Ghassemi et al. Critical Care (2015) 19:118
DOI 10.1186/s13054-015-0801-4

CRITICAL CARE

REVIEW

State of the art review: the data revolution in critical care

Marzyeh Ghassemi[1], Leo Anthony Celi[2*] and David J Stone[3]

Abstract

This article is one of ten reviews selected from the Annual Update in Intensive Care and Emergency Medicine 2015 and co-published as a series in Critical Care. Other articles in the series can be found online at http://ccforum.com/series/annualupdate2015. Further information about the Annual Update in Intensive Care and Emergency Medicine is available from http://www.springer.com/series/8901.

Introduction

Many recent articles highlight the data revolution in healthcare, an offshoot of the vast amount of digital medical information that has now accumulated in electronic medical records (EMRs), and present it as an opportunity to create a 'learning healthcare system'. The generally proposed vision is for a population data-driven knowledge system that generalizes from every patient's life, disease and treatment experiences to impute the best course of action for diagnosis, prognosis and treatment of future patients.

There have also been many articles focusing on the risk that naïve use of Big Data (or data in general) poses. As stated by Zak Kohane of Harvard Medical School, Big Data in healthcare cannot be a simple, blind application of black-box techniques: "You really need to know something about medicine. If statistics lie, then Big Data can lie in a very, very big way" [1].

This paper will discuss the general issue of data in critical care with a focus on the Big Data phenomenon that is sweeping healthcare. With the vast amount of digital medical information that has accumulated in EMRs, the challenge is the transformation of the copious data into usable and useful medical knowledge.

* Correspondence: lceli@bidmc.harvard.edu
[2]Beth Israel Deaconess Medical Center, Harvard-MIT Division of Health Science and Technology, Division of Pulmonary, Critical Care and Sleep Medicine, Cambridge, USA
Full list of author information is available at the end of the article

We are experiencing a rapidly expanding collection of vast amounts of clinical data from routine practice and ambulatory monitoring. Clinicians must already make sense of a diverse variety of data input streams in order to make clinical decisions. Data from our everyday activities (financial transactions, cellphone and Internet use, social media posts), the environment, and even the local government promise to provide even more clinically relevant information (Figure 1), but to what end? And how can increasing amounts of data be incorporated into a system of already overburdened clinicians?

The bottom line is that pertinent quality data add tremendous value, which accounts for their 'unreasonable effectiveness'. There is no way to minimize undesirable variability in practice without the data to substantiate the standardization. The volume and variety of increasingly available Big Data can allow us to interrogate clinical practice variation, personalize the risk-benefit score for every test and intervention, discover new knowledge to understand disease mechanisms, and optimize processes such as medical decision making, triage and resource allocation. Clinical data have been notorious for their variable interoperability and quality, but a holistic use of the massive data sources available (vital signs, clinical notes, laboratory results, treatments including medications and procedures) can lead to new perspectives on challenging problems. While the wetware of the human mind is a wonderful instrument for this purpose, we must design better data systems to support and improve those components of this data integration process that exceed human abilities [2].

Data in critical care

Critical care environments are intense by definition. Decisions in the intensive care unit (ICU) are frequently made in the setting of a high degree of uncertainty, and clinical staff may have only minutes or even seconds to make those decisions. The increasing need for intensive care has spiked the ratio of ICU beds to hospital beds as the ICU plays an expanding role in acute hospital care

BioMed Central

© 2015 Springer-Verlag and BioMed Central. This article is co-published by agreement with Springer-Verlag. Permission for reuse must be sought from the publisher.

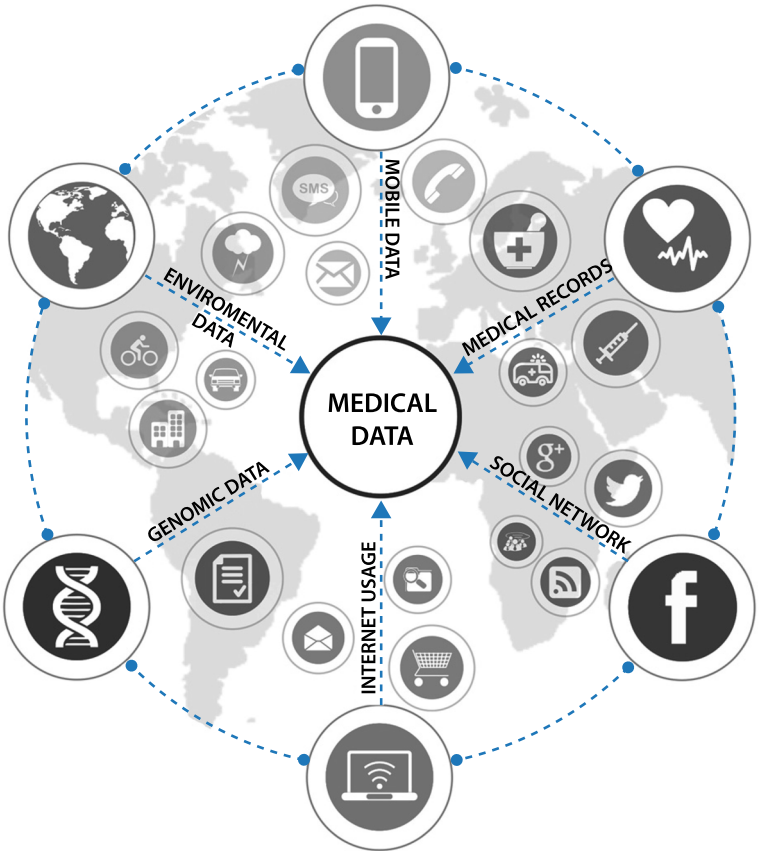

Figure 1 Where Big Data in healthcare come from (figure courtesy of Yuan Lai).

[3]. But the value of many treatments and interventions in the ICU is unproven, with many standard treatments being ineffective, minimally effective, questionably effective, or even harmful to the patient [4]. In a setting where the effects of every intervention are subject to patient and clinical context-specific factors, the ability to use data for decision support becomes very attractive and closer to essential as increasing complexity transcends typical cognitive capabilities.

An example of collected data being used to infer high-level information is the ICU scoring systems in use today. ICU scoring systems, such as APACHE (Acute Physiology and Chronic Health Evaluation), MPM (Mortality Probability Model), and SAPS (Simplified Acute Physiology Score), are all based on the use of physiologic and other clinical data for severity adjustment (Table 1). While these scores are primarily used to assess and compare ICU performance (e. g., by examining the ratio of actual-to-predicted outcomes) they also have use as short-hand indicators of patient acuity [5]. But scoring system value depends not only on the accuracy of the underlying data, but also on clinical trust in the reliability of the data and the predictions based on that data. In 2012, scoring systems were used in only 10% to 15% of US ICUs, despite demonstrated good discrimination and calibration [6].

In practice, clinical prediction must be motivated by the needs of clinical staff, and this must be driven in large part by perceived utility and an increase in technical comfort amongst clinicians. Some of the biggest opportunities for Big Data to make practical gains quickly are focused on the most expensive parts of current clinical practice: Reliable, predictive alerting and retrospective reporting analytics for high-cost patients, readmissions, triage, clinical decompensation, adverse events, and treatment optimization for diseases affecting multiple organ systems [7].

ICU physicians have embraced the value of collecting and storing electronic clinical records, and this has led to partnerships between industrial and academic entities. For example, the commercial APACHE Outcomes database has gathered partial physiologic and laboratory measurements from over 1 million patient records across 105 ICUs since 2010 [8]. The Philips eICU archives data from participating ICUs, and has collected an estimated database of over 1.5 million ICU stays. As an ongoing provider, the eICU adds more than 400,000 patient records per year to its stores, and these data are also commercially available

Table 1 A comparison of intensive care unit (ICU) scoring systems (from [47] with permission)

ICU scoring system	Timing of data collected	Physiological values	Other required data	Total data elements required	Original reported mortality prediction performance
SAPS III	Prior to and within 1 hour of ICU admission	10	Age, six chronic health variables, ICU admission diagnosis, ICU admission source, LOS prior to ICU admission, emergency surgery, infection on admission, four variables for surgery type	26	AUC = 84.8% (n = 16,784)
APACHE IV	First ICU day (16–32 h depending on time of admission)	17	Age, six chronic health variables, ICU admission diagnosis, ICU admission source, LOS prior to ICU admission, emergency surgery, thrombolytic therapy, FiO_2, mechanical ventilation	32	AUC = 88.0% (n = 52,647)
MPM_0-III	Prior to and within 1 hour of ICU admission	3	Age, three chronic health variables, five acute diagnosis variables, admission type (e. g., medical-surgical) and emergency surgery, CPR within 1 h of ICU admission, mechanical ventilation, code status	16	AUC = 82.3% (n = 50,307)

SAPS: Simplified Acute Physiology Score; MPM: Mortality Prediction Model; APACHE: Acute Physiology and Chronic Health Evaluation; AUC: area under the curve; CPR: cardiopulmonary resuscitation; LOS: length of stay.

to selected researchers via the eICU Research Institute [9]. In contrast to these commercial databases, the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) database is open and publicly accessible (Figure 2). Over the past decade, the MIMIC database has collected clinical data from over 60,000 stays in Beth Israel Deaconess Medical Center ICUs, including clinical notes, physiological waveforms, laboratory measurements, and nurse-verified numerical data [10].

Establishing knowledge

Medicine is ultimately based on knowledge, and each of the many ways to establish knowledge has certain advantages and pitfalls. Here, we focus on the randomized controlled trial (RCT), observational studies and what we have termed "dynamic clinical data mining" (DCDM) (Figure 3).

RCTs are the gold-standard for clinical knowledge discovery. But 65 years after the first RCT was published, only 10–20% of medical decisions are based on RCT-supported evidence [11]. When examining the validity of a variety of medical interventions, about half of systematic reviews report insufficient evidence to support the intervention in question. Most treatment comparisons of clinical interest have actually never been addressed by an RCT [12]. The reality is that the exponential combinations of patients, conditions and treatments cannot be exhaustively explored by RCTs due to the large cost of adding even small numbers of patients. Furthermore, the process of performing RCTs often intentionally or inadvertently

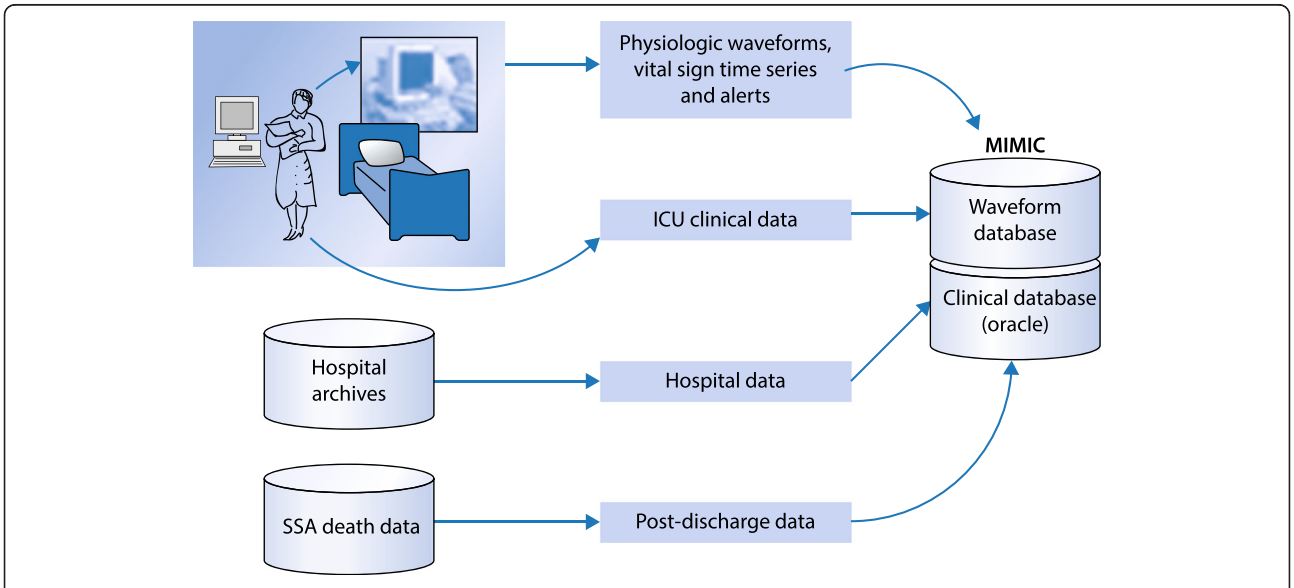

Figure 2 The MIMIC database. SSA: social security administration (figure courtesy of the Laboratory of Computational Physiology, Massachusetts Institute of Technology).

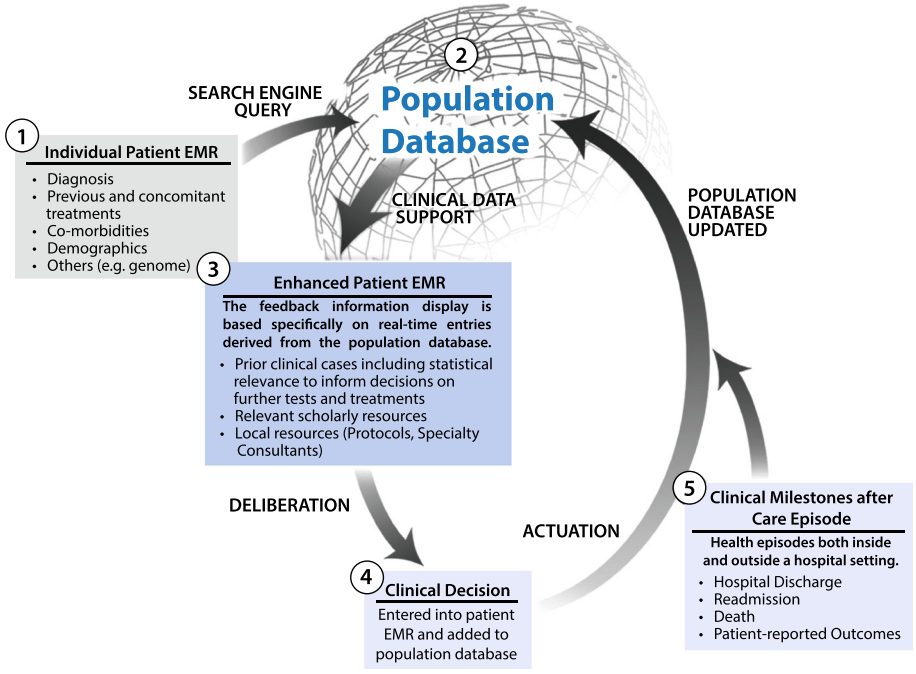

Figure 3 Dynamic clinical data mining. EMR: electronic medical record (figure courtesy of Kai-ou Tang and Edward Moseley, from [20] with permission).

excludes groups of patients, such as those with particular co-morbidities or medications, or of certain ages or ethnic groups. Thus, when trying to make a real decision under practice conditions, the RCT conclusions may simply not be applicable to the patient and situation in hand. This was the driver for the concept of DCDM in which the user of an EMR would be automatically presented with prior interventions and outcomes of similar patients to support what would otherwise be a completely subjective decision (see below).

Recent observational studies on the MIMIC ICU database have yielded many interesting findings. These include the heterogeneity of treatment effect of red blood cell (RBC) transfusion [13], the impact of pre-admission selective serotonin reuptake inhibitors on mortality in the ICU [14], the interplay between clinical notes and structured data on mortality prediction [15], optimization of heparin dosing to minimize the probability of over- and under-anticoagulation [16], long-term outcomes of minor troponin elevations in the ICU [17] and the association between serum magnesium and blood pressure in the critically ill [18], to name a few. But these observations may be specific to the Beth Israel Deaconess Medical Center and need to be validated using databases from other institutions.

Others have examined institution-specific databases, and these studies have yielded findings that have been translated into practice: A recent study at Seattle Children's compared a wide range of performance metrics and translated results into prioritized departmental and enterprise-wide improvements [19].

Celi, Zimolzak and Stone described an operational vision for a digitally based, generalized decision support system that they termed "Dynamic Clinical Data Mining" [20]. The proposed system aggregates individual patient electronic health data in the course of care; queries a universal, de-identified clinical database using modified search engine technology in real time; identifies prior cases of sufficient similarity as to be instructive to the case at hand; and populates the individual patient's EMR with pertinent decision support material such as suggested interventions and prognosis, based on prior treatments and outcomes (Figure 3).

Some of the most clear-cut arguments for Big Data in healthcare are in conjunction with the formulation of fully digitized prevention and pharmacovigilance processes [21] (Figure 4). Clinicians of the future will have to work with user friendly versions of these tools to make timely and informed decisions about the drugs their patients are receiving. In a more general sense, clinicians will have to begin to consider an individual EMR as only part of a patient's record with the remainder of the record consisting of the two-way relationship of the patient's EMR with the entire population database. The essential starting point of the individual patient can be enhanced by the knowledge present in population-level databases, and the resulting information combinations and comparisons used to make informed clinical decisions. In turn, the

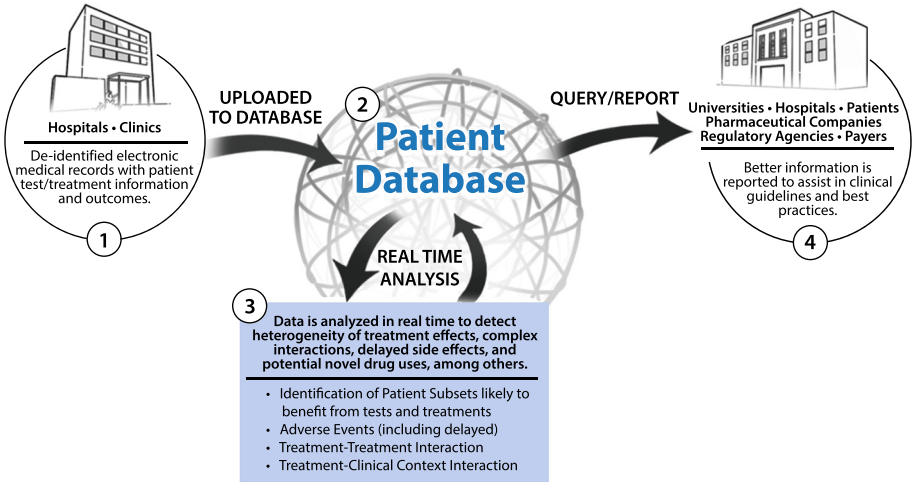

Figure 4 Clinical care optimization: a Big Data model for efficient targeting of tests and treatments and vigilance for adverse events (figure courtesy of Kai-ou Tang and Edward Moseley, from [21] with permission).

information accumulated from individuals benefits the healthcare of the entire population.

Industry is also taking note. National pharmaceutical benefits manager, Express Scripts, can predict which patients may fail to take their medication 12 months in advance, with an accuracy rate of 98% [22]; IBM is modifying their famed Watson system (in tight collaboration with clinicians) for predicting different types of cancer [23]. 23andMe's database has already been used to find unknown genetic markers for Parkinson's disease [24] and myopia [25], and their acquisition of $1.3 million in National Institute of Health funding has shown additional confidence in their goals [26].

The open data movement and medicine

More recently, the open data movement has been quietly sweeping almost every industry, including the specialized domain of healthcare. It calls for data sharing, and by its very nature, requires a degree of accountability as well as collaboration across disciplines never seen before. At the forefront of the open data movement in healthcare is the pharmaceutical industry. In October 2012, GlaxoSmithKline (GSK) announced that it would make detailed data from its clinical trials widely available to researchers outside its own walls, stunning the scientific community [27]. For a company that spends $6.5 billion a year on research and development, it was a sharp turn away from a historic system of data secrecy. In May 2013, the company began posting its own data online. It then invited others to join ClinicalStudyDataRequest.com [28], where GSK and six other drug makers have already uploaded data from nearly 900 clinical trials. The following month, the medical device company, Medtronic, teamed up with Yale University and shared its clinical trials data through the Yale University Open Access Data (YODA) Project [29].

Other important trends in open data are crowdsourcing, data marathons and hackathons, which leverage several newly available phenomena [30]. These include combining publically available, detailed, and de-identified EMRs with crowdsourcing techniques and coordinated hackathons to capture, organize and integrate stakeholder user input from a necessary variety of input sources (Figure 5). The traditional approach to knowledge discovery involves publication in peer-reviewed journals by a very circumscribed group of contributors. This process excluded a number of potentially valuable contributors, such as full time clinical physicians, nurses, medical trainees, and patients, among others.

Hackathons are large-scale events that contemporaneously bring together (physically and/or by teleconferencing) large groups of qualified individuals to collectively contribute their expertise towards a common problem set [31]. Crowdsourcing also focuses large groups of qualified individuals towards a common problem, but allows those individuals to do so asynchronously and in a mobile manner using phones, tablets, laptops and other devices to contribute from any location. With such tools, individual clinical encounters no longer have to be experienced in a silo-like fashion. The clinical 'crowd' can be leveraged to form a 'data substrate' available freely to clinicians and data scientists [4]. This amalgamation of individual knowledge should allow each clinician to address gaps in their knowledge, with the confidence that their decisions are supported by evidence in clinical practice.

In January 2014, the inaugural Critical Data Marathon and Conference was held at the Massachusetts Institute

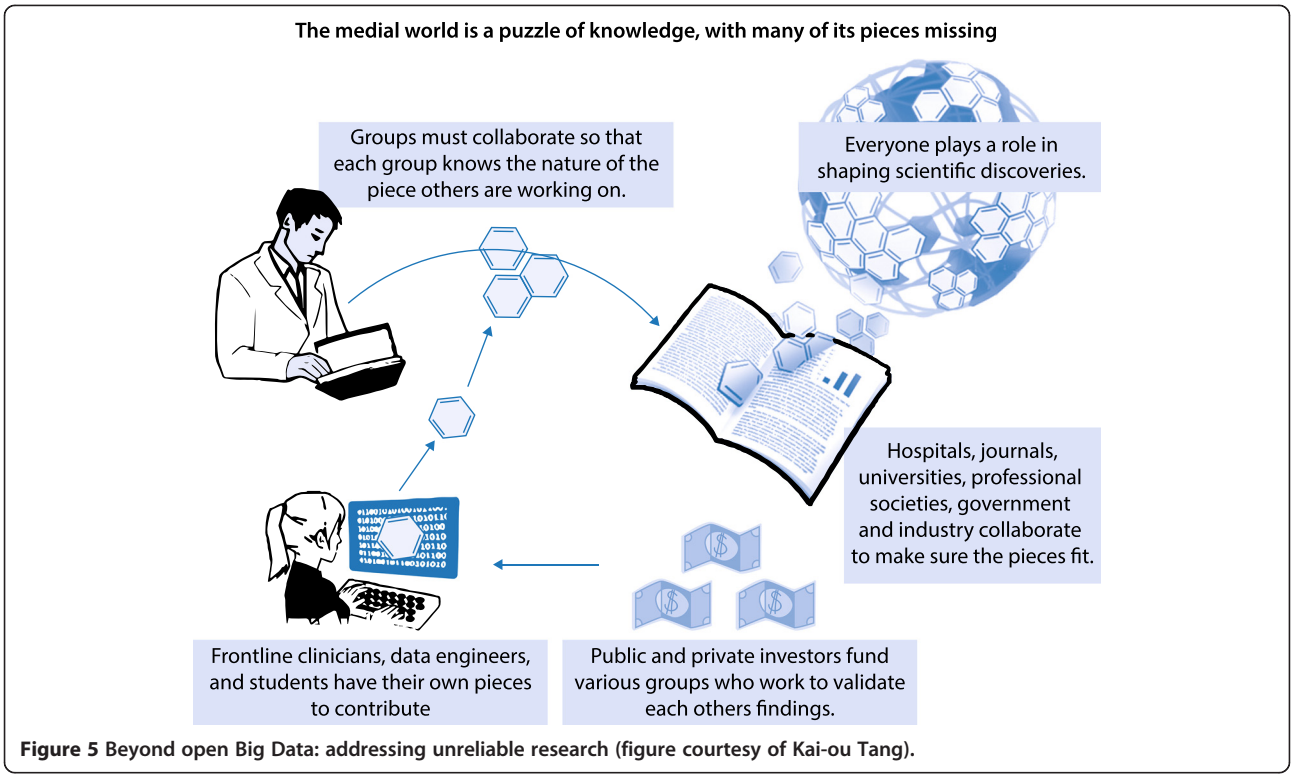

Figure 5 Beyond open Big Data: addressing unreliable research (figure courtesy of Kai-ou Tang).

of Technology [30]. In the data marathon, physicians, nurses and pharmacists were paired with data scientists and engineers, and encouraged to investigate a variety of clinical questions that arise in the ICU. Over a 2-day period, more than 150 attendees began to answer questions, such as whether acetaminophen should be used to control fever in critically ill patients, and what the optimal blood pressure goal should be among patients with severe infection. This event fostered collaboration between clinicians and data scientists that will support ongoing research in the ICU setting. The associated Critical Data Conference addressed growing concerns that Big Data will only augment the problem of unreliable research. Thought leaders from academia, government and industry across disciplines including clinical medicine, computer science, public health, informatics, biomedical research, health technology, statistics and epidemiology gathered and discussed the pitfalls and challenges of Big Data in healthcare. The consensus seemed to be that success will require systematized and fully transparent data interrogation, where data and methods are freely shared among different groups of investigators addressing the same or similar questions [30]. The added accuracy of the scientific findings is only one of the benefits of the systematization of the open data movement. Another will be the opportunity afforded to individuals of every educational level and area of expertise to contribute to science.

From a broader analysis of Big Data, we can try to understand larger patterns by comparing the strength of many signals in large populations. Larger data sets must also herald the advance of shared data sets. There is a critical need for collaborative research amongst many groups that explore similar questions. The association between data sharing and increased citation rate [32], and an increasing commitment by companies, funding agencies and investigators to more widely share clinical research data [33] point to the feasibility of this move. The prospect of using Big Data in an open environment may sound overwhelming, but there have been key steps to encourage this cultural transformation. For example, the Centers for Medicare and Medicaid Services (CMS) have begun to share data with providers and states [34]. As the largest single payer for health care in the United States, CMS has used its vast store of data to track hospital readmission rates in the Medicare program (importantly finding a rapid decline in readmission rates in 2012 and 2013), and combat Medicare fraud (in its first year the system stopped, prevented, or identified an estimated $115 million in improper payments).

As large amounts of shared data become available from different geographic and academic sources, there will be the additional benefit from the collection of data from sources with different viewpoints and biases. While individual researchers may not be aware of their own biases or assumptions that may impact reported results, shared exploration of Big Data provides us with an inherent sanity check that has been sorely lacking in many fields.

Big data *per se*

In a recent analysis of data-driven healthcare by the MIT Technology review, the authors noted that "medicine has entered its data age" [1]. Driven by the promise of an estimated $300 to $450 billion a year [35], companies of all sizes are beginning to fight in earnest to capture and tame the data explosion. Key innovations fall into three major areas: More and more data, especially resulting from mobile monitoring; better analytics using new machine learning and other techniques; and meaningful recommendations that focus on prediction, description, and prevention of poor health outcomes (that are finally captured in an easily accessible format).

The mass of new data rests primarily in the proprietary hands of large entities like insurance companies and care providers. For example, the genomics company 23andMe is famously creating a huge database of genomic data, moving from over 700,000 records towards their goal of tens of millions [26]. Some countries with centralized healthcare systems like Denmark are also beginning to leverage that accessible data [36]. In addition, smaller companies like WellDoc [37] and Ginger.io [38] are beginning to focus on rampant cell-phone penetration to get into the health-data market. Mobile phones can now seamlessly acquire daily patient metrics on meals, exercise, call patterns and other behaviors; WellDoc uses these data to recommend personalized insulin doses based on patients' daily habits, and Ginger.io monitors patients with mental illnesses for the kinds of actions that might indicate a need for help. Other companies provide physical attachments to mobile devices that enrich the possible data types available: CellScope sells an attachment to support remote otoscopy; AliveCor provides electrocardiogram (EKG) signals; Propeller Health attaches to an inhaler to record pertinent data; and there are a slew of others for nearly every imaginable data need [39].

But bigger data require better methods, and better machine learning techniques for clinical data have been a long time in coming. The most intuitive argument (that more data from which to learn cannot be worse, so must be better) is true: There have been empirical demonstrations that predictive models built from sparse, fine-grained data see marginal gains in predictive performance even to massive scale [40]. But there is another less intuitive argument for bigger data: Certain rare trends or behaviors simply may not be observed in sufficient numbers without employing Big Data. Dubbed the 'heavy tail' of data, these rare behaviors are even more difficult to observe as we add more features to our datasets. Intuitively, we can think of datasets as a set of samples out of a larger space; for example, a circle inscribed within a square gets most of the area, leaving only the corners out. But as we move from inscribing a circle within a square, to inscribing a sphere within a cube, the ratio of space in the corners increases [41] (Figure 6). Repeat this to a higher dimension and most of the volume of the cube will be concentrated in its (many) corners. But it is these rare instances (sometimes appropriately referred to as 'corner cases') of behaviors or patient characteristics that machine learning cannot reliably analyze with historically available data sample sizes. The Big Data explosion is finally offering data at a scale large enough to overcome the risks of higher-dimensional spaces when working with healthcare data issues.

Along with Big Data's promise, there have been warnings of over confidence and disaster, labelled by Lazer et al. as "Big Data hubris" [42]. The warning parable told to illustrate this is Google's "Flu Trends" [43]. In 2008, Google launched its Flu Trends, which used the search terms typed into Google to track the progression of influenza epidemics over time. However, this approach was subsequently revealed to have suffered from several known data analysis pitfalls (e. g., overfitting and concept drift) so that by 2012–2013, the prevalence of flu was being greatly overestimated. Other oft-cited risks include misleading conclusions derived from spurious associations in increasingly detailed data, and biased collection of data that may make derived hypotheses difficult to validate or generalize [44].

But avoiding spurious conclusions from data analysis is not a challenge unique to Big Data. A 2012 *Nature* review of cancer research found reproducibility of findings in only 11% of 53 published papers [45]. There is concern that Big Data will only augment this noise, but using larger datasets actually tends to help with inflated

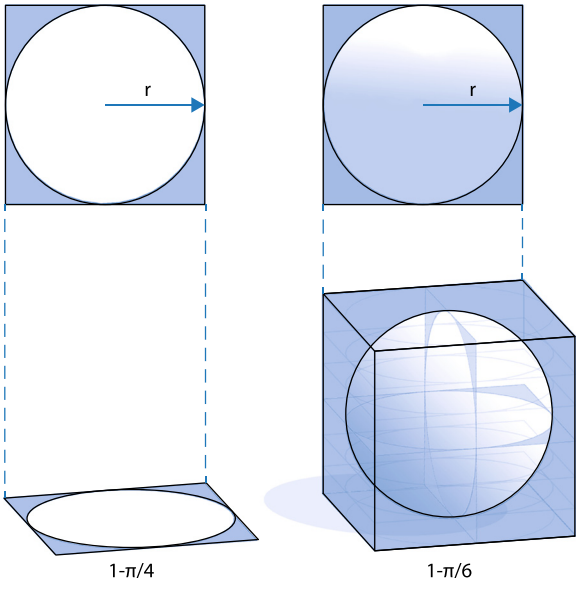

Figure 6 The data space and corner cases (figure courtesy of Yuan Lai).

significance, as the estimated effect sizes tend to be much smaller [46].

The biased collection of data is a non-trivial question. If researchers have large amounts of data that severely oversample certain populations or conditions, their derived hypotheses can be incorrect or at least understandably difficult to validate. The way that current literature is designed, generated, and published creates sequential 'statistically significant' discoveries from restricted datasets. It is not uncommon in the scientific literature to get a different story for a variable's (vitamin E, omega-3, coffee) relationship to outcome (mortality, Alzheimer's, infant birth-weight) depending on what is adjusted for, or how a population was selected. There is little meaning to exploring the impact of one variable for one outcome: it is the big picture that is meaningful.

Conclusion

The benefits of the data explosion far outweigh the risks for the careful researcher. As target populations subdivide along combinations of comorbid conditions and countless genetic polymorphisms, as diagnostic and monitoring device including wearable sensors become more ubiquitous, and as therapeutic options expand beyond the evaluation of individual interventions including drugs and procedures, it is clear that the traditional approach to knowledge discovery cannot scale to match the exponential growth of medical complexity.

Rather than taking turns hyping and disparaging Big Data, we need organizations and researchers to create methods and processes that address some of our most pressing concerns, e. g., who is in 'charge' of shared data, who 'owns' clinical data, and how do we best combine heterogeneous and superficially non-interoperable data sources? We need to use Big Data in a different way than we have traditionally used data – collaboratively. By creating a culture of transparency and reproducibility, we can turn the hype over Big Data into big findings.

Abbreviations

APACHE: Acute physiology and chronic health evaluation; AUC: Area under the curve; CMS: Centers for medicare and medicaid services; CPR: Cardiopulmonary resuscitation; DCDM: Dynamic clinical data mining; EKG: Electrocardiogram; EMR: Electronic medical record; ICU: Intensive care unit; LOS: Length of stay; MPM: Mortality probability model; RBC: Red blood cell; RCT: Randomized controlled trial; SAPS: Simplified acute physiology score.

Competing interests

LC's laboratory has received funding from SAP and Philips Healthcare; DJS was an employee of Philips Healthcare until March 2012.

Declarations

The article processing fee was funded by the National Institute of Health.

Author details

[1]Department of Computer Science and Electrical Engineering, Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Laboratory, Cambridge, USA. [2]Beth Israel Deaconess Medical Center, Harvard-MIT Division of Health Science and Technology, Division of Pulmonary, Critical Care and Sleep Medicine, Cambridge, USA. [3]Departments of Anesthesiology and Neurological Surgery, and the Center for Wireless Health, University of Virginia, Charlottesville, USA.

Published online: 16 March 2015

References

1. MIT editors. Business Report: Data-driven Health Care. MIT Technol Rev. 2014;117:1–19.
2. Celi LA, Csete M, Stone D. Optimal data systems: the future of clinical predictions and decision support. Curr Opin Crit Care. 2014;20:573–80.
3. Vincent JL. Critical care-where have we been and where are we going? Crit Care. 2013;17:S2.
4. Celi LA, Mark RG, Stone DJ, Montgomery RA. "Big data" in the intensive care unit. Closing the data loop. Am J Respir Crit Care Med. 2013;187:1157.
5. Breslow MJ, Badawi O. Severity scoring in the critically ill: Part 2 -Maximizing value from outcome prediction scoring systems. Chest. 2012;141:518–27.
6. Breslow MJ, Badawi O. Severity scoring in the critically ill: Part 1 – Interpretation and accuracy of outcome prediction scoring systems. Chest. 2012;141:245–52.
7. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big Data In health care: Using analytics to identify and manage high-risk and high-cost patients. Health Aff (Millwood). 2014;33:1123–31.
8. APACHE Outcomes. Available at: https://www.cerner.com/Solutions/Hospitals_and_Health_Systems/Critical_Care/APACHE_Outcomes/. Accessed Nov 2014.
9. McShea M, Holl R, Badawi O, Riker R, Silfen E. The eICU research institute – a collaboration between industry, health-care providers, and academia. IEEE Eng Med Biol Mag. 2010;29:18–25.
10. Saeed M, Villarroel M, Reisner AT, et al. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): a public-access intensive care unit database. Crit Care Med. 2011;39:952.
11. Smith M, Saunders R, Stuckhardt L, McGinnis JM, Committee on the Learning Health Care System in America, Institute of Medicine. Best Care At Lower Cost: The Path To Continuously Learning Health Care In America. Washington: National Academies Press; 2013.
12. Mills EJ, Thorlund K, Ioannidis J. Demystifying trial networks and network meta-analysis. BMJ. 2013;346:f2914.
13. Dejam A, Malley BE, Feng M, et al. The effect of age and clinical circumstances on the outcome of red blood cell transfusion in the critically ill patients. Crit Care. 2014;18:487.
14. Ghassemi M, Marshall J, Singh N, Stone DJ, Celi LA. Leveraging a critical care database: selective serotonin reuptake inhibitor use prior to ICU admission is associated with increased hospital mortality. Chest. 2014;145:745–52.
15. Ghassemi M, Naumann T, Doshi-Velez F, et al. Unfolding physiological state: Mortality modelling in intensive care units. KDD. 2014;2014:75–84.
16. Ghassemi MM, Richter SE, Eche IM, Chen TW, Danziger J, Celi LA. A data-driven approach to optimized medication dosing: a focus on heparin. Intensive Care Med. 2014;40:1332–9.
17. Velasquez A, Ghassemi M, Szolovits P, et al. Long-term outcomes of minor troponin elevations in the intensive care unit. Anaesth Int Care. 2014;42:356–64.
18. Celi LA, Scott DJ, Lee J, et al. Association of hypermagnesemia and blood pressure in the critically ill. J Hypertension. 2013;31:2136–41.
19. Kolker E, Kolker E. Healthcare analytics: Creating a prioritized improvement system with performance benchmarking. Big Data. 2014;2:50–4.
20. Celi LA, Zimolzak AJ, Stone DJ. Dynamic clinical data mining: search engine-based decision support. JMIR Med Inform. 2014;2:e13.
21. Celi LA, Moseley E, Moses C, et al. From pharmacovigilance to clinical care optimization. Big Data. 2014;2:1–8.
22. The Runaway Cost of Diabetes. Available from: http://lab.express-scripts.com/insights/drug-options/the-runaway-cost-of-diabetes. Accessed Sept 2014.
23. Edwards C. Using patient data for personalized cancer treatments. Commun ACM. 2014;57:13–5.
24. Do CB, Tung JY, Dorfman E, et al. Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. PLoS Gen. 2011;7:e1002141.
25. Kiefer AK, Tung JY, Do CB, et al. Genome-wide analysis points to roles for extracellular matrix remodeling, the visual cycle, and neuronal development in myopia. PLoS Gen. 2013;9:e1003299.

26. 23andMe Scientists Receive Approximately $1.4 Million in Funding from the National Institutes of Health. http://mediacenter.23andme.com/press-releases/nih_grant_2014/. Accessed Sept 2014.
27. GSK announces further initiatives to advance openness and collaboration to help tackle global health challenges. Available from: http://us.gsk.com/en-us/media/pressreleases/2012/gsk-announces-further-initiatives-to-advance-openness-and-collaboration-tohelp-tackle-global-health-challenges. Accessed Sept 2014.
28. Clinical Study Data Request Site. Available from: https://clinicalstudydata request.com/ (accessed Nov 2014); 2014.
29. Krumholz HM, Ross JS, Gross CP, et al. A historic moment for open science: the Yale University Open Data Access Project and Medtronic. Ann Intern Med. 2013;158:910–1.
30. Badawi O, Brennan T, Celi LA, et al. Making big data useful for health care: a summary of the inaugural mit critical data conference. JMIR Med Inform. 2014;2:e22.
31. Celi LA, Ippolito A, Montgomery RA, Moses C, Stone DJ. Crowdsourcing knowledge discovery and innovations in medicine. J Med Internet Res. 2014;16:216.
32. Piwowar HA, Day RS, Fridsma DB. Sharing detailed research data is associated with increased citation rate. PloS One. 2007;2:e308.
33. Wilhelm EE, Oster E, Shoulson I. Approaches and Costs for Sharing Clinical Research Data. JAMA. 2014;311:1201–2.
34. Brennan N, Oelschlaeger A, Cox C, Tavenner M. Leveraging the big-data revolution: CMS is expanding capabilities to spur health system transformation. Health Affairs. 2014;33:1195–202.
35. Kayyali B, Knott D, Van Kuiken S. The big-data revolution in US health care: Accelerating value and innovation. McKinsey & Company. http://www.mckinsey.com/insights/health_systems_and_services/the_big-data_revolution_in_us_health_care. Accessed Nov 2014; 2013.
36. Saunders MK. In Denmark, big data goes to work. Health Affairs. 2014;33:1245–5.
37. Quinn CC, Clough SS, Minor JM, Lender D, Okafor MC, Gruber-Baldini A. WellDoc™ mobile diabetes management randomized controlled trial: change in clinical and behavioral outcomes and patient and physician satisfaction. Diabetes Technol Ther. 2008;10:160–8.
38. Giles J. Computational social science: Making the links. Nature. 2012;488:448–50.
39. M Health Health and appiness. The Economist Magazine. http://www.economist.com/news/business/21595461-those-pouring-money-health-related-mobile-gadgets-and-apps-believe-they-can-work (Created Feb 1, 2014). Accessed Nov 2014.
40. Junqué de Fortuny E, Martens D, Provost F. Predictive modeling with big data: is bigger really better? Big Data. 2013;1:215–26.
41. Bishop CM. Pattern Recognition And Machine Learning. New York: Springer; 2006. p. 740.
42. Lazer D, Kennedy R, King G, Vespignani A. The parable of Google flu: traps in big data analysis. Science. 2014;343:1203–5.
43. Butler D. When Google got flu wrong. Nature. 2013;494:155.
44. Harford T. Big Data: are we making a big mistake. Financial Times Magazine. http://www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html#axzz3TDz4MSnF. Accessed Nov 2014; 2014.
45. Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. Nature. 2012;483:531–3.
46. Ioannidis JP, Panagiotou OA. Comparison of effect sizes associated with biomarkers reported in highly cited individual articles and in subsequent meta-analyses. JAMA. 2011;305:2200–10.
47. Mayaud L. Prediction of mortality in septic patients with hypotension. PhD Thesis, Oxford University; 2014