JACC: CARDIOVASCULAR IMAGING
© 2017 BY THE AMERICAN COLLEGE OF CARDIOLOGY FOUNDATION
PUBLISHED BY ELSEVIER

VOL. 10, NO. 11, 2017
ISSN 1936-878X/$36.00
http://dx.doi.org/10.1016/j.jcmg.2016.10.013

EDITORIAL COMMENT

Paradigm Shift in Medical Data Management

Big Data and Small Data*

James B. Seward, MD

> *The four stages of acceptance: 1) This is worthless nonsense; 2) This is an interesting, but pervasive point of view; 3) This is true, but quite unimportant; 4) I always said so.*
>
> J.B.S. Haldane (1)

Medicine is at the beginning of a paradigm shift regarding how we store and process larger and larger amounts of data. The change in data efficiency will not be a factor of 50% to 75%, but rather millions of times (2). In this issue of *iJACC*, Omar et al. (3) and a team of expert researchers, statisticians, and practitioners explore the fast-changing landscape of big data. As a reviewer, I am impressed by the content and narrative; however, my first impression is that the era of data informatics will come with significant challenges, including understanding and implementing new strategic assets, finding multiple experts with unique talent, and becoming familiar the convergence of IT and data science. Before we begin the venture, we must understand the basic features of the impending paradigm shift.

SEE PAGE 1291

BACKGROUND

As we enter into the 21st century, we are recognizing that our current view of human disease based on simple relationships has very serious shortcomings (4,5). A disruptive change is absolutely necessary, because our current systems are failing in efficiencies, reproducibility, disease prevention, and affordability. We are informed that computer-assisted statistical learning (i.e., Big Data [6,7]) and quantitative systems biology (i.e., Small Data [8,9]) will become extremely useful for diagnosing and implementing personalized evidence-based diagnosis and management of both overt and emergent disease processes. The whole universe of medicine is about to change.

CHANGE. For more than 60 years, we have known that clinical predictions based on the subjective impressions of trained experts are inferior to computer prediction algorithms (2). There is no exception to this observation in more than 200 studies (9). A larger portion of the health care infrastructure will be able take advantage of computer-based expert knowledge and assisted decision-making. We are destined to become data-dependent and computer-assisted.

DATA. Large amounts of data are analogous to crude oil. More data (e.g., more oil) does not necessarily equate to more information; insight and value are not the same; ability to interpret and use data varies based on the data itself; and implementation of data itself can be challenging. Be aware that most data are noise, just as most of the universe is filled with empty space. *Data analysis* defines highly related data and converts it into usable correlations and associations. *Data analytics* uses refined data to impute insights into the data function and things you may not have otherwise known. Most important is that there is a fundamental difference between Big Data correlations and Small Data causality.

COMPLICATED VERSUS COMPLEX SYSTEMS. Machine learning (Big Data) is a *complicated* form of informatics (e.g., building an airplane) focused on very large volumes of data that may be analyzed to reveal patterns (e.g., diagnostic phenotypes), trends (risk associations), and correlations (outcome) that can relate to disease behavior and relationships (10). Small data is a *complex* form of informatics (e.g., what can the plane do?) using small data sets that can analyze data

*Editorials published in *JACC: Cardiovascular Imaging* reflect the views of the authors and do not necessarily represent the views of *JACC: Cardiovascular Imaging* or the American College of Cardiology.

From the Division of Cardiovascular Disease (Emeritus), Mayo Medical School and Clinic, Rochester, Minnesota. Dr. Seward is a cardiovascular consultant to the RENEW GROUP and has equity in the company from which he does not receive income; and is president of the nonfunctional holding entity EchoMetrics Cardiologist.

behavior and interactions, which are easily understood (transparent), knowledge-based (validated), actionable (cause-and-effect disease management), and informative (monitor cause-and-effect data change) (10,11).

ANALYTICS AND ALGORITHMS

Data analytics and algorithms enable better understanding of data-derived knowledge (12). Analytics is the general science of formulating data associations. Big Data analytics are most commonly used to affirm data-to-disease associations or correlations (e.g., diagnosing diastolic dysfunction) (11). Algorithms are the step-by-step instructions (e.g., recipe) used in computing for archiving desired results. Small Data algorithms are used to define causality and emergence of dynamic outcome relationships (e.g., cause and effect) (2,9,12).

Deductive and *inductive reasoning* are 2 methods of logic used to arrive at a conclusion (8,13,14). Big Data conclusions impute linear deductive associations and correlations (e.g., absolute answer; $2 + 2 = 4$) (6,7). Small Data conclusions impute nonlinear inductive conclusions (e.g., close enough answer; $2 + 2 \approx 3.9$) (11). Human intelligence is principally a series of inferred approximations that are rarely 100% conclusive (7,13). Both humans and computers learn more rapidly and decisively through inductive reasoning. Deductive reasoning can provide certainty, but its scope and utility is more limited (14,15).

CORRELATION VERSUS CAUSALITY. Big Data best computes linear correlations; Small Data can compute nonlinear causality (emergence) (11). Big Data correlations cannot adequately imply causation. It is causation that actually generates a correlation and joint probability distribution. The future of medicine will use Big Data to assist in making a diagnosis and Small Data will direct preventive actions and monitor and quantify management of emergent disease processes.

PRECISION AND ACCURACY. Precision is a measure of reliability and consistency. Both small and big data models have high degrees of precision (reproducibility). Accuracy is a measurement of how closely results agree with the true or accepted value. Adding data to a model increases precision, which can mistakenly masquerade as an accurate conclusion (4,7). Probabilistic small data models, composed of only 2 to 4 highly vetted data, has a high degree of evidence-based accuracy (8).

BIG DATA USE. Big Data is designed to establish syndromic patterns that streamline the number of phenotypes to consider, but lacks the sensitivity in monitoring cause and effect (11,14). Big Data in health care today is largely limited to research because it requires a highly specialized skill set and expense (complicated type system). Uniform comprehensive outcomes data are often missing, which can become an obstacle to widespread adoption (4).

SMALL DATA USE. Small Data systems are user-friendly, requiring only an expert user developer and little statistical expertise and low cost (9). A small algorithm (e.g., 2 to 4 data) can define cause-and-effect modifiers, explain different disease manifestations, quantify risk, and monitor cause-and-effect management (8). The greatest limitation is that clinical medicine has had almost no exposure to complex-type system informatics.

CONCLUSIONS

We have entered an Internet Renaissance that dwarfs anything medical practice has ever seen. The whole infrastructure of medicine is destined to change. Acceptance of this new technology will methodically evolve (1). Prediction should not be done solely for the sake of prediction. Without having the proper framework in place, with a willingness to intervene and the context for meaningful use, prediction is really not very useful.

Some experts in the community will tend to oppose any conceptual change. Change will mostly logically occur by generating additional supporting evidence and not trying to discover falsifying evidence. As the new paradigm solidifies and unifies, it will replace the old paradigm, and a paradigm shift will have occurred.

Are big and small data informatics ready for implementation into my practice? The technology is ready (YES) but (NO) because the means to deliver and interpret computer-assisted data management has yet to be adequately developed. We're at step 2 of the acceptance cascade (1). Teams of experts such as that of Omar et al. (3) will help us navigate the stepwise emergence of computer-assisted decision-making.

ADDRESS FOR CORRESPONDENCE: Dr. James B. Seward, Emeritus: Mayo Medical School and Clinic, Division of Cardiovascular Disease, 102 South Broadway #310, Rochester, Minnesota 55904. E-mail: jim@echo-metrics.com.

REFERENCES

1. Haldane JBS. The Four Stages of Acceptance. Available at: http://www.goodreads.com/quotes/78634-the-four-stages-of-acceptance-1-this-is-worthless-nonsense. Accessed January 4, 2017.

2. Bishop MA, Trout JD. 50 years of successful predictive modeling should be enough: lessons for philosophy of science. Philos Sci 2002;69:S197–208.

3. Omar AMS, Narula S, Abdel Rahman MA, et al. Precision phenotyping in heart failure and pattern clustering of ultrasound data for the assessment of diastolic dysfunction. J Am Coll Cardiol Img 2017;10:1291–303.

4. Ioannidis JPA. Why most published research findings are false. PLoS Med 2005;2:e124.

5. Topol E. The Creative Destruction of Medicine. How the Digital Revolution Will Create Better Health Care. New York, NY: Basic Books; 2013.

6. Deo RC. Machine learning in medicine. Circulation 2015;132:1920–30.

7. Nate Silver. The Signal and the Noise. Why So Many Predictions Fail – But Some Don't. New York, NY: Penguin Books; 2015.

8. Loscalzo J, Barabási A-L. Systems biology and the future of medicine. Wiley Interdiscip Rev Syst Biol Med 2011;3:619–27.

9. Kahneman D. Thinking, Fast and Slow. New York, NY: Farrar, Straus and Giroux; 2011.

10. Allen W. Complicated or complex—knowing the difference in important. Planning Monitoring, Evaluation and Learning - supporting sustainable development. Available at: http://learningforsustainability.net/post/complicated-complex/. Accessed December 23, 2016.

11. Lucas RM, McMichael AJ. Association or causation: evaluating links between "environment and disease." Bull World Health Org 2005;83:792–5.

12. SAS. From Data to Action: A Harvard Business Review Insight Center Report. 2014 http://www.sas.com/en_us/whitepapers/hbr-from-data-to-action-107218.html. Accessed August 10, 2016.

13. Cellier FE, Lopez J. Causal inductive reasoning: a new paradigm for data-driven qualitative simulation of continuous-time dynamical systems. Syst Anal Mode Simulation 1995;18:27–43.

14. Loscalzo J, Kohane I, Barabasi A-L. Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. Mol Syst Biol 2007;3:1–11.

15. Silver DL. Machine lifelong learning: challenges and benefits for artificial general intelligence. In: Schmidhuber J, Thorisson KR, and Looks M, editors. Artificial General Intelligence 2011, 4th International Conference, AGI 2011. Berlin, Germany: Springer-Verlag, 2011:370–5.

KEY WORDS big data analytics, diastolic dysfunction, left ventricular filling pressures, speckle-tracking echocardiography