Published in final edited form as:
Health Aff (Millwood). 2014 July ; 33(7): 1163–1170. doi:10.1377/hlthaff.2014.0053.

Big Data And New Knowledge In Medicine: The Thinking, Training, And Tools Needed For A Learning Health System

Harlan M. Krumholz
Harold H. Hines Jr. professor of Medicine and Epidemiology and Public Health at the Yale University School of Medicine, in New Haven, Connecticut

Abstract

Big data in medicine--massive quantities of health care data accumulating from patients and populations and the advanced analytics that can give it meaning--hold the prospect of becoming an engine for the knowledge generation that is necessary to address the extensive unmet information needs of patients, clinicians, administrators, researchers, and health policy makers. This paper explores the ways in which big data can be harnessed to advance prediction, performance, discovery, and comparative effectiveness research to address the complexity of patients, populations, and organizations. Incorporating big data and next-generation analytics into clinical and population health research and practice will require not only new data sources but also new thinking, training, and tools. Adequately used, these reservoirs of data can be a practically inexhaustible source of knowledge to fuel a learning health care system.

Netflix, the popular entertainment company, is known for making useful movie suggestions to its customers. In 2006, the company embarked on a project to further improve its ability to predict which movies its customers would like.[1] Through an open competition, Netflix offered a $1 million prize to the group that most improved on Netflix's traditional approach, which was based on conventional statistics.

The Netflix strategy for improving service was interesting, in part, for what it did not do. Netflix did not hire psychologists to develop conceptual models of the factors that influence an individual's viewing experience. It did not test hypotheses about the theory of choice or the determinants of genre preference. It did not perform randomized controlled trials to compare ways of presenting information to customers.

Instead, Netflix chose to exploit its data. Netflix provided competitors for the prize with 100 million ratings submitted on almost 18,000 movie titles by almost 500,000 people. The winning teams not only focused on how each person rated movies, but also importantly discovered that an individual's ratings were influenced by factors such as whether the person ranks many movies at a time (which tended to accentuate positive or negative preferences) or by the overall popularity of a movie across raters at a particular point in time. Ultimately, the winners produced an algorithm that increased accuracy of predicting ratings by 10 percent.

The Netflix competition exemplifies a data-driven approach that is emerging from a new era of big data. Big data has been described as the rapidly increasing size of available data, the speed with which it is produced, and the ways in which it is represented.[2] It also can refer not only to the data but to the possibilities of discovering new knowledge by leveraging massive data collections in novel ways. The analytic methods for big data typically depart from traditional statistics and hypothesis testing, and incorporate techniques such as machine learning, a form of artificial intelligence that employs advanced mathematical and computational systems to reveal information from the data, commonly for the purpose of prediction and discovery.

Learning from large, complex data is becoming routine in commercial enterprises. Many companies use immense quantities of information available to them from purchases, searches, and social media to model consumer behavior. Advanced data management and analysis is exemplified by leading information companies, including Google, Facebook, and Amazon,[3] and government organizations, such as the National Security Agency (NSA).[4]

Google leverages the information from trillions of individual pages on the Internet and develops programs and formulas to produce search results that match, within seconds, its users' needs.[5] Amazon not only employs its data collections to accurately suggest products to its customers, but has improved its predictive analytics to the point that it now has a patent for anticipatory shipping, a method that one day could lead to the shipping of products customers are expected to buy based on previous orders and other factors.[6] The NSA is creating methods to provide real-time analyses to rapidly characterize and assess social communications, with the ability to handle trillions of connections.[4] The NSA, Google, Facebook, and others exemplify the use of new approaches such as graph analyses, a method of portraying data in three-dimensional space and as nodes and edges rather than rows and columns. The analysis of this web of information can highlight relationships and the structure of associations, such as those that exist in social networks.[7] In medicine, such methods may improve classification of disease, reveal ways to determine the influence of particular physicians on practice patterns, or predict a patient's clinical events.[8,9]

This article discusses the need for the clinical research enterprise to expand its approaches to generating new clinical and population health knowledge. The approaches, which will involve systematically collecting and harvesting big data from many different sources, will require new thinking, new training, and new methods and tools.

The Stakes

The current medical research enterprise cannot keep pace with the information needs of patients, clinicians, administrators, and policy makers. The flow of new knowledge is too slow, and its scope is too narrow. The medical research community's delay in adopting big data approaches has left it particularly ill prepared for a precision medicine future that is designed to provide personalized information and individualized care.[10]Medicine aspires to a learning health care system, but is failing to rapidly learn from the data being generated through the course of clinical care.

For anyone who is ill and for those providing care to them, the uncertainties almost always outweigh what is known. Patients and clinicians often have too little information to determine in specific instances which strategies to avoid because they fail to provide benefits, and which to embrace because they are most likely to produce outcomes that the patient values. When evidence does exist for a particular decision, it often is not applicable to the person in need. Moreover, the data generated in every day medical practice is largely wasted. To better inform decisions, we urgently need better personalized predictions about prognosis and response to treatments; a deeper understanding of the complex factors and their interactions that influence health at the level of the patient, the health system, and society; enhanced approaches to detecting safety problems with drugs and devices; and more effective methods of comparing prevention, diagnostic, and treatment options.

Medical practice and clinical research are still largely anchored in producing new knowledge through studies that tend to narrow the research question and avoid complexities of real-world practice. Clinical trials, for instance, often exclude complicated patients--those who may have several medical ailments and complex regimens--who are typical of patients that are seen in medical offices. These studies are most commonly focused on a single question, are commonly expensive, and take years to complete. Moreover, most studies are poorly equipped to explore how various factors may interact to influence the result for a particular patient. Meanwhile, data generated every day, for a variety of practical purposes, could serve as a practically inexhaustible source of knowledge to fuel a learning health care system. However, to date, these data are largely wasted as a source of research and rarely investigated, in the course of medical research, with big data analytics.

Despite the potential of big data approaches, the progress in medical research methods pertinent to big data has largely been sequestered in the more basic sciences such as bioinformatics, which has a vibrant community of researchers who employ computers as their laboratories and elevate data science to a foundational skill.[11] They apply novel approaches to massive troves of biological data and, as early adopters of open science, share data assets and provide ample opportunities for cross-pollination of ideas and techniques. Until medicine develops a robust clinical research community that embraces these contemporary opportunities and fully realizes the promise of big data, the large gap between the available evidence, what is needed and what could be generated will persist.

New Thinking

The integration of new approaches will require new thinking on the part of medical authorities regarding the ways in which this type of health and health care research can best contribute to the productivity of the research enterprise.[12–14] Clinical medical research has often subordinated insights that are derived empirically from existing data to those that are based on theory and experiment. Studies that start with exploiting data are often considered inferior to those that first formulate a hypothesis stemming from a presumptive understanding of mechanism. However, many types of research questions can be addressed before understanding the direct cause of disease – or can even provide insights that eventually lead to an understanding of mechanism.

For example, researchers can use approaches that are designed to reveal clusters of patient groups that might suggest new taxonomies of disease based on how similar they are according to a broad range of characteristics, including outcomes.[10] It may be, for instance, that based on biological, clinical, behavioral, and outcomes data there are many more types of diabetes than previously appreciated. The empirical classification could be shown to have value in selecting treatment strategies and predicting outcomes. This knowledge can be useful even in advance of understanding the underlying mechanisms of disease and response to therapy. In fact, in medicine there is precedent for discovering effective therapies (e.g., aspirin) before knowing why it produced a benefit.

Advances in prediction are possible simply by learning from the data and creating approaches that are highly reproducible and consistently perform well in different settings and with different patients. Amazon can predict customer preferences without knowing why customers have those preferences--and the prediction is useful even without a deep knowledge of why patients have those preferences. Similarly, using methods of signal detection, data-driven research may identify better ways to detect safety problems with drugs and devices prior to understanding the underlying cause.[15] The same may be true with predicting epidemics.

New big data methods can turbocharge powers of observation in health care. In the same way the microscope enhanced eyesight, sophisticated mathematical and computational approaches can augment what can be “seen” and understood from massive amounts of data.

Inductive Reasoning

The new way of thinking can embrace inductive reasoning and pattern recognition on an equal basis with deductive reasoning.[16] Much of contemporary medical research involves deductive reasoning, which starts with a general theory and a hypothesis evolving from it, and pursues studies to test conceptual models. Inductive reasoning and pattern recognition, in contrast, begin with observations and builds to specific conceptual models or creates tools that have utility in informing decisions. They key is to test the consistency of the results and ensure the validity of the conclusions.

Many medical research leaders have legitimately been skeptical of empiric work preceding theory due to the possibility of false positive conclusions. For prediction, discovery and signal detection, the value of inductive reasoning and pattern recognition can be self-evident from the value of the information that is generated. If a new method of prediction is better than the current method, then the results should allay concerns about the provenance of the method.

The concern about inductive reasoning is particularly important in settings seeking to infer causal relationships. False positive findings from investigations into genomic associations that started with the data are indeed an example of the hazard of pursuing knowledge about causation without theory.[17] This issue in not solely about big data approaches, though the multiple comparisons in big data approaches may accentuate the risk of false positive conclusions. And yet, experts are not precluding the possibility that big data approaches can ultimately assist in revealing causal relationships, which would enable the comparison of the effectiveness of therapies in real-world use across a diversity of patients that characterize usual medical practice.[18] Nevertheless, the inductive process, for the purpose of understanding causation, is often less certain and conclusions are often expressed in terms of confidence levels rather than as definitive conclusions. Findings can make us more or less confident about whether one factor causes another. For example, no one would conduct a human experiment to test whether smoking causes cancer, but the empiric data offers a strong confidence in the causal relationship. Scientists have developed and refined criteria to help with these types of inferences.[19] We will need to develop criteria that govern the ways to interpret results from millions and trillions of observations that are relevant to decision-making at the bedside. The validation of these results will be critical to instill confidence in their utility in studies the involve causation, such as comparative effectiveness studies.

Complexity

Another issue is the need for new thinking about the importance of research that can account for the complexity of patients and medical decision making. The current paradigm and the available methods often involve a reductionist framework that ultimately fails to provide information that is primed for the complexities of patients and medical practice. Reductionism states that a system can be defined by the sum of its parts and described by its individual components--an approach that likely cannot capture the complexity of the human body, disease, and health care delivery systems. Complex interactions can lead to emergent phenomena that cannot be predicted directly and result from many related factors. The prototypical example is the snowflake, whose appearance cannot be directly predicted by the temperature or clouds--but is a result of the interactions of those influences and others.

Although this broader type of thinking is prevalent in systems biology, it has not yet entered mainstream clinical medical research. The convergence could occur through the vision promulgated by Leroy Hood, a pioneer in molecular biology and systems biology, who has written that patient-activated social networks, big data, and systems medicine, in concert with advanced analytics, is leading to “medicine that is predictive, preventive, personalized, and participatory.”[20] Big data approaches can retain the complexity of the data and illuminate the ways that biological, demographic, clinical and environmental factors interact with each other to influence risk and outcomes.

Machine Learning And Other Advanced Analytic Techniques

The advances in big data will require openness by researchers, funders, and end-users to employing machine learning, data mining, and machine-based algorithms. These techniques, which are already prevalent in biological sciences, use computer programs that help scientists reveal patterns and relationships that might not otherwise be appreciated or anticipated. The researcher can search for patterns without knowing what may emerge. This approach is very different from developing a research project around a specific question and requires stringent methods to validate findings to ensure they are not occurring by chance.

Such investigations are common in other fields, such as astronomy. Joe Bredekamp, senior science program executive at NASA Science Mission Directorate, has written that “The scientific discovery process [in astronomy] is increasingly dependent on the ability to analyze massive amounts of complex data generated by scientific instruments and simulations. Analysis is rapidly becoming the bottleneck, if not chokepoint, in the process. This situation has motivated needs for innovation and for fostering of collaborations with the computer science and technology community to bring advances in those fields to bear on the scientific investigations."[21] Medicine is facing these same challenges and is slowing responding by placing an emphasis on ways to increase the speed and fidelity of what can be learned from data.

Machine learning can also circumvent the confirmation bias that can contaminate investigations directed by scientists with strong, pre-existing ideas. Machine learning can provide input similar to that of a truly independent expert. The results of these types of analyses still need to be evaluated and interpreted by scientists with content expertise in medicine, but they can accelerate the ability to generate useful knowledge.

Data

Another area that needs new thinking surrounds the imperative to use data generated from everyday life. An aspiration of big data practitioners is to leverage data that has low costs of production and is readily available. For example, the vast amount of potential information that is continuously generated from patient experience with health and health care remains untapped. A premise of big data is that those daily interactions, captured through medical encounters, health behaviors, and other data produced for a variety of purposes can be the source material for vast amounts of medical research that can ultimately meet the needs of future patients, clinicians, and other health care professionals. There is the potential to learn from each patient, but only if there is a commitment to making the data available and organizing it appropriately. The new thinking has to embrace the importance of not wasting this potential source of knowledge. The essence of a truly learning health care system will be to learn from its daily experience.

New Training

Medicine is an information profession, and the underlying basis of investigation must increasingly include data science. There is a need to invest in strengthening the skills of clinical investigators, very few of whom are thoroughly trained in data science. New terms such as Hadoop (a programming framework that supports the processing of massive data across many computers), unsupervised learning (analyses that seeks to find hidden patterns within the data), graph analytics (analyses that use graphs to understand relationships and patterns), and natural language processing (analyses enabling computers to derive meaning from human language and thus to extract knowledge from documents) will need to become part of the research lexicon. Such terms will need to become a part of medical curricula, including those used in distance learning and the discourse of knowledge communities.

Informatics skills will also be necessary, given the need to understand how to best generate the reagents (the data) for this research. Curricula for investigators will need to evolve to produce the human capital and cross-field collaboration necessary to pursue this work. New academic tracks will need to be considered as well as new funding mechanisms for previously unconventional research and the integration of new types of expertise into clinical departments. In addition, clinicians' mindsets must change such that they are more comfortable with the evidence that is generated from these new approaches.

New Methods And Tools

Even with new thinking, implementation will not be possible without practical methods and tools, customized for issues germane to medicine. Conventional methods have been very useful, but researchers often contort their questions to fit methods that may not adequately accommodate complexity. New methods of classification, methods of describing disease that go beyond the diagnostic labels we have historically inherited, that are better suited to incorporating this complexity are needed and could allow interventions to be customized for each group. Variations of these methods abound in other fields, but are conspicuously absent in mainstream medical research.

Beyond the development of these tools, health care leaders need to direct attention toward implementation. The new tools need to not only incorporate novel approaches, but also have the capacity for easy and useful application. A path toward achieving this might be through a research commons where massive data assets can reside and where investigators can share and refine tools to interpret them, as has been done in other fields by services such as ZENODO,[22] created by the European Commission's Open AIREplus project. Similar initiatives are being developed by Optum Labs and the Health Care Cost Institute, and there will likely be more over time.[23,24]

These analytic methods and tools may depend on a variety of approaches including geometric data organization and visualization, analytic algorithms, Bayesian networks and graphical models, spatiotemporal analytics, and high-dimensional modeling and interference. These words and concepts, representing advanced analytic strategies, will be unfamiliar to most people involved in clinical research, and their widespread use will require training and incorporation into standard software packages.

The importance of methods and tools to analyze these massive data collections was highlighted in *Frontiers in Massive Data Analysis*, a National Research Council (NRC) of the National Academies Press report published in 2013.[18] The report emphasizes this urgent need and draws attention to complex interactions and causal relationships that might emerge from big data efforts. In particular, it describes the necessity of combining algorithmic (mathematical) and inferential (statistical) perspectives. According to the report, "Harnessing massive data to support causal inference represents a central scientific challenge" and, to date, "causal modeling in massive data has attracted little attention." This guidance has relevance to other approaches to interpreting data as it addresses scalability, complex interactions, and the spatiotemporal characteristics of the data. The document also highlights the need to develop methods that can be used to draw conclusions about causation, which signals the NRC's recognition that such methods may be useful for comparative effectiveness studies.

Challenges

Data assets are growing, but major gaps remain in the quality and quantity of data. In addition, privacy issues mandate the need to balance security of the data with the desire to share them. The imperative is to find ways to safely share data because open science and data sharing are essential to providing the opportunity for replication.[25,26]

Validation of findings becomes an essential component of the research as broad-based investigations include many different analyses. Perhaps the biggest impediment is a medical research culture that has demonstrated little interest in the adoption of new methods.

Insularity of research teams also slows progress. This type of work requires interdisciplinary collaboration and deep investments in learning languages and customs across disciplines.

What The Future Could Hold

Big heterogeneous data have overwhelmed the human researcher’s intuitive ability to see correlations using classical statistical methods. Thus, there is a vital need for algorithms that enhance medical practice and public health – and help researchers discover important relationships. The challenge is to develop algorithms such as those that promote an understanding about which of the hundreds or thousands of dimensions of data are correlated unexpectedly, which are correlated trivially, and which are not correlated. For example, in understanding the factors that are related to a patient’s recovery after surgery, advanced algorithms can produce predictions of the results, incorporating biological, clinical, demographic, and psychosocial information in addition to physician, hospital, health care system, and geographic factors.

In addition, many of these factors are connected to and interact with other factors in ways that are poorly understood. Moreover, recovery can be described in different ways, based on symptoms, function, resource use, and survival. When a patient has an adverse outcome, which of the many dimensions of data are important and which can be ignored? In the past, researchers would seek to identify a parsimonious set of variables, well chosen for their likely relationships with what is being studied, and minimize the complexity to isolate a single effect--as if health and disease could be understood through the illumination of single effects.

Novel approaches that leverage these data assets could initiate an era of remarkable new insights that transform medicine. They could extend the ability to pursue classes of problems such as prediction and signal detection. Such information could allow the preemption of health care problems and properly target interventions. In the same way that Amazon does anticipatory shipping, health providers could do anticipatory interventions as the data begin to indicate that patient risk is rising.

The potential of big heterogeneous data to help understand causal relationships should also be considered. For example, research could support studies to compare outcomes of different treatments. Better characterization of patient profiles could elevate the ability to control for factors that can confound studies and lead to false conclusions. With better profiling of individuals, it would be possible to match patients who are and are not treated in certain ways to determine the effect of specific strategies. Given that it is not possible to conduct enough trials to cover all the types of patients, methods that leverage actual experience could be critical to producing evidence applicable to the full range of patients.

In our research team's own research, we are discovering the value of using machine learning techniques. For example, we have used these methods to phenotype hospitals, defining groups that have similar profiles and performance. In one study, we evaluated the way that hospitals abandoned an expensive medication that was shown to be potentially harmful and found very different patterns among hospitals, with some groups characterized by rapid change in practice in response to new information about the drug and others changing more slowly.[27] The insights about the hospitals were only possible with comprehensive data and advanced analytics that revealed similarities among hospitals that otherwise would not appear to be alike.

We also used massive data collections from a large number of hospitals and employed automated routines to classify hospitals based jointly on their financial and clinical performance.[28] We determined value phenotypes among groups of hospitals, exposing similarities among institutions that, on the surface, did not appear to have much in common.[27,28] The next step is to understand what hidden factors influenced the similar performance. This type of analysis opens the way for investigations of the determinants of these differences, perhaps leveraging methods that have illuminated factors that produce positive deviance and ultimately developing interventions to improve performance.

These types of investigations are amenable to queries and may better exist in interactive formats. In the future, the products of scientific inquiry may evolve from a static journal publication to a more dynamic platform for presenting and updating results. This type of interactive platform could enhance the relevance of the research to decision makers who want to know how the results change in response to changes in some parameters, as they seek to customize the information to their needs.

To achieve this future, there is a need for initiatives, such as the big data to knowledge centers being funded by the National Institutes of Health,[29] to provide funding for interdisciplinary teams of researchers, in collaboration with patients and other stakeholders, to produce new methods and tools. Progress will occur rapidly if the value of the information from the new research is clear, the products improve the health of patients, and perhaps even improve the value of health care.

Conclusion

This is a historic moment in medicine. There is a remarkable opportunity to promote medicine as an information science and strengthen the foundations of a learning health care system, defined by the Institute of Medicine as one "designed to generate and apply the best evidence for the collaborative healthcare choices of each patient and provider; to drive the process of discovery as a natural outgrowth of patient care; and to ensure innovation, quality, safety, and value in health care."

To achieve this vision, it is imperative to fundamentally augment the ability to learn from experience and produce knowledge with the speed, efficiency, relevance, and quality necessary to fulfill the needs of patients, clinicians, researchers, policy makers, and health care systems. Massive repositories of potential knowledge, populated by data from health care visits, devices, administrative claims, and biospecimens, are increasingly available.

The promise of massive data assets lies not merely in their size, but in the way they are used. For the clinical medical research enterprise to achieve its potential, it needs to catch up with the world around it and reflect the complexity within it. Adequately used, these reservoirs of data can be a practically inexhaustible source of knowledge to fuel a learning health care system.

Acknowledgments

Harlan Krumholz is supported by Grant No. U01 HL105270-04 (Center for Cardiovascular Outcomes Research at Yale University) from the National Heart, Lung, and Blood Institute. Krumholz is a recipient of research grants from Medtronic and from Johnson and Johnson, through Yale University, to develop methods of clinical trial data sharing and to chair a cardiac scientific advisory board for UnitedHealth.

Notes

1. Netflix. Netflix prize [Internet]. Los Gatos (CA): Netflix; 2009 Sep. [cited 2014 May 28]. Available from: http://www.netflixprize.com/
2. Laney, D. Deja vvvu: others claiming Gartner’s construct for big data [blog on the Internet]. 2012 Jan 14. [cited 2014 May 28]. Available from: http://blogs.gartner.com/doug-laney/deja-vvvue-others-claiming-gartners-volume-velocity-variety-construct-for-big-data/
3. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., et al. Big data: the next frontier for innovation, competition, and productivity [Internet]. New York (NY): McKinsey and Company; 2011 May. [cited 2014 May 28]. Available from: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation
4. Marko, K. The NSA and big data: what IT can learn [Internet]. InformationWeek. 2013 Jul 18. [cited 2014 May 28]. Available from: http://www.informationweek.com/big-data/big-data-analytics/the-nsa-and-big-data-what-it-can-learn/d/d-id/1110818?
5. Google. How search works [Internet]. Mountain View (CA): Google; [cited 2014 May 28]. Available from: http://www.google.com/insidesearch/howsearchworks/
6. Bensinger, G. Amazon wants to ship your package before you buy it [blog on the Internet]. New York (NY): Wall Street Journal; 2014 Jan 17. [cited 2014 May 28]. Available from: http://blogs.wsj.com/digits/2014/01/17/amazon-wants-to-ship-your-package-before-you-buy-it/
7. Lazer D, Pentland A, Adamic L, Aral S, Barabasi AL, Brewer D, et al. Social science. Computational social science. Science. 2009; 323(5915):721–3. [PubMed: 19197046]
8. Barnett ML, Christakis NA, O’Malley J, Onnela JP, Keating NL, Landon BE. Physician patient-sharing networks and the cost and intensity of care in US hospitals. Med Care. 2012; 50(2):152–60. [PubMed: 22249922]
9. Barnett ML, Landon BE, O’Malley AJ, Keating NL, Christakis NA. Mapping physician networks with self-reported and administrative data. Health Serv Res. 2011; 46(5):1592–609. [PubMed: 21521213]
10. National Research Council. Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease. Washington (DC): National Academies Press; 2011.
11. Yip KY, Cheng C, Gerstein M. Machine learning and genome annotation: a match meant to be? Genome Biol. 2013; 14(5):205. [PubMed: 23731483]

12. Krumholz HM. Outcomes research: myths and realities. Circ Cardiovasc Qual Outcomes. 2009; 2(1):1–3. [PubMed: 20031804]
13. Krumholz HM. Grant applications with a result-based orientation. Circ Cardiovasc Qual Outcomes. 2013; 6(5):507–8. [PubMed: 24021690]
14. Krumholz HM. How do we know the value of our research? Circ Cardiovasc Qual Outcomes. 2013; 6(4):371–2. [PubMed: 23838103]
15. 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(3):404–8. [PubMed: 23467469]
16. Platt JR. Strong inference: certain systematic methods of scientific thinking may produce much more rapid progress than others. Science. 1964; 146(3642):347–53. [PubMed: 17739513]
17. Morgan TM, Krumholz HM, Lifton RP, Spertus JA. Nonvalidation of reported genetic risk factors for acute coronary syndrome in a large-scale replication study. JAMA. 2007; 297(14):1551–61. [PubMed: 17426274]
18. National Research Council. Frontiers in massive data analysis. Washington (DC): National Academies Press; 2013.
19. Hill AB. The environment and disease: association or causation? Proc R Soc Med. 1965; 58(5): 295–300. [PubMed: 14283879]
20. Hood L. Systems biology and p4 medicine: past, present, and future. Rambam Maimonides Med J. 2013; 4(2):e0012. [PubMed: 23908862]
21. Way, MJ.Scargle, JD.Ali, KM., Srivastava, AN., editors. Advances in machine learning and data mining for astronomy. Boca Raton (FL): Chapman and Hall/CRC; 2012.
22. ZENODO. About ZENODO [Internet]. Geneva (Switzerland): ZENODO; [cited 2014 May 28]. Available from: https://zenodo.org/about
23. Optum Labs. A groundbreaking approach [home page on the Internet]. Eden Prairie (MN): Optum; [cited 2014 May 28]. Available from: http://www.optum.com/optumlabs.html
24. Health Care Cost Institute [home page on the Internet]. Washington (DC): HCCI; [cited 2014 May 28]. Available from: http://www.healthcostinstitute.org/
25. Krumholz HM. Open science and data sharing in clinical research: basing informed decisions on the totality of the evidence. Circ Cardiovasc Qual Outcomes. 2012; 5(2):141–2. [PubMed: 22438459]
26. Krumholz HM, Ross JS, Gross CP, Emanuel EJ, Hodshon B, Ritchie JD, et al. A historic moment for open science: the Yale University Open Data Access project and Medtronic. Ann Intern Med. 2013; 158(12):910–1. [PubMed: 23778908]
27. Partovian C, Li SX, Xu X, Lin H, Strait KM, Hwa J, et al. Patterns of change in nesiritide use in patients with heart failure: how hospitals react to new information. JACC Heart Fail. 2013; 1(4): 318–24. [PubMed: 24621935]
28. Xu X, Li SX, Lin H, Normand S-LT, Kim N, Ott LS, et al. ‘Phenotyping’ hospital value of care for patients with heart failure. Health Serv Res. 2014 Forthcoming.
29. National Institutes of Health. Big data to knowledge (BD2K) [Internet]. Bethesda (MD): NIH; [cited 2014 May 28]. Available from: http://bd2k.nih.gov/#sthashXczGHAVfdpbs