Big Data and Medicine – A Big Deal?

Viktor Mayer-Schönberger[1], Erik Ingelsson[2]

[1] Oxford Internet Institute, University of Oxford, Oxford, UK;

[2] Division of Cardiovascular Medicine, Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA.

Addresses for Correspondence:

Viktor Mayer-Schönberger

1 St Giles; Oxford; OX1 3JS United Kingdom.

Phone: +44 (0)1865 287210; E-mail: viktor.ms@oii.ox.ac.uk

Erik Ingelsson, MD, PhD, FAHA

300 Pasteur Dr, mail code: 5773; Stanford, CA 94305; USA.

Phone: +1-650-656-0089; E-mail: eriking@stanford.edu

Word count: 248 (abstract), 6,228 (main text)

Short Title: Big Data and Medicine

Abstract

Big Data promises huge benefits for medical research. Looking beyond superficial increases in the amount of data collected, we identify three key areas where Big Data differs from conventional analyses of data samples: (1) data is captured more comprehensively *relative* to the phenomenon under study; this reduces some bias but surfaces important tradeoffs, such as between data quantity and data quality; (2) data is often analyzed using machine learning tools, such as neural networks rather than conventional statistical methods resulting in systems that over time capture insights implicit in data, but remain black boxes, rarely revealing causal connections; and (3) the purpose of the analyses of data is no longer simply answering existing questions, but hinting at novel ones and generating promising new hypotheses. As a consequence, when done right, Big Data analyses can accelerate research.

Because Big Data approaches differ so fundamentally from small data ones, research structures, processes and mindsets need to adjust. The latent value of data is being reaped through repeated reuse of data, which runs counter to existing practices not only regarding data privacy, but data management more generally. Consequently, we suggest a number of adjustments such as boards reviewing responsible data use, and incentives to facilitate comprehensive data sharing. As data's role changes to a resource of insight, we also need to acknowledge the importance of collecting and making data available as a crucial part of our research endeavors, and reassess our formal processes from career advancement to treatment approval.

Much has been made recently of the analysis of massive amounts of data points, often termed Big Data, to gain novel insights into society, culture and human nature. In the health context, use cases range from analyzing Internet search queries to predict the spread of the flu, to exploring large databases of health records to discover indicators for unknown negative drug interactions.[1-3] They include "training" a computer with lots of example images to identify skin cancer at a rate equal or better than a human dermatologist,[4] self-learning systems to help radiologists in differential diagnosis of lung cancer,[5] and image recognition algorithms that use Instagram photos to predict depression.[6]

Opinions about the usefulness of Big Data approaches vary widely. While some assume that Big Data means that any large data sets can be magically transformed into novel insights, others warn that more data merely means more noise in which any true signals drown, and many more are wary of yet another methodological shift in medical research and practice. In this review article, we take a look at the core building blocks of the Big Data approach, and how it differs from more conventional ones; and we look at the implications, challenges and opportunities that come with Big Data.

1. Qualities of Big Data

No simple, widely accepted definition of Big Data exists. Early Big Data work has suggested three "V"s (volume, velocity and variety) of data as defining qualities.[7] While initially helpful as an illustration, later work suggested that the "V"s capture epiphenomenal or consequential elements of Big Data, rather than its defining qualities. Therefore, today the usefulness of the "V"s in capturing Big Data's essence is disputed.[8]

The exact contours of a definition for Big Data elude us, but certain key characteristics have emerged that encapsulate Big Data's key qualities and delineate Big Data from more conventional uses of empirical data. In the following, we describe three such characteristics focused on data, methods, and purpose, and examine their salience particularly in the context of medical research.

(a) Data

The first, and perhaps most obvious, way to understand Big Data is as a massive shift in our ability to collect and analyze data.

Through all of human history, humans have attempted to make sense of the world they live in by observing – essentially capturing and examining data. Being data-focused is nothing new. But working with data has been time-consuming and expensive – often prohibitively so. Consequently, humans collected as little data as necessary and assumed that this dearth of data was a fundamental constraint to human discovery. The development of data sampling, especially randomized sampling, is a response to this assumption of "data poverty": by offering the opportunity to reason about the whole through only looking at a small part, it is designed to provide useful insights from as little data as possible. But this approach has structural shortcomings: the sample needs to encapsulate the essence of the whole. Because this is difficult to do purposefully without knowing the whole, for almost a century, scientists have resorted to randomized sampling.[9] But even a perfectly random sample will capture the whole only with a limited degree of likelihood, as the sample – by definition – cannot contain all the detail of the whole. Consequently, this results in errors.

The dispute over match-fixing in sumo wrestling offers a case in point. For many years, observers believed that match-fixing was taking place at sumo contests, undermining the fairness of the sport. An analysis of a purposeful data sample – focusing on the championship bouts that matter most - revealed no evidence of foul play. Neither did an analysis of a random sample drawn from all sumo contests. Only when researchers collected and analyzed data from every single sumo match over a decade, they discovered evidence that sustained match-fixing was indeed taking place, but not where everyone was looking, and so highly concentrated that the random sample failed to sufficiently capture it.[10]

Focusing on data samples rather than all of the data has always been a choice dictated by the cost associated with collecting and analyzing data. Of course, with sufficient resources, it has always been possible to go beyond the sample. For instance, in the 19th century, an exhaustive analysis of millions of data points extracted from hundreds of ship logs led to a vast improvement in predictions of prevailing winds and currents on the Oceans, greatly reducing shipping times. But the effort took many human "computers" working for almost a decade.[11] Needless to say, very few research endeavors have had the luxury of nearly endless resources.

Thanks to recent advances in digital technology, however, the ability to gather and examine massive amounts of data has become far cheaper and faster. This has fueled data collection at an unprecedented scale globally. The best estimates we have suggest that in the two decades between 1987 and 2007 alone, the amount of data in the world grew one hundred times.[12] This growth has continued, as more recent estimates suggest a doubling of the total amount of data in the world every two to three years. Equally important is a shift in the composition of this data. If in the year 2000, three quarters of data was still analog and only one quarter of data was digital, by 2015, analog data accounts for less than one percent (**Figure 1**). As this predominance of digital data has greatly improved our ability to gather and process data, the resource cost – in terms of time and money - to use not just small samples of data, but all of it, has plummeted.

The change in the use of data is twofold. First, vastly more data is being captured and analyzed than ever before. The absolute number of data points that can and are being used is growing exponentially. But it is not just the absolute growth of data that matters. Arguably more important, second, is that thanks to digital technologies, researchers now capture substantially more data *relative* to the phenomenon they are studying than before. The success in discovering match fixing at sumo wrestling rested not simply on an increase in the absolute number of data examined, but in the fact that relative to the phenomenon to be researched – presumed match-fixing over a ten-year period – the successful analysis utilized all available data and included information about every sumo match.

While small data samples always miss information – either because it was purposefully not collected or because the random sample just wasn't large enough to include the salient data points – collecting data comprehensively offers an informational improvement. There is no certainty, of course, that more comprehensive data will lead to insights. Researchers may still fail to include the relevant data, but using more routinely trumps using less, and with decreasing cost differentials, there is less reason in the future to settle for data samples. Of course, it may be impossible to collect *all* data of relevance for a medical condition; because of the complexity and stochastic behavior of biological processes, prohibitive costs, as well as ethnical or practical challenges related to data collection. But even with these constraints in place, in the future, medical researchers will routinely collect and analyze more comprehensive data from a larger number of individuals, and when possible analyze all data collected, rather than just data samples.

Comprehensive data collection does not mitigate all selection bias. How the phenomenon a researcher aims to study is defined, and thus what data is and is not being collected is still a decision open to error. The focus on the phenomenon and what different types of data may capture it, however, prompts researchers to face perhaps more directly than before the crucially important issue of delineating and making explicit the object of their research. In the medical field, most conventional evidence-based research compares data collected from a variety of different patients. The phenomenon under review thus is the reaction or behavior of an "average" human being in a particular context. This may be useful when trying to gain insights that are generalizable to many (but certainly not all) humans. An alternative approach could focus on just one particular human being and gathering relevant data over time, as was done in the Snyderome.[13] This way, insights are for a specific individual; they may not be generalizable, but they may be highly salient for understanding what is happening to this particular human. Neither approach is better than the other (and of course often data is collected across patients and across time), but being forced to define what phenomenon is under review highlights and makes transparent the specific aim of a research undertaking, and what insights are generalizable in what way.

This shift towards comprehensive data use has consequences for data quality. With samples, data quality was highly relevant. If only a limited number of data points are used to extrapolate to the whole, capturing even a handful of these data points inaccurately may result in erroneous conclusions. If, on the other hand, a massive amount of data points is captured and a relatively small number of them are erroneous may not be similarly problematic. What in practice used to be a simple principle – that data quality matters most – is no longer so straight-forward in a Big Data context; it turns into a more complex trade-off. Sometimes it may be better to combine data from different sources (such as different sensors) even at varying degrees of data collection quality, if in return the heterogeneity of the data sources guards against measuring biases. And sometimes, even data of limited quality may be better if one has a huge amount of it, compared with only a small amount of data at high quality. For instance, in the context of machine translation (using data and software to translate from one language into another), utilizing orders of magnitude more training data albeit of very varying quality beat a small training corpus of high quality. Big Data experts call this the "unreasonable effectiveness of data".[14] Another illustration where more data of lesser quality beats less data of higher quality is the rapid discovery of novel genetic loci associated with complex human traits over the past decade. Early criticism of genome-wide association studies suggested that the data is too noisy – both that the

individuals studied were too heterogeneous, and that the phenotypes were lacking in detail or did not reflect the underlying disease well enough. However, we have now learned that large meta-analyses of heterogeneously sampled individuals with relatively blunt phenotypic measures – such as body mass index and waist-hip ratio rather than exact distribution of body fat measured by computed tomography – can provide hugely important insights regarding the underlying biology, not offered by previous small, hypothesis-driven analyses of very well-characterized individuals with more exact phenotyping.[15-17]

As the approach to discovery moves from samples to comprehensive data use, a further set of questions gain currency: when to stop collecting data, and what to do with "new" data? Because samples at best capture much (but never all) of the essence of a phenomenon, samples are good either at representing phenomena that do not change, or with dynamic phenomena, they have to be retaken and reevaluated in regular intervals. But doing so is costly and prompts challenging statistical questions (including on sample selection). Unsurprisingly, much empirical research in the natural sciences has focused on uncovering the "iron" laws of nature that do not change. Even there, however, seemingly immutable law of nature – think only of Newton's law of gravity – had to be reformulated – think of Einstein's law of gravity - as we uncovered more of nature's complexity. In the medical field, the object to be investigated is generally highly dynamic and changing, even if we'd assume the underlying fundamental laws of physics don't. The conventional way to deal with this dynamic has been to generalize (and thus lose individualized precision), or to acknowledge change through regular resampling. In the context of Big Data, these questions become harder to answer; conventional standard responses may no longer be appropriate, and researchers may – erroneously – be tempted to assume the comparative immutability of phenomena because of mistaken beliefs driven in part at least by simple convenience.

Researchers at Google predicting the spread of the flu virus using search query data offers a case in point. Initially, their prediction model worked well. Published research in 2013, however, showed that Google's flu forecasting system erred badly in predicting the spread of the seasonal flu in December 2012, foreseeing almost twice as many flue cases as public health authorities later confirmed existed.[18]

Google's prediction rested on the assumption that humans impacted by the flu would search the Internet (through Google) for relevant flu information. In building the model, Google used a

years of past search request data and official flu spread data from the Centers for Disease Control. As a result, Google's model captured common human information search behavior in the U.S. in the years before 2009, when Google made its model public. In the years after, Google simply applied the model to new search queries to render its flu predictions.

The model incorporated human behavior – what humans search online. Unlike general laws of physics, such behavior is not immutable. For instance, our search behavior online changes as the Internet becomes more integrated in our daily routines, as more diverse sociodemographic groups go online, and as access shifts from the desktop to mobile devices. This resulted in a behavioral change, and a growing chasm between the model and reality, leading to false predictions. The Google engineers had erroneously assumed (or perhaps hoped) that search behavior would remain unchanged. When Google's engineers modified their model based on data from more recent years, the accuracy of the prediction improved significantly. For dynamic phenomena, static models are inappropriate. As reality changes, so must the model. This requires regular reevaluations utilizing new data.

Underlying is a broader, more general point on the progress of scientific discovery. Many phenomena change over time. But even for those that are truly immutable, the human ability to understand them comprehensively may change, much like we moved from Newton's view on gravity to Einstein's. Perhaps that, too, is but an approximation of what is, getting us closer to the complex reality we perceive. It fits our current needs, but in the future, we may discover yet another version that's even better in capturing the world. What is true for the laws of physics, applies with much greater magnitude to the black boxes in biology that we begin to shed light in with the help of Big Data. The emergent methods of Big Data analysis are often iterant, adaptive, and learning; as more data becomes available and is fed into the system, the system modifies itself, it "learns". But as the model is still learning, the analysis isn't "done" and the results are only preliminary. This tentativeness isn't new; it has always been a foundation of scientific discovery, but Big Data highlights it. One is never done collecting and analyzing data and revisiting the current model and hypothesis, because every additional data point is but an opportunity to learn further and get closer to capturing reality. This is not just a philosophical point, but rather one that permeates how research is being conducted and employed. If in the past, we hoped for iron-clad natural laws; in the future, at least temporarily, we need to be more cautious. Humbly we will have to accept that our learning is but a summary of what we know rather than eternal truth; and, practically speaking, data collection and analysis will need to

continue if we want to progress. It also requires us to find pragmatic answers to questions such as after how many new data points that we collected we should reevaluate our models and redo our analysis (**Figure 2**).

(b) Methods:

The need to tease out maximum insights from a small number of data points has led to the development of a rich toolkit of statistical methods. This has helped discovery in a data-deprived world. Given what they were designed for, these methods may not be ideal anymore when comprehensive data becomes available at low cost. For instance, with conventional statistical tools, we may find a small enough "signal" in any large enough data set, even if it actually is just noise.[19] This is why globally Big Data researchers are working on and building new data analysis methods aligned with large data volumes. While much work remains to be done, notable advancements have been made.

In particular, so-called machine learning approaches have been gaining traction in the Big Data context recently. The underlying idea with these systems is no longer to "teach" a system a particular insight through a concrete algorithm of how the system should behave, but rather let the system learn itself through the analysis of a large volume of training data. The system's model of what it observes through the data evolves as more training data is fed into the system. With such a strategy, the initial model matters little as it evolves over time and with the help of data. The technical concepts of many of these machine learning systems, for instance neural network theory, have been around since the 1970s, but only recently through the availability of massive volumes of data, and huge computing capabilities have these systems begun to deliver on their earlier promises.[20]

Big Data experts have robust discussions about exactly what variation of a machine learning system offers the best fit for what data context. Some systems are supervised, and their success depends critically on tweaking by a human; others achieve success through purely unsupervised machine learning. It is a highly dynamic field, fueled by frequent innovation. Most recently, for example, systems that emulate a form of decaying memory while learning look particularly promising. Experts agree, however, that the most crucial ingredient to success is neither the choice of the right system (and thus learning methodology), nor the role of humans, but the availability of appropriate training data.

Importantly, once a system has been sufficiently trained for its model to capture much of the essence of a particular phenomenon and can be used, its use produces (feedback) data and together with other data can continue its own evolution and refinement. Simply put, when the streams of data are sufficient, the learning system never stops to learn; not only getting better and better, but also adjusting to any changes in the underlying phenomenon.

Because Big Data analytics works quite differently from conventional analysis of "small" data, such as samples, not only are different methods necessary, but also different expertise. Experts trained in traditional statistical methods are comparatively ill-equipped to engage in and manage Big Data analysis, and conventional statistics courses and programs at universities are producing traditional statisticians. This is why many Big Data projects employ experts versed in machine learning, neural networks, and many other cutting-edge Big Data methods. But these experts are still scarce: demand is growing while supply is still limited, as universities grapple with the methodological shift. For the short to medium term, therefore, as much as it is elemental for a Big Data research undertaking to have a methodological expert on board, it will be a challenge for many of these projects to find and retain them.

More generally, these Big Data methods lead to two distinct methodological features of Big Data analysis that have significant consequences far beyond the concrete choice of a particular analytical method. These features are the inductive nature of most successful Big Data systems, and their (at least partial) agnosticism to causal understanding. Due to their iterative "learning" nature, such systems are neither proving nor applying hypotheses about a particular phenomenon that humans conjured. Rather, in most cases the hypotheses emerge from the learning process. It's an inductive approach to human reasoning and scientific discovery. But the system has no explicit "understanding" of the underlying working of the phenomenon it captures; knowledge and insight is captured in the idiosyncratic model that emerges. This does not help us much directly to understand the underlying causal influences, but it does accurately describe reality and predict what output would result from a given input.

(c) Purpose

The shift in methods mirrors a further change that the comprehensive use of data brings about – one of purpose. The inductive method many Big Data projects feature facilitates a broadened

role for empirical data analysis. The conventional approach was for data to be used to evaluate human hypotheses and where possible, falsifying them (according to Popper[21]). Hypothesis testing has been a costly process. Comprehensive data use speeds up the process and makes it eminently more affordable. This was Google's approach when predicting the spread of the flu through search requests. Rather than hoping to *a priori* come up with the combination of search terms that would best correlate with the distribution of the flu, Google tested all 50 million frequent search terms, and a total of 450 million models combining terms. The model with the best fit consisted of 45 search terms, each of which was a perfectly plausible choice.[3] However, it would have been extremely unlikely for a human-generated hypothesis to contain exactly these 45 terms. The net result is a very significant acceleration in the discovery process as hypothesis are no longer evaluated through a manual process involving human hypothesis generation, but through an automated process that tests exhaustively. The strategy is not without pitfalls; tendencies to "over-fit" are of particular concern and care must be taken to avoid these traps.[18]

More importantly perhaps, Big Data aids in human hypothesis generation, too, by stimulating the human capacity to ask the right questions, i.e. to put forward the most promising hypothesis. This is done by taking comprehensively collected data about a particular phenomenon, then subjecting the data to analysis employing Big Data analytics to uncover distinct and unusual patterns in the data, and to use these patterns to prompt hypothesis generation, either by alerting humans, or more and more frequently through iterative machine learning. The hypotheses generated can then be evaluated using more traditional hypothesis testing approaches.

An example of this concept is a Big Data analysis of data from a language learning app that showed a surprising and unknown learning challenge for a specific language concept. The analysis was not undertaken with a hypothesis to uncover exactly this learning difficulty, but the patterns discovered during analysis refocused the attention of the researchers in that direction.[20] A prominent example in biomedicine is the advent of genome-wide approaches and other –omics methods. Before 2005, genetic associations studies were hypothesis-based – variation in a candidate gene was studied in relation to a human trait.[22] Thousands of such candidate gene studies were published, but an astonishing small fraction of these associations have been robust and replicable. After the introduction of hypothesis-free genome-wide association studies, thousands of robust genotype-phenotype associations have been uncovered, informing human biology and giving leads to development of new medical therapies based on

previously unknown pathways that the human mind was unable to come up with *a priori*. This paradigm shift in genomics is an example of the power of Big Data approaches, and may harbinger changes that can impact biomedicine more broadly.

The hypothesis-generating strategy emphasizes the usefulness of inductive approaches for novel discovery; it is not, however, the beginning of an inductive reasoning revolution, or the end of the deductive method. As is well established in the medical sciences and their focus on individual cases, induction has always been playing an important role for discovery. Neither is deduction disavowed forever. Rather, Big Data strengthens a pragmatic strategy towards scientific discovery, in which inductive methods can often reveal patterns that stimulate hypothesis generation, followed by deductive methods to test these hypotheses.

Big Data analytics are often associated with a further shift in focus – away from exposing causality. Behavioral psychologists see humans hard-wired to find answers to the question of "why". Only by understanding root causes, the argument goes, will we be able to shape reality, and therefore unearthing linkages between causes and effects is the ultimate aim of scientific discovery. In contrast, typical Big Data approaches are correlational; they do not prove or demonstrate actual causality.

This isn't a unique feature of Big Data. Much research based on the analysis of quantitative empirical data using conventional statistical methods is correlational. Even traditional A/B testing, whether in the context of double-blind studies or carefully designed lab experiments, offers no "proof" of causality. Any differences between the treatment group and the control group may, at least theoretically, be simply randomness and thus a sign of coincidence rather than evidence of efficacy. This is often forgotten when correlational Big Data approaches are criticized for lacking causal explanatory value.

The exchange about correlation versus causality does resurface, however, a larger debate that has been looming in the sciences in general, and in medicine in particular, for at least over a century. It is a debate about the value of different types of discovery, and what is sufficient to lead to recommendations regarding behavior or treatment.

Perhaps one of the first large-scale debates in the medical sciences about the value of correlation and causality took place when hygienist Ignaz Semmelweis in 1847 instituted mandatory chlorine

hand washing at the maternity ward of the hospital he was working in Vienna, and the mortality rates dropped dramatically. Semmelweis proposed an odd and incorrect underlying cause. Skeptical of his causal explanation most of his colleagues throughout Europe resisted for years to wash their hands with chlorine, causing (as we today know) tens of thousands of unnecessary patient deaths.[23]

Big Data correlational insights raise similar issues: Pragmatically, what do we accept as sufficient evidence to act? How high is the burden of proof? There is no simple answer to these questions, but highlighting Big Data's limitation to correlational insights may force us to revisit our current solutions and protocols and to rethink whether they are still adequate.

In the meantime, there is active research under way to improve correlational methods to intuit at least some causal insights from them, and there have been recent advances in the development of a formal mathematical way to express causality (which, perhaps surprisingly to some, we have not had).[24] Big Data may not be "locked" into a purely correlational straight-jacket forever, but in the medium term we will have to accept that much like many statistical methods in wide use, Big Data approaches are mostly correlational.

This does not imply that causal investigations are going to be replaced by purely correlational ones. Rather, in the future, researchers will have to consider the best fit given the purpose of their investigation. For example, in contexts such as prediction, prognostication and diagnosis, uncovering causality is unlikely to be of central importance. For many instances of drug development on the other hand, causal understanding will continue to be crucial.

Often overlooked in the debate, correlational analysis is not without value for causal investigations. It can act as a cognitive filtering device, which identifies those explanations that seem most likely to have caused a specific effect. These can then be subjected to a thorough conventional causal investigation. As examining causal linkages is very costly, having with Big Data a useful filter in place that helps select the most promising causal hypotheses for further investigation is not only efficiency-enhancing, but also a valuable contribution to the quest to uncover causality.[8] Eventually, it may lead Big Data to become a crucial part of a more staged discovery process, in which a correlational result informs an ensuing causal inquest.

In summary, while there is no comprehensive and concise definition of Big Data, changes in data, methods, and purpose of research undertakings encapsulate different perspectives of a shift towards Big Data approaches in the sciences in general, and in the medical field in particular (**Figure 3**).

2. Implications:

Increasingly embracing Big Data in medical research and practice has a number of important implications that center on the management of data and its role within the profession. In particular, Big Data's qualities contradict some of the standard assumptions in the field, whether regarding data quality or the level of causal evidence required. This prompts the need for, or at the very least the discussion about structural and procedural adaptations, but also a mental flexibility to question long-held habits.

(a) Data management

When using data is costly, data is gathered when there is a concrete need for and routinely disposed of or disregarded after that use. In contrast, Big Data approaches highlight the importance and value of data reuse for novel purposes. We will collect not because we already know what to use it for, but because we might need it in the future. This implies a different attitude towards data: not as something that is needed swiftly for a pressing decision, but something that holds potential future value for the individual, and for society at large. It also suggests that data be stored far beyond it has been used once because of the strong incentive for reuse (**Figure 4**).

Incidental data collection without concrete purpose and long-term data retention, however, run counter to the fundamental principles of information privacy and potentially can violate national and international laws protecting health information privacy like HIIPA in the U.S. or GDPR in the EU.[25, 26] An example of such a clash between the scientific interest of creating a valuable resource for future research and laws regarding data privacy is the conflict surrounding the Swedish cohort study LifeGene.[27] The goal of the LifeGene study was to enroll 500,000 Swedes and follow them longitudinally for at least 20 years, to enable studies biomedical research questions, many of which were unknown at the start of the study.[28] The Swedish privacy regulator found the broad consent collected from study participants unlawful and the study was

halted in 2011. It could only be restarted in 2014 after extended legal battles and revised privacy legislation. Very similar in design and scope as the widely praised UK Biobank study,[29] LifeGene's troubles are a stark reminder of the importance of alignment between data collection efforts and (dated) data privacy laws, while maintaining participants' and society's trust in comprehensive data collection efforts.

The conventional paradigm of privacy protection that focused on collection limitation and constrained data gathering in practice to an activity requiring concrete notice and specific consent is frequently at odds with the Big Data view of data's value and how this value can be reaped. It is tempting, but far too simplistic, to assume that conventional privacy regimes are optimal policies to protect vulnerable patients against overreaching data collectors. Existing information privacy regimes are themselves constructs of a "small data" world, in which cost impeded comprehensive data use; they were palatable policies only in that context. In contrast, as comprehensive data use becomes feasible, stunting potentially life-saving discovery through artificial constraints of how data can be utilized is no longer an effective policy to advance health care and requires additional justifications to continue. Privacy experts around the world have suggested an alternative to constraining data collection and tying it to a particular purpose: to shift the focus of privacy protection to an assessment of the purpose of a concrete data analysis. Permission would be afforded as long as no individual would be obviously harmed by the consequences of the analysis and if the analysis is ethical, even if it entails a reuse of data collected for a different or no apparent purpose.[30] Privacy legislation in a growing number of nations, including in the EU, is inching towards this "use based" alternative privacy approach. But further and more decisive steps towards such a policy shift may be needed.[26]

As the focus on responsible data use gains traction, the debates over health data will likely transition from being firmly rooted in the context of informational privacy (and control) to one of accountability of data users, even irrespective of the will of the individual. Accountable health data usage is a duty vis-à-vis society rather than just individual patients. This requires the development of rules regarding data usage, as well as processes of assessing data use against such rules, paired with suitable enforcement mechanisms. It is possible, for instance, that we'll see data ethics in medicine to be taken on by specialized boards, modelled after ethical review boards more generally. Alternatively, ethical review boards' remit may be widened to include questions on usage of medical data.[31]

Information privacy laws are often cited as the prime hurdle that stunts comprehensive data use in health care. Equally constraining however are limitations in the management of comprehensive health data. Big Data approaches necessitate a caring and curating stance vis-à-vis data, in which the emphasis is on proper capturing, labeling, and storage to facilitate later discovery and repeat utilization rather than fast one-off analysis. A byproduct of such thinking is the need for more capable data management *infrastructures* in health care. But such infrastructures are less physical than informational. It matters less on what storage devices the data is held or through what pipes it flows, than how well it is annotated, deposited, and its content made usable. Consequently, the structural emphasis of future data management will be on enabling appropriate data ontologies, ensuring meta-data availability, and facilitating data reuse through comprehensive discoverability tools and necessitates planning and the allocation of sufficient resources, from equipment to staffing.

A third difficulty is organizational and structural. Data is the raw material to scientific discoveries, but there has been little recognition of those that expended effort in collecting and curating the raw material. The fear is that others free ride if they are given easy access to data, reaping most of the benefits without having had to pay (in time, energy and resources) for collection. Hence, little incentive exists for groups to share their data even within the same organization. The incentive is even lower for data to be shared across organizations, resulting in wasteful duplication of data collection efforts and limited data utilization. This was already suboptimal, but in the context of Big Data, it is a highly inefficient practice. Alternative strategies are needed. These include the creation of concrete incentives, including economic ones, for groups to share data within and between organizations. On the policy side, government grants and subsidies could be tied to mandates making data collected with the help of these funds openly accessible to others. It would be a suitable extension of "open access" policies for journal publications for the data age to not just make the results of data use accessible to the general public, but the raw material that led to these discoveries as well.[32] There is already development in this direction: more and more scientific journals require deposition of the raw data in connection with the publication.[33, 34] Additionally, a number of initiatives have enabled open access to very rich data biomedical data, such as the UK Biobank.[29] However, the overall development is still slow, and most biomedical data remains unavailable for secondary purposes by others than those that collected it in the first instance.

(b) Data's Role

When discovery is built in substantial part on data, not only to validate a particular hypothesis, but to generate hypotheses, and as researchers embrace the value of data through its purposeless collection and repeat reuse, data's overall role in the process of scientific discovery changes. Rather than supporting human ingenuity, it will more often seen as a crucial resource of insights. Accordingly, a wide variety of existing policies reflecting a far more limited role of data will have to be adjusted.

For instance, currently, the rules and conventions regarding authorship of research reporting on scientific discoveries primarily reflects the role of individuals coming up with the project idea, conducting the analyses and writing up the results. We have less developed ways to acknowledge the contributions of researchers collecting and curating data that led to these new discoveries. This seems increasingly inappropriate. We may want to consider publication rules that better reflect the new division of labor, and the important contributions of those that manage data.

Similarly, until now data collection and curation have largely been disregarded in the assessment of research careers; scientific discoveries and their dissemination were most of what counted. When roles – of data, but also of humans – adjust in the Big Data context, the disrespect afforded to data management no longer seems appropriate for career decisions such as promotion and tenure. Here, too, adjustments are in order.

The shift in the role of data necessitates policy adjustments beyond changes in publication practices and career assessments of researchers. For instance, in the context of evaluating the efficacy of new forms of diagnosis or treatment, regulatory agencies may want to reconsider the role of data, and what exactly they require data to deliver as evidence of drug efficacy. We are not calling for a radical modification of evidence-based research, but believe it is time to consider limited alternatives to the highly formalized approval process in health care that is currently in place, and - arguably for Big Data approaches - is slowing down unnecessarily the pace of discovery.[35]

3. Conclusions and a Look into the Future

Big Data promises to reshape medicine. It's driven by the availability of comprehensive datasets, a shift in methods, and a change in the role of purpose (**Figure 5**). It requires medical

researchers to rethink current practices of data management and, more broadly, the role data plays in medical research. This will necessitate a much better understanding of Big Data's inherent limitations, as well as important adjustments in the policies and processes of medical science.

There is an even larger promise, however, associated with Big Data approaches in medicine. It is that with a shift in the role and use of data, we'll be able to bridge a tension between numbers-driven medical research and doctors' tending to specific patient cases. If evidence-based medicine in the small data age was successful, but also tainted by its focus on samples, data quality and summarizing statistics; Big Data approaches, if done right, provide us with the opportunity to refocus more strongly on the individual, thus helping medicine to strengthen its human touch.

Conflicts of Interest

Erik Ingelsson is a scientific advisor for Precision Wellness and Olink Proteomics for work unrelated to the present project.

References

1 White RW, Tatonetti NP, Shah NH, Altman RB, Horvitz E. Web-scale pharmacovigilance: listening to signals from the crowd. *J Am Med Inform Assoc* 2013; **20:** 404-8.

2 Dugas AF, Hsieh YH, Levin SR*, et al.* Google Flu Trends: correlation with emergency department influenza rates and crowding metrics. *Clin Infect Dis* 2012; **54:** 463-9.

3 Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. *Nature* 2009; **457:** 1012-4.

4 Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; **542:** 115-8.

5 Dhara AK, Mukhopadhyay S, Dutta A, Garg M, Khandelwal N. Content-Based Image Retrieval System for Pulmonary Nodules: Assisting Radiologists in Self-Learning and Diagnosis of Lung Cancer. *J Digit Imaging* 2017; **30:** 63-77.

6 Reece AG, Danforth CM. Instagram photos reveal predictive markers of depression. *EPJ Data Science* 2017; **6:** 15.

7 Laney D. 3D Data Management: Controlling Data Volume, Velocity, and Variety. *META Group Report*. 2001.

8 Mayer-Schönberger V, Cukier K. *Big Data*. Boston: Houghton Mifflin Harcourt. 2013.

9 Neyman J. On the Tow Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *J Royal Statistical Society* 1934; **97:** 558-625.

10 Duggan M, Levitt SD. Winning Isn't Everything: Corruption in Sumo Wrestling. *American Economic Review* 2002; **92:** 1594-605.

11 Lewis CL. *Matthew Fontaine Maury: The Pathfinder of the Seas*. U.S. Naval Institute. 1927.

12 Hilbert M, Lopez P. The world's technological capacity to store, communicate, and compute information. *Science* 2011; **332:** 60-5.

13 Chen R, Mias GI, Li-Pook-Than J*, et al.* Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* 2012; **148:** 1293-307.

14 Halevy A, Norvig P, Pereira F. The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems* 2009; **24:** 8-12.

15 Locke AE, Kahali B, Berndt SI*, et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* 2015; **518:** 197-206.

16 Shungin D, Winkler TW, Croteau-Chonka DC*, et al.* New genetic loci link adipose and insulin biology to body fat distribution. *Nature* 2015; **518:** 187-96.

17 Fox CS, Liu Y, White CC*, et al.* Genome-wide association for abdominal subcutaneous and visceral adipose reveals a novel locus for visceral fat in women. *PLoS Genet* 2012; **8:** e1002695.

18 Lazer D, Kennedy R, King G, Vespignani A. Big data. The parable of Google Flu: traps in big data analysis. *Science* 2014; **343:** 1203-5.

19 Silver N. *The Signal and the Noise: Why So Many Predictions Fail--but Some Don't*. Allen Lane. 2012.

20 Mayer-Schönberger V, Cukier K. *Learning with Big Data* Houghton Mifflin Harcourt. 2014.

21 Popper K. *Conjectures and Refutations. The Growth of Scientific Knowledge*. London, U.K.: Routledge. 1963.

22 Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* 2017; **101:** 5-22.

23 Codell Carter K, Carter BR. *Childbed Fever: A Scientific Biography of Ignaz Semmelweis*. New Jersey, U.S.: Transaction Publishers. 2005.

24 Pearl J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press 2009.

25 Terry N. Protecting Patient Privacy in the Age of Big Data. *University of Missouri-Kansas City Law Review* 2012; **81**.

26 Mayer-Schönberger V, Padova Y. Regime Change? Enabling Big Data through Europe's New Data Protection Regulation. *Columbia Science & Technology Law Review* 2016; **17**.

27 Lind A-S. LifeGene: Case closed? 2014. Available from: http://www.crb.uu.se/biobank-perspectives/item/?tarContentId=496824

28 Almqvist C, Adami HO, Franks PW*, et al.* LifeGene--a large prospective population-based study of global relevance. *Eur J Epidemiol* 2011; **26:** 67-77.

29 Sudlow C, Gallacher J, Allen N*, et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015; **12:** e1001779.

30 Cate FH, Mayer-Schönberger V. Notice and consent in a world of Big Data. *International Data Privacy Law* 2013; **3:** 67-73.

31 Fiske ST, Hauser RM. Protecting human research participants in the age of big data. *Proc Natl Acad Sci U S A* 2014; **111:** 13675-6.

32 Walport M, Brest P. Sharing research data to improve public health. *Lancet* 2011; **377:** 537-9.

33 Barsh GS, Cooper GM, Copenhaver GP, Gibson G, McCarthy MI, Tang H, Williams SM. PLOS Genetics Data Sharing Policy: In Pursuit of Functional Utility. *PLoS Genet* 2015; **11:** e1005716.

34 Pham-Kanter G, Zinner DE, Campbell EG. Codifying collegiality: recent developments in data sharing policy in the life sciences. *PLoS One* 2014; **9:** e108451.

35 Master SR, Mayer-Schonberger V. Learning from our mistakes: the future of validating complex diagnostics. *Clin Chem* 2015; **61:** 347-8.