The problem with Big Data in Translational Medicine. A review of where we've been and the possibilities ahead

Les Jordan

Bad Rabbit Consulting, 100 Spring St., Foxboro, MA 02035, United States

Abstract

Understanding the parallel evolutions of Big Data and Translational Medicine, and the types of disruptive technology that bring them together, requires a look back at their evolution and a discussion of the hindrances in applying big data techniques to translational medicine. We will then take a look into the future, at the concept of the “Complete Health Record” and how that may change the very nature of translational medicine.

© 2015 Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Contents

1. Introduction .............................................................. 0
2. The evolution of Big Data ........................................................ 0
3. Big Data and Translational Medicine today ................................................. 0
   3.1. Right data ............................................................ 0
   3.2. Right person ........................................................... 0
   3.3. Right time ............................................................ 0
4. Art of the possible ............................................................ 0
5. Emerging technologies .......................................................... 0
6. The drive to the Complete Health Record ......................................... 0
7. Conclusion ............................................................... 0
References .................................................................. 0

1. Introduction

It has been 15 years since the Human Genome project announced that it had sequenced 90% of the human genome. It took another three years until they announced the successful completion of the project, having sequenced 99% of the human genome (The Human Genome Project, 2003). Compare that to today, where an individual's genome can be sequenced in a matter of three days (Illumina Corp., 2015).

With those advances, it is now technically possible to create a patient registry enabled by a patient's Complete Health Record, combining genomic data, the vast amount of data available through Connected Medical Devices, the ever increasing number of personal fitness devices (through the Internet of Things), along with data from the electronic medical record (EMR) and other types of available patient data. One would think that the insights gained from the Complete Health Record would solve a wide range of health problems (see Fig. 1). But here is where the problem of improving patient care through translational medicine and the problems of big data intersect: the problem is not (necessarily) in the integration of the data: as with all big data problems, the problem is one of availability and analysis, in the finding and understanding of signals, trends, causal (and not corollary) relationships, and then turning those insights into actionable information.

To understand how to overcome these issues, and the truly disruptive technology that I believe lies on the horizon, we’ll approach this review in three parts: a look at the development of the concept of “big data”, a discussion of the hindrances in applying big data techniques in our environment today and then a look into the future, the concept of the “Complete Health Record” and how that may change the very nature of translational medicine.

2. The evolution of Big Data

When I started working in this industry in the late 1980s, running statistical analyses against large data sets on massively parallel...
computers, the concern was as much about processor speed (and number of available processors) as it was about the sheer volume of data. How long would the computing job run? When would we need to switch tapes or disks to load in more data? Consider the CrayY-MP C90 supercomputer, state of the art for crunching data in 1991, had 16 CPUs, with 2 Gb of central memory, ran at 16 Gigaflops and had 16 Gb of SSD storage (Cray Research, Inc. 1991). My current laptop has more raw processing power, central memory and storage than that Cray!

In the late 1990s came the availability of multiple networked PC processors, referred to as “High Performance Computing (HPC)” with around the same amount of processing power of a Cray, but with hundreds of terabytes of storage and a considerably lower cost. And while organizations were busy implementing HPC clusters and networking vast amounts of data, vendors started offering Software as a Service (SaaS), which itself evolved into storing vast amounts of data, vendors started offering Software as a Service (SaaS), which itself evolved into storing vast amounts of data in the Cloud. The issue evolved from technical limitations of physical storage or processing power to such questions as: “How much compute power and storage space can be purchased?” and “How quickly can the system be spun up?”

And the big question: how to integrate the data.

In the early 2000s, while SaaS and the Cloud were evolving, came a focus on data interoperability and integration standards. HL7, CDISC, Continua Alliance. SAFE BioPharma. IHE. All focused on enabling interoperability and integration. In 2006, I was evangelizing the concept of interoperability and integration standards and one of the major goals of this was to encourage organizations to embrace standards, citing the benefit “Do you have your next blockbuster drug in your discarded portfolio and don’t know it?”

The implication of that statement is by using standards you can have access to data, and views across data, which were previously hidden, stored in individual silos. And that once this data was made available to a wider range of researchers, additional insights could be gained, simply through the mindshare of additional researchers and the possibility of new ways of looking at data.

When it comes to these standards, we’ve largely arrived at a place where it is possible to realize that vision. By the time of this article in 2015, the industry has largely moved from evangelizing the need for developing standards to implementing mature standards that have been around for years, easing integration to a degree not possible just a decade ago.

3. Big Data and Translational Medicine today

And while we now have the ability to process and store massive amounts of data, and the standards and technologies that enable the integration of that data, there are still problems in Big Data in general and in applying Big Data techniques to Translational Medicine more specifically that hinder the vision of finding those critical insights.

Critics of Big Data use the oft cited phrase common to many information technology initiatives: the problem with Big Data is about the right data, delivered to the right people at the right time (Forte Wares, 2014). I’ve personally used this phrase for years, mostly when speaking about user experience and user interfaces, but it is equally applicable to Big Data, and especially when applied to Translational Medicine.

Let’s use that paradigm to test a common application of translational medicine: the patient registry.

3.1. Right data

Not only questions the veracity of the data, but the applicability of the data and (perhaps most importantly) the accessibility of the data.

- If the registry includes data from fitness trackers, was the data really generated by that patient?
- For registries driven by electronic medical records, is the information up to date?
- Is the right information captured for the type of study you are trying to recruit?
- Do you have access? While you may have good EMR data sets, have you formed agreements with fitness data providers like FitBit or patient focused data aggregators such as 23AndMe or PatientsLikeMe? Have you solved the problem of correlating patients from those services with the EMR data you already have?

3.2. Right person

When thinking about the right person, we need to think not only about the analysts themselves, but also consider the analysis technology that the analyst can use to provide insights.

- Do you have medical specialists on hand who understand all the sources of data in the registry?
- Do you have technology that can pose questions in the right way, interrogating data across domains, utilizing the Complete Medical Record, and can look at relationships between genomic, fitness, medical records, and patient adherence? Across patients? Across populations?

3.3. Right time

One of the tenets of Big Data is an acknowledgement that the underlying data is rapidly changing.

- Can you find patients to recruit for your trial, at precisely the right point in that patient’s health timeline, defined by the interaction between their recent fitness, current adherence pattern, and current medical record?
  - Was there a life-event that negatively impacted a patient’s adherence pattern years ago, but not currently?
  - Have you controlled for the patient’s current adherence history in the inclusion/exclusion criteria of a clinical trial?

Those are tough questions, not only showing how necessary it is to use the Complete Health Record and Big Data analysis techniques to drive insights, but also showing how difficult it is to use patient registries if the concept of the Complete Health Record is not taken into account.
4. Art of the possible

Yet there are emerging capabilities and trends which show using Big Data analysis techniques against the Complete Health Record, or at least real-world data, can drive insights not possible a few years ago.

Let’s consider the art of the possible, from some real world use cases that I’ve personally witnessed:

- A biopharmaceutical company has a drug going off patent in five years. By interrogating large EMR data sets, across geographies, with tens of millions of patient lives, other co-indications for that drug are detected through improvement in symptomology. These co-indications are in heretofore unstudied therapeutic areas, completely unrelated to the original FDA approval. The improvements in symptomology are detected by the analysis, not necessarily looking for a specific symptom, but using signal detection to look across indications, across symptoms.
- When planning large Phase III studies, or groups of studies, it is possible to take patient registry data, EMR data sets, and operational data from previous studies to better plan the protocol, understanding the profile of the patient AND the profile of the investigational site. In the end, this develops a better protocol, greatly enhances patient recruitment and leads to fewer protocol amendments. The end effect is saving time and money in the execution of the trial while ending up with better data for your submission. The best part: it is all done graphically, enabling the user to focus on the questions to ask the data, and not on how to format the questions or the underlying data set.

5. Emerging technologies

There are examples of emerging technologies and services that indicate there may well be solutions to some of the analysis issues. Companies such as Tamr and Mark Logic are taking novel approaches to solving the right data, right person, right time problems.

Tamr (http://www.tamr.com) utilizes a curated workflow along with machine learning. A curation workflow identifies and maps data the system doesn’t understand. The system then learns from that curation, applying machine learning algorithms, reducing curation intervention in subsequent data sets. Tamr offers an adapter for CDISC, enabling the conversion, validation and packaging of clinical study data into CDISC format (Tamr Inc., 2015).

Mark Logic (http://www.marklogic.com) is paving the way in Enterprise NoSQL solutions, which enable integration, storage, analysis and search across multiple, complex data types, including both structured and unstructured data (Mark Logic, 2015). This enables researchers to ask new questions as opposed to simply testing hypothesis.

6. The drive to the Complete Health Record

While the technology to drive insights from disparate sources and types of information, to drive the right data to (and for) the right person at the right time is certainly evolving in a positive direction, perhaps the greatest problem faced in driving insights using Big Data methodologies is the problem of availability, specifically the availability of the Complete Health Record.

Consider the different types of data enumerated in Fig. 1. Many of those data types are available for the purposes of building patient registries: Provider Focused Electronic Medical Records (EMR), Connected Medical Devices, Genomic Information. But a vast amount of data is not available in a single place for the patient, let alone for research purposes. Nor is it certain that the patient has access to ‘all’ their information, even though it is “theirs”. Indeed, this raises the question of ownership (which we won’t deal with in this article): is the patient’s data theirs or is it the physician/providers data?

From the researcher standpoint, where can they get data sets that include not just the EMR or Genomic information commonly in a patient registry, but also gain insights from patient created data, or data created from the increasing number of Connected Medical Devices, not just the devices provided by the provider or physician?

The possibility of using the Complete Health Record in patient registries, to drive a complete view of the patient, even possible?

The possibility does exist to build patient registries and other research data sets from the Complete Health Record of a patient, but as an industry we have work to do to get us there. Many countries have created centralized medical records. In the US, we have been moving towards connecting health records, even if only regionally. But even those efforts only connect Provider Focused data and completely ignore Connected Medical Devices and patient generated data.

There are existing technologies which allow the collection of patient focused as well as provider focused data. Microsoft’s HealthVault (Microsoft, 2015) enables patients to pull together their Complete Health Record, including data not just from provider-focused EMR and medical devices, but also pharmacy information, patient focused Connected Medical Devices, fitness information and a host of other patient focused health data (Full Disclosure: the author served as CTO for Microsoft’s Life Sciences Industry Unit from 2007 to 2013).

When provider organizations adopt this technology, integrating it with their EMR, it opens a door for greater availability of data, with the possibility of access to patient focused health data and not just provider focused health data. But that is also the downside of this approach: the requirement for the provider to enable access to their EMR, even if only through the download of an HL7 CCD (Continuity of Care Document) and the requirement for the patient to grant access to that data.

Microsoft places a heavy (but necessary) emphasis on privacy, requiring the patient to opt-in to the use of their data. If the patient is unaware of the ability to use their data for research purposes, if the apps don’t exist to take advantage of that patient data, then that Complete Health Record becomes useful only for the patient.

A different approach has been taken by Apple with their ResearchKit (Apple Inc., 2015). While the details and uses of the ResearchKit are emerging at the time of this article, early uses show promise. The approach taken by Apple with their ResearchKit is to enable the creation of apps specifically for collecting research data. Participating research organizations can contribute to the open source framework, create apps to gather data from patients and then publicize those apps within their target patient communities. The more patients that contribute, the richer the data. This approach has the downside of not capturing provider-focused data, nor the inability to integrate EMR data (at least as of this writing), but has the upside of gathering data from increasing numbers of patients.

What’s needed is a combination of these approaches. An approach where research organizations can create apps, focused on collecting not just the individual data types they create or utilize, but apps that can access all the different types of data available to the patient, including provider-focused health data. Apps that take into account not just the immediate data-points researchers THINK are of interest, but apps that have access to the vast array of patient data, have access to the patient’s Complete Health Record, the intersection of EMR, PHR, Connected Medical Devices, pharmacy information, fitness information and patient behavioral data.

These combined approaches enable the “holy grail” of Big Data in Translational Medicine: the ability to comb through data, to identify insights, enable signal detection, and find patterns that can create new and novel research questions rather than the traditional method of creating hypothesis and then generating the questions to test those hypotheses.

7. Conclusion

The standards exist for integrating both provider focused and patient focused health data, for the most part. The technology exists, and more
advanced technology is emerging that enable researchers to ask out-of-
the-box questions of the data, questions that weren’t envisioned when
they created their initial data-sets. Platforms exist for the capturing of
patient focused data and storing that data alongside provider focused
EMR data.

What’s needed to solve the problems with Big Data in Translational
Medicine is a combination of approaches: research focused apps com-
bined with access to a population of patients Complete Health Records,

enabling greater numbers of patients to contribute and allowing
researchers to have access to complete, all-encompassing health data,
which will drive insights and the ability to detect patterns, rather than
simply testing hypothesis, ultimately driving more timely and targeted
therapies to market.

References

Apple Inc., 2015. Apple Research Kit Apple.com. [Online] https://www.apple.com/
researchkit/.

Cray Research, Inc., 1991. The Cray Y-MP C90 Supercomputer Craysupercomputers.com.
[Online] http://www.craysupercomputers.com/downloads/CrayC916/CrayC916_Brochure001.pdf.

Forte Wares, 2014. Failure to Launch: From Big Data to Big Decisions. Forte Wares
[Online] June 19, http://www.fortewares.com/Administrator/userfiles/Banner/
forte-wares-pro-active-reporting_EN.pdf).

Illumina Corp., 2015. Illumina Expands World’s Most Comprehensive Next-Generation
Sequencing Portfolio. Illumina News Center ([Online] January 12, http://www.
illumina.com/company/news-center/press-releases/press-release-details.html?

newsid=2000979).

Mark Logic, 2015. Solutions for Life Sciences. Mark Logic ([Online] http://www.marklogic.
com/solutions/life-sciences/).

Microsoft, 2015. What Can You Do With HealthVault. Healthvault ([Online] https://www.
healthvault.com/us/en/overview).

Tamr Inc., 2015. Clinical Data Conversion (CDISC): Simpler, Scalable Conversion. Tamr
([Online] http://www.tamr.com).

The Human Genome Project, 2003. The Human Genome Project Completion: Frequently
Asked Questions. National Human Genome Research Institute ([Online] April 14,
http://www.genome.gov/1100043).