The myths of ‘big data’ in health care

D. J. Jacofsky
From The CORE Institute, Phoenix, Arizona, United States

‘Big data’ is a term for data sets that are so large or complex that traditional data processing applications are inadequate. Billions of dollars have been spent on attempts to build predictive tools from large sets of poorly controlled healthcare metadata. Companies often sell reports at a physician or facility level based on various flawed data sources, and comparative websites of ‘publicly reported data’ purport to educate the public. Physicians should be aware of concerns and pitfalls seen in such data definitions, data clarity, data relevance, data sources and data cleaning when evaluating analytic reports from metadata in health care.

Cite this article: Bone Joint J 2017;99-B:1571–6.

What are ‘big data’, and what are their challenges?
‘Big data’ is a term for data sets that are so large or complex they cannot be processed using traditional methods. Billions of dollars have been spent on attempts to build predictive tools from large sets of poorly controlled metadata. Companies often sell reports at a physician or facility level based on these various sources of claims data and comparative websites of publicly reported data which purport to educate the public appropriately and responsibly.

Challenges for big data include their analysis, capture, curation, inconsistency, incompleteness, relevance, data sharing, storage, transfer, visualisation, questioning, data security and privacy.1,2 In a 2012 Harvard Business Review publication about big data, Shah et al3 wrote “investments in analytics can be useless, even harmful, unless employees can incorporate that data into complex decision making. At this very moment, there’s an odds-on chance that someone in your organisation is making a poor decision on the basis of information that was enormously expensive to collect.”

Prior to analysis, data must be well structured. However, considering the wide array of sources of data in health care and the unstructured nature of the vast amount of information, analysis is extremely difficult. Data sets can be terabytes in size and originate from many heterogeneous sources. Hence, real world data are inconsistent, incomplete and filled with noise. Privacy in health care is a major concern with outsourced data, especially in countries with strict privacy laws regarding healthcare information. Data are generated in different formats, which adversely affect their analysis, management and storage. As databases in the real world are highly susceptible to errors, pre-processing techniques such as scrubbing, integration, transformation and reduction should be applied. These all increase costs and in today’s health care, experts are required to do the work thereby largely defeating the purpose intended by automation. To date, the essential landscapes of big data have not been unified and big data cannot be processed using existing technologies and methods effectively. There are currently no procedures and standard tools to search and analyse large, mixed data sets effectively and accurately.4,5 Healthcare data are an order of magnitude more complex than many other industries, adding to the overwhelming challenge before us.

Data set reliability in health care
For high volumes of data to be reliable, data sets must have structured data fields with data clarity, common data definitions, relevance and consistency. Unfortunately, in health care, each of these requirements tends to be suboptimal, or to not exist at all.6

Structured data fields
Generally, structured data refers to information with a high degree of organisation, such
that inclusion in a database is seamless and easily searchable by simple, standard search engine algorithms. Structured data are those that have been organised into a formatted repository, typically a database, so that its elements can be made addressable for more effective processing and analysis. In a database, each field is discrete and its information can be retrieved either separately or with data from other fields in a variety of relationships. Data are increasingly amenable to processing as they are increasingly structured. Structured, as it relates to data, is not specifically defined, but can be implied in each given setting. Examples of structured data include financial spreadsheets, vital signs, laboratory values and certain claims data such as the diagnosis at admission, age, gender, address, referring provider and type of insurance. Structured data are slowly being built into many platforms of electronic medical records, but comes at the expense of specificity and completeness. Physical examination templates, for example, may be built to contain discrete, structured data fields. These improve the ability to search and analyse these data, but physicians may use the pre-populated examination templates even when certain differences in examination exist. Likewise, a single datum point for maximum flexion, for example, may not allow for the important clinical note that flexion is more painful after 90° in a certain patient. Nuance, often critical to accurate diagnosis, is generally lost in structured data fields.

Conversely, unstructured data make compilation and data searching (or ‘data mining’) a time-consuming, energy-intensive and expensive task. If it were possible to convert unstructured data to structured data instantly or easily, searching through unstructured data would not be problematic, but this is not the case. Data with some form of structure may still be characterised as unstructured if its structure is not helpful in the needed interpretation at hand. Many believe that over 70% of all data are, and will be, unstructured, despite the current disadvantages.7

In health care, unstructured clinical data largely live where they are created: at the point of care in the form of free-text or semi-structured documentation in medical records. For decades, provider organisations have focused on using electronic medical records to capture clinical documentation from physicians and then drive these data into downstream workflows, hoping to gain deeper functionality in billing and decision support. Although some electronic medical record platforms allow for some structured data fields, billions of clinical notes with unstructured data or free text fields are created annually because that is how physicians can best describe a clinical encounter with a patient. Clinical notes, operation notes, pathology notes, radiology reports, radiographs, arthroscopic images and outside records are some examples of unstructured data. It is therefore clear that most clinically important information is unstructured. Although future technologies may allow effective and predictable searches of these data, to date, they have been excluded from automated analysis and reporting. The ability to search unstructured text fields accurately is often claimed by young companies today, yet to date it has not successfully sought out healthcare data in a clinically meaningful way. The work is complex and the variables are enormous. However, like all technology, this too will probably develop to a point of seamless integration.

Data clarity

Data clarity is of critical importance, without which much of the value of its management is lost. Data must be unambiguous to be of a high value. The use of standardised and valid definitions is fundamental to the measurement of healthcare outcomes and/or adverse events.

One example of how the integrity of data can be destroyed by the lack of a clear definition is in the relatively common diagnosis of infection. Infection can be a devastating complication, especially after arthroplasty. The original data in health care generally come from one of two sources: diagnostic codes from payer claims or coders employed in the hospital. Imagine the varied responses you would receive if you questioned a group of orthopaedic surgeons as to how they specifically and definitively diagnose a deep infection in a patient with an arthroplasty. There would be many answers and clarity would be lacking. Now imagine adding Emergency Department physicians and internal medicine providers to the group of respondents. The following example is not an uncommon scenario. A patient presents to the Emergency Department with a red, swollen knee eight days after total knee arthroplasty on a Friday night. The Emergency Department physician admits the patient with a diagnosis of infection after calling the general medicine service on call. The patient is treated with antibiotics and the internist consults the orthopaedic surgeon on their Saturday morning rounds. The orthopaedic surgeon who had performed the arthroplasty consults and discovers that the patient has a mild haemorrhosis, so antibiotics are discontinued and the patient is discharged with conservative care. Many databases that use coding data as their source will now erroneously contain the diagnosis of infection in this arthroplasty and will probably link it to both the facility and the surgeon.

Even data sets about seemingly simple diagnoses such as wound infection are unreliable. Bruce et al reviewed 112 prospective studies in great detail, including the definition and measurement of wound infection. Of these peer-reviewed studies, 26 had no definition despite reporting rates of infection in the manuscripts. A total of 82 studies fulfilled the criteria for both the appraisal and definition of infection, but used many different definitions of infection, ranging from ‘the presence of pus’ to detailed criteria that stratified the severity and components of a wound infection. Furthermore, the literature search referenced five different, nationally recognised definitions: three published by groups within the United Kingdom (including the Surgical Infection Society Group) and two published by the Centers
for Disease Control in the United States. These definitions were each published following multidisciplinary consensus by groups of experts in the field of wound infection, yet all are different.9-13

Comparison between providers or hospitals related to deep vein thrombosis (DVT) is another example of commonly used definitions of data that can be highly problematic. For example, providers may categorise DVT based on different tests, with different sensitivities for diagnosis including ultrasound, venography and magnetic resonance angiography, based on ordering tests due to variable triggers (e.g., symptomatic versus non-symptomatic), reflecting the inclusion or exclusion of partial non-occlusive clots, or based on the inclusion or exclusion of chronic versus acute thrombosis. Add to this the variables of anatomical site and size, it becomes clear that these data will be extremely difficult to search fairly and objectively, even manually, by an expert.

Kraus et al14 pointed out that even defining the baseline characteristics of a patient is difficult. There would be considerable variation if one was trying to normalise data sets simply based on pre-existing conditions. Arthritis, for example, is a common diagnosis leading to an orthopaedic operation. However, arthritis is a spectrum of conditions with varying anatomical, phenotypic, molecular and aetiological variables. The pre-existing diagnosis of hypertension made by the primary care physician may be accurate or inaccurate, as the definitions of this may vary between providers. Severity is a scale and aetiologies vary.

Newer diagnostic code criteria, such as the tenth revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10),15 create a very granular classification of diagnoses. This is a step in the right direction, but the cumbersome scope of this system gives insight into how robust a system of this nature would need to be if it were to include all comorbidities, complications, implants and symptoms with the same level of granularity as the ICD-10 does for the diagnosis. Therein lies the challenge of collecting usable data without the system getting so bogged down that physicians become further thwarted in reporting requirements. Additionally, complex classifications lead to increased human errors. Epelboym et al16 described the complications of reporting in the American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP) and found that in 249 pancreatectomies, there was disagreement in the reporting of complications 27.3% of the time between their review of the cases and the documentation of the clinical coder for ACS-NSQIP. They also noted 34 cases in which a complication was reported and none had actually occurred.16 It should be noted that unlike claim-based data, the complications for ACS-NSQIP are coded by clinical personnel, such as nurses or other trained medical professionals. This highlights the rate of errors to be expected in claims data given the rate of errors even when these highly trained individuals are responsible for the reporting. The strengths and weaknesses of each individual database are outside the scope of this manuscript, but despite its strengths as a large clinically based outcomes database, even ACS-NSQIP is not specific to an individual orthopaedic patient. The misclassification of data is probably minimised by the ACS-NSQIP Data Definition Committee, but outcomes beyond 30 days, implant data, nonunion and functional information like range of movement is not collected in this database.

As the collection of data becomes more automated and objective, and as physicians come together to control and standardise definitions more broadly, the definitions and clarity of data may improve.

Relevance of data
For data to be highly relevant, they must portend improved clinical decision-making and be timely. In the United States, both The Centers for Medicare and Medicaid Services (CMS) and companies such as Premier report data being used to manage the behaviour of providers and in some cases, affect income. Unfortunately, these data are largely claims-based and reveal information that is both minimally actionable and often many months, or even a year out of date. For example, attempts to manage bundled payments in the Medicare Bundled Payments for Care Improvement pilot programme are made difficult because information is reconciled and received in a packaged form nearly a year after the event has occurred. Hospital administrators, or a Chief Medical Officer, are giving physicians information about complications that are months old. Being told that you have an increased rate of infection in January when it is now April, not being given a root cause for the high rate, and being unaware of whether the increased rate still exists attenuates its usefulness.

Even programmes dispensed by the CMS or Department of Health and Human Services that directly impact payment to physicians, such as Medicare Access and the CHIP Reauthorization Act of 2015, affect the pay of physicians a full two years after the work is done. This makes direct correlation for providers between the work and the reward difficult to appreciate, makes a feedback loop for continuous improvement attenuated and therefore makes incentives harder to use to change behaviour when compared with more immediate mechanisms of reward.

Consistency of data
In addition to simply having inconsistent definitions, many database systems to which the same question is applied will give very different results. This lack of inter-system reliability implies that one must question the value of the output from any of the systems. If you asked the same calculator the same mathematical question with the same inputs and got many different answers, it would be impossible to know if any single answer was correct. This is the state of the consistency of healthcare data today.
The National Hospital Discharge Survey and the Nationwide Inpatient Sample (NIS) both collect sample data and are used in orthopaedic research. Bekkers et al.\(^\text{15}\) reported that there were significant differences between these two commonly used data systems and showed that 26 out of 42 factors in demographics, comorbidities and adverse events after total hip arthroplasty in these databases differed by more than 10%. They concluded that “analyses of large databases can have limited reliability and should be interpreted with caution.” The Vascular Quality Initiative (VQI) and ACS-NSQIP are two independently maintained databases that have emerged as the primary tools for the assessment of the quality of vascular surgery in the United States. A study in 2014\(^\text{18}\) reviewed all patients undergoing lower limb bypass surgery during the same five-year period between 2007 and 2012. A total of 240 patients had their data captured in both these databases. They found moderate to good agreement for pre-operative comorbidities except for congestive heart failure and hypertension, for which agreement was poor. Importantly, agreement for post-operative complications such as unplanned return to the operating theatre, wound infection and change in renal function was poor. They concluded that “VQI and NSQIP data sets cannot be used to compare adjusted patient outcomes between institutions.” In a similar orthopaedic study, Bohl et al.\(^\text{19}\) compared NIS with NSQIP to explore the reliability between databases of information about demographics, comorbidities and adverse events in patients undergoing stabilisation of fractures of the hip. There was more than a twofold discrepancy between the databases for this same patient population for coagulopathy, anaemia, non-morbid obesity, renal failure and urinary tract infection. They concluded that “two databases commonly used for orthopaedic research can identify similar populations of operative patients but may generate very different results in terms of comorbidities and adverse events.”

**Cleaning data**

Data cleaning is a process of detecting, diagnosing and editing faulty data. The details of how exactly this process should work are outside the scope of this paper, but it is important to recognise this concept as a required and distinct activity. In clinical research, errors occur despite the careful design and conduct of studies and the implementation of error-prevention strategies. Data cleaning is meant to identify and correct these errors, or at least minimise their impact on the results of the study.\(^\text{20}\) Statistical societies recommend that the description of data cleaning becomes a standard part of reporting statistical methods.\(^\text{21}\) Exactly what to report and under what circumstances remains mostly unanswered. In practice, it is rare to find any statements about the methods of data cleaning or the rates of error in medical publications. In the guidelines from The Society for Clinical Data Management\(^\text{22}\) for good clinical data management practices, they state: “Regulations and guidelines do not address minimum acceptable data quality levels for clinical trial data. In fact, there is limited published research investigating the distribution or characteristics of clinical trial data errors. Even less published information exists on methods of quantifying data quality.” Data cleaning can be considered a three-stage process, involving repeated cycles of screening, diagnosing and editing of suspected abnormal data.\(^\text{20}\) Inaccuracy of a single measurement and datum point may be acceptable, and could simply be related to the inherent technical error of the instrument which made the measurement. Hence, data cleaning should focus on the errors which constitute a major shift in or beyond the population distribution. Data cleaning must be based on the knowledge of possible technical errors and expected ranges of normal values. Some errors deserve priority, but which ones are most important is highly study-specific. In most clinical epidemiological studies, errors that need to be cleaned at all costs include missing gender, gender misspecification, date of birth errors and date of examination, duplications and/or merging of records and results which are biologically impossible. For example, in nutritional studies, errors of dates lead to errors of the reporting of age, which in turn lead to errors in weight-for-age scoring and to the misclassification of subjects as under or overweight. Errors of gender and date are particularly important because they contaminate many derived variables.

In the screening phase, it is convenient to distinguish four basic types of oddities: lack or excess of data; outliers, including inconsistencies; strange patterns of distribution; and unexpected results. In the diagnostic phase, the purpose is to clarify the true nature of the worrisome data points, patterns and statistics. Possible diagnoses for each of these are as follows: erroneous; true extreme; true normal (i.e. the expected normal range was incorrect); or idio-pathic (i.e. no explanation found but still suspect). Some data points are clearly logically or biologically impossible.\(^\text{23}\) The diagnostic phase is labour-intensive and the budgetary and logistical and personnel requirements are typically underestimated or neglected at the design stage of the study. Very little, if any, of this work is done in healthcare data management and searching. In the treatment stage, identified errors, missing values and true (extreme or normal) values must be addressed and one must decide what to do with problematic observations. The options are limited to correcting, deleting or leaving the observations unchanged. There are some general rules for which option to choose in order to minimise the subsequent amplification of errors.\(^\text{20}\)

In summary, when evaluating big data systems, one should question how inevitable errors in data will be screened, diagnosed and treated. The lack of a clear strategy for data cleaning in a system should be viewed as a warning sign that the system is inherently prone to errors and that the design of the system itself may have been developed by individuals who lack fundamental knowledge about the creation of accurate data management systems.
Registries
Worldwide, registries are a worthwhile and noble endeavour to track the societal impact of different implants or techniques over time, but they will not provide the ability to predict how best to manage an individual. Although a great step forward in orthopaedics, much of these data are so fundamentally flawed that it will continue to prove impossible to build meaningful predictive tools from them. The protocols used by different physicians and hospitals are varied and non-standardised. The data sets are woefully inadequate and are missing critical information. Likewise, they are fraught with errors of coding that in some hospitals are in excess of 20%. However, most importantly the fundamental tenant of data searching is that standard definitions must be used to challenge registries, much like all other healthcare data sets. In the case of many data sets, there is no standard for any definitions, nor review of reported complications to see what fits defined criteria. What is the exact definition of foot drop, for example, and does it matter if it lasts one day or one year? What is the definition of a superficial thrombosis versus a DVT? Does the size and need to treat matter? How is infection defined?24 This is not in any way to say that registries do not provide much positive value at a societal level. They have predicted implants that fail early and measuring patient-reported outcome measures has given important information about outcomes that matter to patients. These outcomes confirm, for example, that arthroplasty is an effective treatment for disabling conditions of the hip and knee. In addition, these outcomes provide information about an important minority of patients who do not improve as expected or express dissatisfaction with the results of the operation.25,26 However, these large predictions for groups of patients in different cohorts do little to allow someone to compare one doctor with another from these data sets, nor allow a physician to know how best to manage a particular patient with their set of comorbidities and demographic profile sitting in front of them in the clinic. A registry with widespread adoption and good data governance can successfully demonstrate trends and influence future orthopaedic practice, but correlations can be easily misinterpreted as causalities given the errors of data and the nature of the barriers metadata can assemble and amplify.27 Many studies have shown that registries, like all big data systems, lack patient level predictive specificity and an active system for continuous evaluation of the quality, clarity, relevance and validity of data.28-32

Solutions
Many believe, myself included, that the key to the future is standardised, evidence-based order sets, used universally and coupled with systems that track complications, outcomes and patient factors based on specific definitions. Combine this with the tracking of the behaviour of physicians, the compliance of providers with order sets and protocols and patient-reported outcomes, you have something that will begin to allow statistical searching of data, and permit the production of truly predictive tools for individual patients. The CORE Institute has been standardised for many years for most procedures and has developed unique custom tools to track complications, provider behaviour and outcomes in real time both in and out of the hospital. These data have allowed the Institute to develop validated predictive tools that are being used to predict for a specific patient prior to an intervention, what care is required and what can be modulated to improve the cost of care and outcomes for that patient.

These tools have allowed dramatic reductions in the permember per-month costs of groups of patients, in some cases by 50%.33 This reduction was driven by decreased readmissions, decreased waste, decreased rates of complications and decreased use due to preventive and wellness programmes such as a bone health programme. These findings will only improve with the technology which is available to us. The future will probably include wearable devices for physicians that will, for example, allow them to know which arthroplasty patient is missing their functional and recovery benchmarks after discharge and needs therapy, and who is meeting benchmarks each day with no costly interventions at all. Applying the right intervention to the right patient at the right time is far better than applying all interventions to all patients all the time.

Summary
The notion of tracking data is certainly a good one and this paper is not intended to persuade physicians otherwise, but the value of the activity must be justified with actionable, reliable information. Adverse surgical events and medical complications contribute significantly to cost, morbidity and mortality. However for many reasons, monitoring and recording clinically relevant and accurate data is almost universally imprecise and, at times, simply incorrect. The analysis of big data is often shallow compared with the analysis of smaller data sets.34 Big data is both a ‘buzzword’ and a vague term, but at the same time, an obsession with entrepreneurs, consultants, scientists and the media.35,36 Big data often pose the same challenges as small data, and adding more data does not necessarily solve problems, but may magnify them. Big data may also introduce new problems, such as the problems of multiple comparisons: simultaneously testing a large set of hypotheses is likely to produce many false results that mistakenly appear significant. Ioannidis37 argued that “most published research findings are false” due to essentially the same effect: when many scientific teams and researchers each perform many experiments and process the same data in many ways, the likelihood of a significant result being actually false grows quickly. This is highly problematic when only positive results from the data set are published. Furthermore, the results of analysing big data are only as good as the model on which they are predicated. In the most recent famous American example, big data took centre stage in attempting
to predict the results of the 2016 United States Presidential Election. Forbes, 38,39 for example, predicted that, “If you believe in Big Data analytics, it’s time to begin planning for a Hillary Clinton presidency and all that entails.”

The bending of the cost curve in health care is surely needed and long overdue. The shift from volume to value is inevitable, as it has been in every other industry. The path to get us there may not be straight and there will almost certainly be new obstacles and challenges to overcome along the way, but the road must be travelled. Big data solutions will certainly be a future cornerstone of this transformation. New technologies that attempt to solve the need for individual patient level and predictive data will continue to emerge. Regardless of whether physicians directly play a role in the development of new technology for the management of big data, progress will be limited until they work to solve the problems that only they can control. Medical societies and provider groups must come together to solve the problem of poor clarity of data due to a lack of standardised definitions and reporting, and patients and physicians alike must begin to embrace the reality of being measured. Problem of poor clarity of data due to a lack of standardised definitions and provider groups must come together to solve the Big data solutions will certainly be a future cornerstone of this transformation. New technologies that attempt to solve the need for individual patient level and predictive data will continue to emerge. Regardless of whether physicians directly play a role in the development of new technology for the management of big data, progress will be limited until they work to solve the problems that only they can control. Medical societies and provider groups must come together to solve the problem of poor clarity of data due to a lack of standardised definitions and reporting, and patients and physicians alike must begin to embrace the reality of being measured.

No benefits in any form have been received or will be received from a commercial party related directly or indirectly to the subject of this article.

This article was primarily edited by J Scott.

References

36. Harford T. Is big data: we are making a big mistake? Financial Times. 2014. https://www.ft.com/content/21a6e7d8-b479-11e3-a09a-00144feabdc0 (date last accessed 03 August 2017).