Big Data and Small Data: Reflections on data science, statistical modeling, and financial and health care reforms

by Naihua Duan[*]

Lai (2013) is a landmark contribution to the literature on the important innovations resulting from the emergence of "big data". This commentary offers some reflections to elaborate on several important topics discussed in Lai (2013).

Patient-Centered Outcomes Research Institute (PCORI)

As Lai (2013) noted, one of the important provisions of the U.S. Patient Protection and Affordable Care Act (PPACA) was the establishment of the Patient-Centered Outcomes Research Institute (PCORI, http://www.pcori.org/). PCORI's mission and vision are as follows:

> MISSION: The Patient-Centered Outcomes Research Institute (PCORI) helps people make informed healthcare decisions, and improves healthcare delivery and outcomes, by producing and promoting high integrity, evidence-based information that comes from research guided by patients, caregivers and the broader healthcare community.
>
> VISION: Patients and the public have information they can use to make decisions that reflect their desired health outcomes.

It is important to note that PCORI is not just another research institute. It is an innovative institute with a unique mission and vision that are focused on serving consumers, namely, "people", "patients and the public". In contrast, traditional research in medicine (often motivated by FDA requirements) has been focused mainly on serving developers, such as pharmaceutical and medical device enterprises. The change in focus from the traditional supply side to the relatively new demand side has important implications for the nature and methodology for the research agenda.

Accordingly, one of the priorities for PCORI-sponsored research is on research methodologies needed for its mission and vision:

> Developing and improving the science and methods of patient-centered outcomes research (PCOR) is a central part of PCORI's work. Better methods will produce more valid, trustworthy, and useful information that will lead to better healthcare decisions, and ultimately to improved patient outcomes.

In November 2013, PCORI released its first Methodology Report (PCORI Methodology Committee 2013), to serve as guidelines for PCORI itself and the broad community of researchers who conduct Patient-Centered Outcome Research (PCOR). In particular, an important component of this report (Section III) is a set of standards that provide specific guidance for the design and conduct of individual PCOR projects.

Following the release of this report, PCORI (2014) issued a funding announcement, "Improving Methods for Conducting Patient-Centered Outcomes Research," to provide funding up to US$17Million to address gaps in methodological research relevant to conducting PCOR. The funding provided has the potential to advance methodologies for PCORI's new and challenging research agenda focused on serving end-users. Whether this potential indeed materializes

[*] Division of Biostatistics, Department of Psychiatry, Columbia University
E-mail: naihua.duan@columbia.edu

remains to be seen; but this is definitely a good first step.

With the emphasis on methodology, the establishment of PCORI through the PPACA legislation is an exciting new development, not only for clinical, health services, and public health research domains, but also for applied mathematicians, including statisticians, psychometricians, informaticians, and computer scientists, who are interested in advancing the methodologies needed in the "brave new world" of PCOR, and the applications of these methodologies.

"Big Data"

As Lai (2013) noted, the emergence of "big data", largely driven by the rapid revolution in information technology (e.g., large volumes of online shopping), made a tremendous impact on the methodologies and applications of statistics and related "data science" disciplines.

In my personal experience, every time I log onto my Amazon.com account, I am amazed with the products Amazon recommends for my consideration, ranging from "*Doing Bayesian Data Analysis*" to "*Complete Catalan with Two Audio CDs*;" both products are indeed tempting for my interests. Big Brother Amazon knows me well, to the credit of ingenuous statisticians. Indeed, Amazon (2014) describes its Machine Learning Science Career Category as follows:

> Working with Big Data
>
> Machine Learning Scientists work in the research and development of algorithms that are used in adaptive systems across Amazon. They build methods for predicting product suggestions (recommendations) and product demand (forecasting), and explore Big Data to automatically extract patterns (large-scale machine learning and pattern recognition). Some Machine Learning positions we hire for:
>
> Machine Learning Scientist
> Manager, Research Science
> Data Sciences Lead

One of Amazon's job openings in this category, for a Research Statistician, is described as follows:

> Job Description
>
> PhD in Statistics or related field, expertise in experimental design, sampling, survival and duration modeling, analyzing large experimental and observational data sets. The person filling the position will provide expertise in the design and interpretation of experiments using an existing experimental platform, will also devise new experiments and extensions of the existing platform. The ideal candidate will have deep understanding of computational methods and computer implementation of statistical methods. Expected expertise includes at least one of the major statistical programs (STATA, SAS, SPSS), one of the lower level languages having statistical packages (R, MatLab) as well as some familiarity with C++, or JAVA. Hadoop, PIG or Hive experience useful but not required. The individual chosen will also interact with Amazon's Machine Learning community.
>
> Basic Qualifications
>
> Research Statistician:
> PhD in Statistics or related field, expertise in experimental design, sampling, survival and duration modeling, analyzing large experimental and observational data sets.
> Deep understanding of computational methods and computer implementation of statistical methods.
>
> one of the major statistical programs (STATA, SAS, SPSS)
> one of the lower level languages having statistical packages (R, MatLab)

Kudos to the Machine Learning Scientists at Amazon, for your great work to find "*Doing Bayesian Data Analysis*" and "*Complete Catalan with Two Audio CDs*" for me. Your good work has helped make life more enjoyable for me, and many of my peers. Of course your good work has also brought good results to Amazon's bottom line. You are exemplary of "big data" statisticians working successfully in the trenches.

It is probably fair to characterize "big data" statistics as supply side statistics, serving primarily the needs for supplier such as Amazon, in a way analogous to the traditional research in medicine that has been focused mainly on serving developers, such as pharmaceutical and medical device enterprises. The emerging "small data" statistics, discussed in the next section, offers a different paradigm.

"Small Data" Is the Real Revolution?

While the "big data" enterprise is thriving, both intellectually and financially, the notion of "small data" has also emerged in recent years, focused on serving consumers in a direct way, similar to the focus on Patient Centered Outcome Research. Pollock (2013) argued that we should "Forget big data, small data is the real revolution", with the reasoning that the real revolution in information technology "is the mass democratisation of the means of access, storage and processing of data."

Remember the good old days when the main frame computer was the way to go? With the "hundred flowers" that bloomed all around the globe on personal computers, tablets, smartphones, etc., one has to wonder where have the main frame computers gone?

While the emergence of "big data" is driven by the revolution in information technology, the emergency of "small data" is also driven by the same revolution, especially the democratization of information technology that empowers just about everyone to participate in the data enterprise. The days are long gone when one has to rely on gatekeepers for essentially all data functions: to rely on the operator to make a phone call, to rely on a banker to handle a financial transaction, to rely on the travel agent to book hotels and flights, to rely on stock brokers to purchase/sell stocks and other investment instruments, etc. Instead of the gatekeepers monopolizing on those data functions, just about everyone nowadays can perform those data functions on his/her own.

Estrin (2013) observed: “We leave a trail of digital data crumbs as we go about our days. With access and good apps, we could make sense of this “small data” to help get a clearer picture of our personal health.” This desire among consumers to “get a clearer picture of our personal health” has fostered several “small data” movements, such as PatientsLikeMe.com and QuantifiedSelf.com, that facilitate patients to keep track and makes sense of their own “data crumbs”.

The difference between “big data” and “small data” does not necessarily lie in the size of the dataset. With more than 220,000 members as PatientsLikeMe.com claims, the combined database at PatientsLikeMe.com is by no means “small.” There are compelling advantages for patients participating in PatientsLikeMe.com to share their experience with other patients, Patients Like Me, to learn from each other. At the same time, the “big data” at Amazon.com is assembled from chunks and chunks of “small data” from me and my peers.

What really distinguishes “big data” and “small data” lies in the “mass democratization” referred to in Pollock (2013) as the “real revolution.” PatientsLikeMe.com describes its patient-centered enterprise as follows:

> We want to democratize the process of monitoring disease progression and return the results to patients as quickly as possible, so they are empowered to make the best decisions.

The driving force for the “big data” enterprise, such as Amazon.com, is to serve the supplier. The steering wheel is in the hands of the Big Brother – a benevolent Brother, but nevertheless a Big, Centralized Mainframe. The driving force for the “small data” enterprise, such as PatientsLikeMe.com and QuantifiedSelf.com, is to serve the consumer. While patients might choose to pool their data together for a common good, the combined data are, as Pollock (2013) describes, “small pieces loosely joined” instead of a centralized enterprise.

Single-Patient Trials (SPTs): A “Small Data” Methodology

For clinical applications, a good example of “small data loosely joined” is the methodology of single-patient trials (SPT, a.k.a. n-of-1 trials) (see, e.g., Duan et al. 2013; Kravitz, Duan, et al. 2014). The SPT is a multiple cross-over trial conducted within an individual patient to inform the current patient’s treatment decision.

As a hypothetical example, consider a patient with lower back pain who has been taking the over-the-counter medication ibuprofen (Advil) for some time; the results are not totally satisfactory. Therefore she and her clinician are considering switching to naproxen (Aleve) in the hope of getting better results. In order to make an informed decision, they choose to consider an SPT, switching back and forth between ibuprofen and naproxen from week to week to find out which agent provides better results, to determine a treatment plan for the future.

The experimental units for this SPT are time periods such as weeks. The assignment of treatment (ibuprofen or naproxen) can be randomized, or counter-balanced such as ABBABAAB…, to control for possible confounding with time trend. With either approach, what is important is to achieve good balance between the A weeks and B weeks, so that the results of the SPT are based on comparing apples to apples.

There are many parameters in such a patient-centered SPT that can be determined by the patient, to accommodate her preferences. It should be noted that the usual practice of standardization of the protocol in traditional clinical trials does not necessarily apply here – while traditional clinical trials need to be standardized in order to produce generalizable knowledge that can be applied to future patients, the primary objective for this patient-centered SPT is to serve the needs for the current patient, therefore this SPT should be customized to match the needs and preferences for the current patient. As an example, the outcome measures should be selected to reflect what matters to the current patient. Some patients might give the highest priority to symptom reduction; some might prioritize functioning; some might prioritize side effects. Similarly, the duration of the trial needs also be set according to patient preference. Some patients might like to finish the SPT within a few weeks and move on, accepting a higher level of uncertainty in the results of the trial; some patients might have more patience and prefer to have a longer trial to achieve a higher level of precision. (Some patient education material is needed here, to help the patient and her clinician decide how to balance the duration of the SPT and the level of uncertain to be tolerated.)

Such an SPT will produce a chunk of “small data” or what Estrin (2013) refers to as “a trail of digital data crumbs.” With suitable statistical utilities, those data crumbs can provide useful clinical knowledge specific to the current patient to inform her treatment decision.

It should be noted that the SPT is not a panacea for all clinical conditions. The conditions suitable for the SPT are discussed in Duan et al. (2013, Table 1): chronic conditions (such as lower back pain) that require long term treatment (not surgery!), so that the knowledge gained from the SPT can inform future long term treatment outcome; heterogeneity of treatment effects (one size does not fit all, some patients might do better with ibuprofen, some might do better with naproxen, some might be tied between the two agents), so that there is a need for personalized knowledge to inform each patient's treatment decision; rapid onset and washout, so that the outcome observed in an A (B) week indeed reflects the patient's response to treatment A (B); etc.

Gabler et al. (2011) reviewed the medical literature for SPTs published between 1985 and December 2010 and identified 108 studies reporting on 2,154 participants. So there are at least 2,154 patients who have fostered their "small data." There are of course possibly many other patients who utilized SPTs but were not reported in the medical literature.

Is there a potential "market" for the SPT as a clinical decision tool? Several recent pilot studies (Brookes et al. 2007, Kravitz et al. 2009, Nikles et al. 2010) examined the acceptability for the SPT among potential end users (patients and clinicians, and other stakeholders). The results are generally encouraging for this personalized approach for clinical decision-making. Further research in this area is needed to inform the design and implementation of future SPT programs.

Combining "Small Data" Chunks

An ensemble of "small data" chunks from similar SPTs can be "loosely joined," in the expression of Pollock (2013), to achieve a variety of benefits beyond what can be accomplished with an individual SPT alone.

First, the current patient may benefit from the joint consideration of the results from her own SPT and the results from other SPT from patients similar to her with similar clinical conditions comparing similar treatment options. If the results from her own SPT are consistent with the results from the other SPTs, she would have more confidence in the treatment decision. If the results from her own SPT are different from the results from the other SPTs, she and her clinician need to discuss how to interpret the two sets of results for her treatment decision. Statistical methods based on the Bayesian/empirical Bayesian approach (Nikles et al. 2011, Zucker et al. 1997, 2006, 2010), sometimes known as *borrow from strength* (Higgins and Whitehead 1996), can be use to combine the results from the current patient's own SPT and the aggregated results from the ensemble of all SPTs, weighting the two parcels of information according to the balance between the level of uncertainty within the current patient's own SPT and the variation across patients for the results of the SPTs. If the current patient's own SPT has a low level of uncertain, e.g., if this SPT was conducted over a long duration with a large number of crossovers, more weight is given to the current SPT. If there is little variation across patients, i.e., one size does fit all, more weight is given to the aggregate results.

The second reason for combining the "small data" chunks across patients is to use the experience from previous SPTs to inform the design of the current SPT. For example, with the Bayesian approach, the knowledge gained from previous SPTs can be used to construct the prior distribution for the current SPT, which can then inform the determination for the duration of the current SPT, taking into consideration the needs and preferences for the current patient.

Furthermore, the "small data" chunks can be combined across patients to produce aggregate estimates for the comparison between treatments A and B, to inform treatment decisions for future patients who choose not to participate in their own SPT, or do not have an opportunity to do so. In addition, the combined SPT data can also be used to assess the heterogeneity of treatment effect (HTE, Kravitz et al. 2004), namely, how does the difference between A and B vary across patients?

SPT for Personalized Medicine

The SPT can be used as a tool to personalize treatment, to improve the overall effectiveness of the treatment protocol, in the presence of HTE. Such applications of the SPT can be used in the absence of known biomarkers – each SPT can stand on its own and provide information on the treatment decision for the specific patient. With an ensemble of SPTs, it is also possible to examine the individual-specific treatment effects across patients to explore the presence of biomarkers or other factors that might indicate a plausible pattern of HTE that deserves further investigations. Alternatively, the SPT can be applied in the presence of known biomarkers hypothesized to indicate HTE, to validate or reject the hypotheses and inform future treatment decisions.

Methodological Developments for SPT

Duan et al. (2013) recommended methodological development to expand the design methodology for SPTs, such as the use of response-adaptive designs and sequential stopping rules, similar to those discussed in Lai (2013) for parallel group trials. Those designs need to be adapted to take into consideration the multi-level panel structure for the SPTs. There are indeed many challenging methodological research questions for application-minded statisticians and data scientists to resolve and contribute towards further developments in this promising patient-centered clinical domain (Duan et al. 2013, Section 5).

Discussions

There are many commonalities between "big data" and "small data". While "big data" is akin to supply side economics, while "small data" is more like demand side economics, the success of a marketplace requires the supply and demand to meet on a common ground. There is a promising potential for applied mathematicians (statisticians, psychometricians, informaticians, computer scientists/engineers, etc.) to work together and contribute towards this fast-paced revolution that is likely to shape our lives in the latter part of the twenty-first century.

References

Amazon.com. 2014. "Machine Learning Science." http://www.amazon.jobs/jobs-category/machine-learning-science.

Brookes S. T., Biddle L., Paterson C., Woolhead G. and Dieppe P., "Me's me and you's you": Exploring patients' perspectives of single patient (n-of-1) trials in the UK. Trials (2007), 8:10.

Duan N., Kravitz R. L. and Schmid C. H., Single-patient (n-of-1) trials: A pragmatic clinical decision methodology for patient-centered comparative effectiveness research. *J. Clin. Epidemiol.* **66**(8 Suppl) (2013), S21–8.

Estrin D., "Small data, N = me, Digital Traces." 2013. Presented at TEDMED 2013, Washington, DC. http://smalldata.tech.cornell.edu/narrative.php.

Gabler N. B., Duan N., Vohra S. and Kravitz R. L., N-of-1 trials in the medical literature: a systematic review. *Med. Care.* **49**(8) (2011 Aug), 761–8.

Higgins J. P. and Whitehead A., Borrowing strength from external trials in a meta-analysis. *Stat. Med.* **15** (1996), 2733–49.

Kravitz R. L. and Duan N., eds, and the DEcIDE Methods Center N-of-1 Guidance Panel (Duan N., Eslick I., Gabler N. B., Kaplan H. C., Kravitz R. L., Larson E. B., Pace W. D., Schmid C. H., Sim I., Vohra S.). *Design and Implementation of N-of-1 Trials: A User's Guide.* AHRQ Publication No. 13(14)-EHC122-EF. Rockville, MD: Agency for Healthcare Research and Quality; February 2014. www.effectivehealthcare.ahrq.gov/N-1-Trials.cfm.

Kravitz R., Duan N. and Braslow J. T., Evidence Based Medicine, Heterogeneity of Treatment Effects, and the Trouble with Averages. *Milbank Q.* **82**(4) (2004), 661–687. Erratum in: *Milbank Q.* **84**(4) (2006), 759–60.

Kravitz R. L., Paterniti D. A., Hay M. C., Subramanian S., Dean D. E., Weisner T., Vohra S. and Duan N., Marketing therapeutic precision: Potential facilitators and barriers to adoption of n-of-1 trials. *Contemp. Clin. Trials* **30**(5) (2009), 436–45. Epub 2009 Apr 16.

Lai Z. L., Data Science, Statistical Modeling, and Financial and Health Care Reforms. *Notices of the ICCM* **1**(2) (2013), 47–57.

Nikles J., Mitchell G. K., Clavarino A., Yelland M. J. and Del Mar C. B., Stakeholders' views on the routine use of n-of-1 trials to improve clinical care and to make resource allocation decisions for drug use. *Aust. Health Rev.* **34**(1) (2010), 131–6.

Nikles J., Mitchell G. K., Schluter P., Good P., Hardy J. and Rowett D., et al. Aggregating single patient (n-of-1) trials in populations where recruitment and retention was difficult: the case of palliative care. *J. Clin. Epidemiol.* **64** (2011), 471–80.

PCORI (Patient-Centered Outcomes Research Institute) Methodology Committee. 2013. "The PCORI Methodology Report." pcori.org/research-we-support/research-methodology-standards.

PCORI. "Improving Methods for Conducting Patient-Centered Outcomes Research." 2014; http://www.pcori.org/funding-opportunities/funding-announcements/improving-methods-for-conducting-patient-centered-outcomes-research-spring-2014-cycle/.

Pollock R., "Forget big data, small data is the real revolution." 2013. http://www.theguardian.com/news/datablog/2013/apr/25/forget-big-data-small-data-revolution.

Zucker D. R., Schmid C. H., McIntosh M. W., D'Agostino R. B., Selker H. P. and Lau J., Combining single patient (N-of-1) trials to estimate population treatment effects and to evaluate individual patient responses to treatment. *J. Clin. Epidemiol.* **50** (1997), 401–10.

Zucker D. R., Ruthazer R., Schmid C. H., Feuer J. M., Fischer P. A. and Kieval R. I., et al. Lessons learned combining N-of-1 trials to assess fibromyalgia therapies. *J. Rheumatol.* **33** (2006), 2069–77.

Zucker D. R., Ruthazer R. and Schmid C. H., Individual (N-of-1) trials can be combined to give population comparative treatment effect estimates: methodologic considerations. *J. Clin. Epidemiol.* **63** (2010), 1312–23.