candidates is highly valuable.

Big Data fuels the already existing debate about how much evidence for causality we need before we feel confident enough to act on it. This has been an ongoing discussion in all sciences; the medical field in particular is no stranger to this debate. At its core, it is the question about when and on what basis to act.

If, for instance, as a result of a Big Data analyses a subtle pattern in data would be discovered that with a high degree of likelihood could predict the onset of a future illness, should cardiologists give patients with this pattern in their data medication even though the causal link between data and illness remains unstudied? How much proof do we require before we act?
To provide guidance when answering such difficult questions, experts acting for scientific associations and regulatory bodies have developed rules and procedures that aim to ensure appropriate decisions. These guidelines stem from a small data context, in which the fundamental issues on data usage seemed largely settled. As Big Data is providing researchers with not just new tools, but a new perspective on the role of data, these rules and procedures may soon come into question.

Researchers at the University of Ontario offer a case in point. They identified patterns in data of prematurely born babies that indicate an elevated risk of an infection much before conventional symptoms manifest themselves. They do not know, however, whether and what the causal linkage is between the data pattern and the later illness.\(^ {12,13}\) Should we accept this purely correlational pattern as sufficient to treat such babies preemptively? How much additional evidence do we require, and under what circumstances?.

In the 19th century, Ignaz Semmelweis’ suggestion to wash hands with chlorine before treating patients was disregarded for years, because Semmelweis could only show correlations, and his suggested causal connection was not only unproven but turned out to be simply wrong. His pragmatic intervention, however, worked. In contrast, the reluctance of his colleagues to implement his measure may push medical professionals to confront a reality, in which some clinicians act pragmatically even though they at best know what, or what kind of data, is needed, but more often whenever it becomes available and easy to gather.

Consequences for the field

If the traditional scientific method is being adapted thanks to Big Data, it will have significant consequences.

Further emphasis on data

The most obvious consequence is that data, already a core resource for cardiologists, will receive further emphasis. As the cost of using data plummets, researchers have a strong incentive to collect as much data as possible, to combine different data sets so as to improve comprehensiveness, and—keeping in mind affordable storage—to reuse the data for different purposes in the future. Each one of these entails a significant change from current behaviours.

Today data are being gathered with a particular purpose in mind. Collection takes place after the question one wants to answer has been defined clearly. In contrast, in the future, researchers will want to collect data whenever the cost of doing so is comparatively incidental. This means that data are no longer collected only when it is needed, but more often whenever it becomes available and easy to gather.

Because relative scale matters in Big Data analysis, a much bigger incentive exists going forward to combine data sets, even of differing qualities. In the past, such a combining of data has happened relatively infrequently. In part, this was because of regulatory restrictions. But in part, it was also due to organizational sensibilities and the desire to retain control over the fruits of one’s research efforts. As the incentives to combine data strengthen, they eventually may outweigh desires to keep data sources apart, even if that entails some loss of direct control.\(^ {15}\)

Until now, data collected for a particular research project also were rarely reused. Here, too, regulatory constraints played a role, but equally important has been the rather high cost of data storage. As a consequence, data were routinely discarded after they had been used for the purpose they were collected. That, too, is changing as the economics of data storage facilitate repeated reuse.

Taken together, this fosters a renewed debate about the appropriate choice of one of the current core mechanisms to protect privacy—limiting the purpose of data use, and linking it to individual consent at the time of collection.\(^ {16}\) But likely consequences go beyond the immediate issue of privacy. If data are valuable because it can be combined and reused, researchers need to think about appropriate organizational structures and processes of such reuse.\(^ {17}\)

How can a cardiologist beginning a particular research project avail herself of existing data sets from colleagues elsewhere? How can such data be found, and under what conditions can it be reused? How will data have to be prepared for reuse (for instance by ‘anonymizing’ it or adding sufficient meta-data)?\(^ {18,19}\) Is mandatory access to one’s data for other researchers (‘open data’) going to be the next requirement by funding authorities? And how can research organizations ensure that research data are shared at the least within their organization? These and similarly nascent but pressing questions will keep not just the professional leadership of the discipline busy for years to come.

Redefining the comparandum

We define a data point to be out of the norm by comparing it with others. In the medical field, there are two distinct ways to compare a data point. It either can be compared to data points gathered earlier or later, thereby relating data across time, or one can compare data with all instances or cases. Both approaches have pros and cons, but in the medical field the latter is often the only method available. This is because patients come to see doctors when they feel sick, so the data gathered from them cannot easily be compared with data when they felt healthy because such data were not collected then. Thus, one had to compare their data with data of other humans or more precisely with a fictitious ‘average human’. This is not ideal, as humans differ from one another, but in the absence of data from a patient’s past it was often the best option.

In the future, however, with powerful yet affordable sensors built into digital wearables such as watches, and affordable data storage and management tools, humans will be able to collect data about their body regularly even when feeling perfectly healthy. In fact, innovative software developers have been reutilizing existing sensors even in smartphones to gather some of that information. With the availability of the necessary gadgets, individuals are getting curious about their health data and begin to actively collect it. The ‘quantified self’ movement of individuals measuring much of their daily routines and activities is arguably the most visible part of this trend.\(^ {20}\)
This will result in rich and comprehensive data sets for each individual and across time, providing doctors with the necessary empirical basis to compare an individual patient's health not only against an average patient but against the individual’s own past healthy self.

This may prompt more customized and thus targeted diagnoses as well as treatments. In a few years, standard dosages administered to patients will likely be replaced by a more personalized approach. It may also get us to reassess the definition of core concepts such as what constitutes an ‘illness’: perhaps as a deviation of the status quo compared with the past rather than often just a deviation of the individual from the general.

Such personalized medicine is already taking shape—to an extent—and is much discussed both within the profession as well as the general public. But it is important to understand that much if not all such personalization rests on the foundations that Big Data provides to offer a better comparandum in the data collected.

**Empirically approximating reality**

Scientists have traditionally striven for absolute truth, for understandings that transcend context and time. This will not change. But in the life sciences, with often highly complex phenomena that we do not yet fully grasp, doctors will take on board that insights may come in small portions; that rather than arriving swiftly at an absolute truth they will develop an iterative, process-based understanding of a patient’s medical reality. Each additional data point collected may get them a tiny bit closer, but full and comprehensive understanding may remain illusive, at the very least in practical terms.

This pragmatic and iterative approach is a far cry from the ebullient hopes perhaps two decades ago that the inner workings of the human body will get deciphered rapidly. It reflects humility vis-à-vis a complex, diverse, and dynamic reality, in which absolute truths are fewer and further apart than were naively anticipated.

At the same time, it is not an unfamiliar situation in the evolution of human discovery and understanding. Whenever a paradigmatic shift in perspective and outlook occurred in the past, it prompted first the realization of the shortcomings of our existing understanding, and the tremendous amount of work, but also of opportunities of insights that lay ahead. The Big Data era is no different in that respect. A more iterative, process-oriented approach, however, also necessitates an infrastructure of continuous data collection and analyses that is very different from what is in place today. This is challenging, even for engineers at Google.

This emerging emphasis on complexity and iteration also re-focuses the attention of the medical community on something familiar: the human, the specific case, and its specificities. In this sense, the promise of Big Data is that the focus on the individual can now be paired successfully with the power of empirical data and its analysis rather than being pitted against it.

**Conclusions**

Taken together, the growing importance of data, the ability to compare on an individual level, and the reminder of discovery as an iterative process approaching a complex reality will profoundly shape life sciences and medical practice. In areas such as cardiology that have long collected swaths of empirical data, these changes will likely be felt first, as new opportunities but also distinct challenges beckon. What is key, though, is to realize not only that change is imminent, but that this change is paradigmatic, necessitating a rethinking of much of the established processes, structures, and even institutions through which the profession engages in research, advances discoveries, and tends to patients.
Author’s contributions

V.M.-S. acquired the data; conceived and designed the research; drafted the manuscript; and made critical revision of the manuscript for key intellectual content.

Conflict of interest: none declared.

References