Big Data in Nephrology: Friend or Foe?

Terry Ketchersid

Fresenius Medical Care North America, Danville, Va., USA

Key Words
Big data · Genomics · Predictive analytics

Abstract
The phrase ‘big data’ has arrived in today’s lexicon with great fanfare and some degree of hyperbole. Generally speaking, big data refer to data sets that are too complex to be successfully interrogated using standard statistical software. A wide variety of business sectors has utilized big data to garner competitive advantage within their respective markets. Medicine and nephrology, in particular, have been late to this table. This is beginning to change, however, as data scientists begin to work with these large data sets, developing predictive models that permit us to peer into the future. Coupled with an expanding understanding of genomics, predictive models constructed with the assistance of big data may soon provide us with a powerful tool to use as we provide care to patients with renal disease.

Introduction
What do a wine-loving Princeton economist [1], a baseball general manager [2] in Oakland, California, and a computer [3] that plays Jeopardy better than anyone on the planet have in common? All three examine very large data sets and make predictions or forecasts based on patterns in the data. The phrase ‘big data’ has become increasingly common over the past 5 years. It can loosely be defined as a data set that is so large and complex that it becomes difficult to work with using standard statistical software [4]. Business sectors as diverse as financial services, retail and the airline industry utilize big data to garner competitive advantage in their respective markets. They do so by interrogating these large data sets utilizing increasingly sophisticated analytic tools, generating insights that lead to better decisions. A common misconception surrounding big data is that the data themselves bring value to the firm, when in fact in the absence of analytics, the data are of little use.

The Current Evolution of Data

There are a couple of important aspects of data we need to put into perspective in order to understand the full impact of big data, those are volume and velocity. The widespread adoption of the internet has created a remarkably efficient vector for the creation of data. In Big Data: A Revolution That Will Transform How We Live, Work, and Think, the authors write ‘in 2013 the amount of stored information in the world is estimated to be...
around 1,200 exabytes, of which less than 2 percent is non-digital’ [5]. An exabyte is a unit of information equivalent to one quintillion (10^18) bytes or one billion gigabytes of data. The volume of data generated to date is astronomical.

Velocity creates another challenge regarding big data. One source suggests the amount of data is doubling every 2 years [6]. Clearly, the volume of the world’s data is expanding at an exponential pace. The velocity of data accumulation within a specific market sector or even within a single enterprise in a sector can begin to overwhelm existing infrastructures. Concurrently, the cost of data storage is plummeting. Terabyte storage devices are commercially available today for less than USD 100. As it becomes more economical to store massive amounts of data, the intersection of volume and velocity creates challenges for conventional analytics and gives rise to the opportunities championed by those behind the big data craze.

### Big Data and Nephrology

Health care is one of the last sectors to reap the benefits of big data [7]. Nephrology as a discipline is not yet well represented in this race. A recent Google search of ‘big data and nephrology’, accessed on October 7, 2013, yielded 2.67 million results, but the results were a bit disappointing. The page one results included a couple of blog posts, an upcoming presentation at the 2013 meeting of the American Society of Nephrology, and an editorial in a peer-reviewed journal. The editorial entitled ‘Better understanding live donor risk through big data’ was published in the Clinical Journal of the American Society of Nephrology [8]. This editorial examined a study which investigated the perioperative risk for live renal transplant donors using the US Department of Health and Human Services Agency for Healthcare Research and Quality's Healthcare Cost and Utilization Project Nationwide Inpatient Sample (NIS). The NIS is a large national de-identified database of all payer inpatient care that consists of a stratified sample of 20% of the nonfederal hospitals in the USA. According to the editorial, the authors of this study discovered several inconsistencies between the self-reported complication rates (reported by transplant centers to the Organ Procurement and Transplantation Network) and those harvested from the NIS data set. The editorial notes that current registry limitations are likely to persist and that within this context the utilization of large de-identified data sets will be necessary to supplement our understanding of postdonation health.

Not represented within a simple Google search is the vast data set that supports the work done by the United States Renal Data System (USRDS). The USRDS utilizes data collected by the Centers for Medicare and Medicaid Services for beneficiaries with renal disease. Among other things, USRDS publishes an annual data report based on this very large data set [9]. The underlying data present its share of challenges [10]; nevertheless, this type of analysis represents another example of using big data to examine past performance.

The USRDS experience and the live renal donor paper cited above are examples of using big data to ‘look back’ and examine what has happened in the past. From an analytic perspective, it is important to recognize two distinct features. Davenport et al. [11] published the analytic framework depicted in Table 1, which brings these features into focus. The first feature is the time frame. When analyzed, big data can help us understand what happened in the past, what is happening right now, or what is likely to happen in the future. The second important feature is to make the following distinction: are the results of the analysis providing information or insight?

<table>
<thead>
<tr>
<th>Past</th>
<th>Present</th>
<th>Future</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information</td>
<td>Insight</td>
<td></td>
</tr>
<tr>
<td>What happened?</td>
<td>How and why did it happen?</td>
<td>What's the next best action?</td>
</tr>
<tr>
<td>(reporting)</td>
<td>(modeling, experimental design)</td>
<td>(recommendation)</td>
</tr>
</tbody>
</table>

Table 1. Analytic framework

This last distinction is critical. The majority of the data representations we see in nephrology today are simply visual representations of historical data. We are inundated with historic run charts depicting hemoglobin trends or catheter rates, frequently looking at averages as op-
posed to distributions and outliers. These views of history are important, but one might argue the majority of the ‘analytics’ available to nephrologists today fall into the ‘past information’ cell in Davenport’s framework. As we will see later, it is the ‘future insight’ cell, the home of predictive analytics, which holds the most exciting promise for nephrology.

The Dilemma of Randomized Controlled Trials

Randomized controlled trials (RCTs) have long been recognized as the gold standard for medical research. In the search for truth regarding whether exposing patients to a particular therapy or treatment paradigm results in better primary or surrogate outcomes, simply assemble a large enough pool of subjects and randomly divide them into two groups, one exposed to the therapy under study and the other group, the controls, receiving a placebo. If both the subjects and those conducting the study have no clue what group the subjects are in (the crème de la crème known as a randomized double-blind controlled trial) even better. RCTs facilitate moving beyond identifying correlation and occasionally this framework leads to the derivation of causality.

The RCT has been around for many years [12] with the first published study related to the treatment of tuberculosis appearing in print in 1948 [13]. Forty years later, as a young and impressionable medical trainee, the author and his classmates eagerly consumed the results of as many RCTs as we could. In 1993, a group of like-minded individuals gathered in the UK and birthed what would become the Cochrane Collaboration [14]. Archibald Cochrane, the namesake of the organization, was a Scottish physician who became a champion for the RCT very early in his career [15]. Today, the Cochrane Central Register of Controlled Trials contains approximately 500,000 records and the collaborative has published over 5,000 systematic reviews. Not surprisingly, the concepts of evidence-based medicine and clinical practice guidelines have matured in conjunction with the growth of RCTs.

Population Averages versus Individual Patients

Of course, RCTs are expensive, they take a long time to conduct, and they frequently exclude patients with attributes that one frequently sees in the practice of medicine. Strict inclusion and exclusion criteria within the RCT can make it difficult to generalize the results of the study. Clinical guidelines are established to manage patient populations that consist of the ‘average’ patient. But increasingly, there are calls for personalized health care or care customization as opposed to broadly applying clinical guidelines to every unique patient [16]. The rising interest in pragmatic trial designs [17] and the utilization of observational data are occurring in concert with the establishment of very large data sets. This is where the possibilities of ‘big data’ become very interesting.

A related piece of this big data puzzle involves the human genome. The human genome, which consists of roughly 3 billion base pairs, was successfully sequenced by an international collaborative in 2000 [18]. Since that time, the cost of sequencing a person’s genome has plummeted. Commercially available products can now perform this remarkably complex process for about USD 5,000 [19]. Couple that with the fact that you could store the sequence of your genome on that inexpensive terabyte drive and it becomes clear the economics of this process are rapidly becoming quite favorable. As sequencing costs plummet, the bottleneck becomes the actual analysis of the contents of an individual’s genome. Some would argue that although the costs of sequencing have substantially declined, the promise of medical cures has not arrived [20]. In spite of this pessimism, there have been some remarkable success stories directly related to genome sequencing [21].

Data → Insight → Action

While much remains to be learned within the realm of genomics, there is another parallel path we should consider in the debate surrounding big data. That subject revolves around the intriguing discipline of predictive analytics. In his 2013 book entitled Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie or Die, Eric Siegel [22] defines predictive analytics as ‘Technology that learns from experience (data) to predict the future behavior of individuals in order to drive better decisions’. The sequence of events is as follows: data lead to insight and insight leads to benefit or action. Practically speaking, predictive models create the insight. What happens under the covers in the insight bucket is a series of remarkably complex calculations and machine learning techniques that are under the purview of a new breed of professionals. Data scientists are mathematically grounded statistics quants who build predictive models using large data sets [23]. The attributes that form the basis of these models include the usual suspects.
like tobacco use to predict the risk for coronary artery disease. But they may also include attributes previously hidden from view, like the impact of seasonal variation on important facets of care in the end-stage renal disease population recently discovered by the MONDO (Monitoring Dialysis Outcomes) initiative [24]. These previously unknown attributes are the bounty available for harvest in this process, a harvest that is not possible in the absence of big data.

Suppose you had access to several continuous years of medical claim data, laboratory data, medication prescription data and the entire electronic health record for a large multispecialty practice. Could you build a model that would predict a primary care patient’s risk of developing congestive heart failure over the next 12 months? That is exactly what researchers at the T.J. Watson Research Center of IBM did when they teamed up with epidemiologists at Geisinger Medical Center [25]. From the 30,000 foot view, this process involves examining very large historical data sets, identifying the appropriate mix of attributes, building the model using various combinations of these attributes and then validating the model against a subset of the original data set that was not part of the model construction process.

Note these models are not intended to be clairvoyant. They create value if the insight they provide is better than what we can achieve based upon our intuition. In other words, predictive models can help us make better decisions by taking some of the guesswork out of the decision-making process. Again notice ‘big data’ are necessary, but not sufficient, to complete this process. Simply accumulating large data sets is of no value if we are not in a position to analyze the data in a way that generates insights we can act upon.

Friend or Foe?

For the past 65 years, RCTs have been important to the advancement of our understanding of many areas within medicine. In conjunction with meta-analyses and other studies, these scientific experiments form the basis of evidence-based medicine and are frequently cited in the development of clinical practice guidelines. Not uncommonly, however, the patient in front of us is not the ‘average patient’ the guidelines address. There are many examples of this today, including our efforts to measure the delivery of dialysis. Kt/V targets are not only part of established practice guidelines, but the measure now forms the basis of one of the metrics used by Centers for Medicare and Medicaid Services to measure quality [26]. And yet, does it make sense to apply the same target to a 120-kg, 30-year-old, African-American male and a 50-kg, 80-year-old, Caucasian female? Some would argue treatment time and ultrafiltration rate might be a better approach, but in the context of the big data discussion, perhaps individualized treatment targets would be more appropriate.

Are big data the friend or foe of nephrology? I tend to believe the former. We face a tremendous number of unanswered questions as we deliver care to patients with renal disease. Our practice guidelines, while beneficial at the population level, do not always apply to the patient seated in the chair across from us. Granted, the insights garnered by applying analytics to big data typically identify correlation. But as we expand our understanding of the importance of genomics and we begin to develop and deploy nephrology-related predictive models, we may find ourselves in a position to make smarter clinical decisions for each unique patient we see.

References


Department of Health and Human Services: Centers for Medicare & Medicaid Services: 42 CFR Parts 413 and 414. Medicare Program; End-Stage Renal Disease Prospective Payment System and Quality Incentive Program; Ambulance Fee Schedule; Durable Medical Equipment; and Competitive Acquisition of Certain Durable Medical Equipment, Prosthetics, Orthotics and Supplies; Final Rule. Federal Register 70228/vol 76, No 218/November 10, 2011/Rules and Regulations.