Pharm Med (2015) 29:87–92
DOI 10.1007/s40290-015-0090-x

LEADING ARTICLE

Big Data in Pharmaceutical R&D: Creating a Sustainable R&D Engine

Peter Tormay[1]

Published online: 21 March 2015

Abstract Over the last 20 years, productivity in the pharmaceutical industry has been diminishing because of constantly increasing costs while output has overall been stagnant. Despite many efforts, productivity remains a challenge within the industry. At the same time, healthcare providers quite rightly require better value for money and clear evidence that new drugs are better than the current standard of care, making a complex situation even more complex. With the implementation of 'Big Data' initiatives trying to integrate data from disparate data sources and disciplines that are available in life science, the industry has identified a new frontier that might provide the insights needed to turn the ship around and allow the industry to return to sustainable growth.

Key Points

In order to reinvigorate the pharmaceutical drug pipeline, companies need to take better advantage of the available data.

'Big Data' relates to large data sets that are highly complex. Data complexity is the key challenge in implementing Big Data approaches.

Integration of disparate data in the pharmaceutical industry will help to identify and validate new drug targets, support early identification of safety and efficacy issues, and improve patient stratification.

✉ Peter Tormay
peter.tormay@capish.com

[1] Capish Nordic AB, Stortorget 9, 211 22 Malmö, Sweden

1 Introduction

Do we need 'Big Data' in R&D and, if so, how can it help to overcome the challenges currently facing R&D productivity? It is undeniable that pharmaceutical R&D, as the engine of the pharmaceutical industry, has not been running smoothly over the last two decades. The approval of new molecular entities (NMEs)—products that are based on small chemical molecules or biologics, without a previous marketing authorization for a particular indication—has been more or less flat over the last two decades. The cost of bringing these medicines to market has been constantly rising over the same time period. More worrying, though, is the fact that the revenue anticipated from these new medicines is not going to make up for the shortfall created by recent patent expirations. This is putting the profitability of many companies at risk, making the current situation not sustainable [1].

This so-called innovation gap can be attributed to several internal challenges. Many promising drug candidates fail in phase II and phase III—later stages of the clinical development process [2]. These high attrition rates at a time when projects have already incurred high costs make for very expensive failures. Identification of new safety concerns or issues with the efficacy of the drug at this late stage results in an unfavourable risk/benefit relationship, thus rendering these projects commercially not viable. Moreover, the complexity of the clinical development process is constantly increasing with the implementation of new procedures. The Tufts Center for the Study of Drug Development showed that the overall execution burden grew by 54 % in the period 2004–2007 compared to the period 2000–2003 [3].

At the same time, there is also increasing external pressure on pharmaceutical companies. To start with,

patents for some of the best-selling drugs have recently expired, thus threatening the ability for sustained growth [4]. This is coupled with a changing therapeutic landscape to address clear unmet medical needs, resulting in projects with a lower probability of success [5]. This also means that most low-hanging fruits have been picked, particularly in those therapeutic areas that the industry has focused on in the last decade [6]. Increasing regulatory hurdles are also not helping the problem, although the impact on drug development is not entirely clear [7]. Moreover, regulatory approval is nowadays not enough, as the healthcare sector is moving away from a fee-for-service model to a value-based model through health technology assessments—for instance, by the National Institute for Health and Care Excellence in the UK or the Institute for Quality and Efficiency in Health Care in Germany. Pharmaceutical companies have to provide real-world evidence that new drugs that come on the market are better than existing therapies or the competition in order to get reimbursed. Productivity is therefore no longer just a function of R&D efficiency; it is also a function of R&D effectiveness [1].

The industry has looked at many ways to stem the decline in productivity, starting with increased R&D spending, followed by major consolidations, in-licensing, acquisitions and R&D reorganization—but to no avail [6].

Looking at all of these factors, it becomes evident that the root cause actually lies somewhere else: lack of data or lack of appropriate analysis of the available data. High attrition rates in late-stage clinical trials could, for instance, be avoided if the relevant information was available earlier or if the available information could provide clues as to whether a drug will actually perform as expected in clinical practice. The probability of success of current projects within complex therapeutic areas could be increased through better understanding of the underlying disease mechanism. In particular, the understanding of real-world effectiveness is tied to better insights into market requirements and real-world performance.

This review provides an overview of how Big Data and Big Data initiatives can advance the clinical development process to improve productivity in the pharmaceutical industry.

2 Big Data

The definition of Big Data is most often associated with the '3 V's' provided by Gartner [8]. Big Data involves high-volume, high-velocity and high-variety information assets, which require new forms of processing to enable enhanced decision-making, insight discovery and process optimization. In particular, in the context of pharmaceutical R&D, two other dimensions are highly relevant—namely, veracity and variability. Obviously, the Big Data movement is possible only because of the incredible advances in information technology (IT) and the different ways in which information and data can be captured.

The most interesting dimension, but also the most challenging, is variety. There are many different types of data that are highly relevant. When it comes to understanding disease mechanisms and drug discovery, the main focus has been on genomic data. Since the publication of the first human genome in 2004, the cost of sequencing has greatly gone down because of the establishment of new techniques. Several human genome reference projects have been launched, such as the 1000 Genomes Project [9] or the 100,000 Genomes Project [10]. These projects will make genetic information—together with other phenotypic as well as medical information—available to help and identify new drug targets by linking particular genes and their products to individual diseases. This is greatly aided by the availability of existing genome-wide association studies looking at single nucleotide polymorphisms (SNPs), insertions and deletions, as well as more pronounced rearrangements and their association with different diseases [10–13].

In recent years, data from other sources have been receiving more and more attention. In addition to genomic data, other -omics data have moved into the spotlight. Proteomics and metabolomics, as well as epigenetics and an integrated view of all of these disciplines, are gaining more and more traction. Also, the impact of lifestyle choices is now starting to be factored in.

On the other end of the value chain, electronic health records and other patient-related information in registries, hospital administration databases and payer databases are the focus of interest to establish real-world evidence for the effectiveness and the value of a particular medicine. For instance, Pfizer conducted a cohort study using the Health Improvement Network database in the UK to establish whether switching patients from atorvastatin (Lipitor) to simvastatin has a negative effect [14]. Sanofi undertook a similar approach with its diabetes drug Lantus to establish that Lantus was not associated with an increased risk of cancer [15] after it was rejected by the German health authority [16]. In 2011, AstraZeneca partnered with Healthcore, the analytics arm of WellPoint, to establish a partnership to conduct research, which will include prospective and retrospective observational studies on disease states, as well as comparative effectiveness research. It will analyse how medicines and treatments already on the market are working in a number of disease areas, with a special emphasis on chronic illnesses. It will also provide insight into the types of new therapies most needed for treating and preventing disease [17].

With the advent of personalized medicine, the patient is moving more and more into the spotlight. Increasing

importance is being put on patient-reported outcomes, including those posted on social media such as Twitter, Facebook and patient forums. With technological advances, the use of automated sensors and smart devices is becoming more and more prevalent. In particular, smartphones are becoming point-of-care diagnostic tools through the development of new healthcare-related apps, as well as add-on diagnostic sensors that use the smartphone as an enabling platform.

In addition to these external resources, pharmaceutical companies have a vast array of internal data, ranging from basic laboratory research to elaborate clinical trial programmes, which have not been fully analysed and sit idle in corporate data silos. Several organizations are now starting to make some of their clinical data available to outside researchers for further analysis. Project Data Sphere (http://www.projectdatasphere.org), for instance, is aimed at making historic phase III comparator arm cancer data and analytic tools broadly available [18], while several large pharmaceutical companies have joined forces and made their data available to interested researchers via http://clinicalstudydatarequest.com [19]. Other initiatives include an agreement between Johnson & Johnson and the Yale School of Medicine to provide a mechanism to make clinical trial data more widely available [20].

Another element that is often highlighted is velocity. Velocity refers not only to the ability to access data quickly but also to how fast data change over time and new information becomes available. While real-time access is not critical—at least not in the context of gaining insight into disease mechanisms or better clinical trials and better treatment options—the notion of change is clearly relevant. Topics need to be regularly revisited to evaluate any changes in the available data that might lead to new insights and inform new knowledge.

From an R&D perspective, veracity or data quality is also very important. Nevertheless, for most of the data sources currently in use, there are mechanisms in place to ensure quality standards, which will benefit even further through better use of the available data. At the same time, the introduction of patient-reported outcomes (including those posted on social media), as well as self-service diagnostics, will require a more careful approach and probably will require further validation through more conservative channels.

3 Big Data Challenges

The main challenges the industry is facing are associated with the variety of data. First of all, no single organization or company has all of these data available. It is therefore important for companies, the healthcare system and also the academic community to work together. This has been recognized, and many pre-competitive or non-competitive collaborations are taking shape [21].

While excellent systems exist to analyse different data types in isolation, real value can be gained from integrating the data into one harmonized, unified knowledge base.

However, this is where the issues begin. Different data types are stored in different data sources, and these data sources are not necessarily compatible. Data can be structured (as in clinical trial management systems or electronic data capture systems) or completely unstructured (such as free-text documents or patient-reported outcomes posted on social media). Even if the data are structured, the structure of one data source is not necessarily compatible with that of another data source. Another big challenge is the use of different terminologies and taxonomies. For instance, ALT and ALAT both refer to ‘alanine aminotransferase’—or is it ‘alanine transaminase’? Do we talk about ‘gender’ or ‘sex’?

In order for disparate data sources to be consolidated and integrated into a single view of the world, it is important that they are harmonized into a single data framework. Unfortunately, there are several standards in use. While the life science community is now focusing on CDISC (Clinical Data Interchange Standards Consortium) and MedDRA (Medical Dictionary for Regulatory Activities), healthcare systems are more inclined to use Snomed CT (Systematized Nomenclature of Medicine—Clinical Terms), HL7 (Health Level 7, a set of international standards for transfer of clinical and administrative data between hospital information systems), LOINC (Logical Observation Identifiers Names and Codes, a universal standard for identifying medical laboratory observations) and ICD 9 or 10 (International Classification of Diseases Version 9 or 10). Efforts are therefore needed to establish semantic interoperability between these standards or to create a system that can absorb all of these standards into a single common format. The advantage of the latter would be that all other standards would be mapped to the common format. This would alleviate the fact of having to map all standards to all other standards [22].

4 Big Data Information Model

While the information in these disparate data sources and types is certainly heterogeneous, it is also clear that it is all intrinsically connected as it is related to the knowledge domain of medicine. In this respect, this information can be considered to be a large-scale knowledge network of interconnected information units, somewhat akin to the semantic web. Key to the semantic web is the linking of information through meaningful relationships. These

relationships are described in the Resource Description Framework (RDF) through so-called triples—simple sentences composed of a subject, predicate and object, with the subject and object being linked through the relationship expressed in the predicate. In order to overcome the challenge of different terminologies and data structures, the semantic web also introduces the concept of ontology—basically, a structured, well defined framework that models the underlying concepts and relationships explicitly. Medical information lends itself to such an ontology-based approach, and the use of semantic web technology in life science has been well documented [23]. An information model taking advantage of linked information can be simplified by enclosing relevant information pertaining to the same event in a self-contained information unit, providing all necessary information to understand this individual event and linking these self-contained information units instead [22].

5 Data Analytics

Gaining insight from Big Data is all about relevance and context. Therefore, any Big Data analytics project needs to start with a clear question. In this respect, Big Data analytics is like finding a needle in a haystack. In order to have any chance of finding this needle, it is important that you know exactly what this needle looks like. Once relevant data have been identified, the next step is to develop and apply the right analytical methods and models, so that the right conclusions can be drawn from the data in the context of the original question. Since the ever-increasing flood of different data types makes the identification of relevant data increasingly difficult, this is an iterative process where each previous iteration will inform future evaluations.

Data visualization is an important aspect in dealing with data analytics. The old saying "A picture is worth a thousand words" clearly applies. Big Data analytics—or any data analytics, for that matter—is about understanding trends, correlations and patterns. As with data standards, there are also initiatives to standardize some of these visualizations to provide a good foundation.

In order to achieve meaningful insights and identify actionable results, data analytics also needs to move from descriptive business intelligence models to predictive models and ultimately to prescriptive models. Descriptive models are purely aimed at analysing what happened in the past and giving you a good understanding of what was. Predictive models add another layer to this and try to gain insights into how these data might help you to better understand what will happen in the future. Predictive models are trying to provide insights into potential future states. Prescriptive data analytics adds again another layer that aims to provide recommendations on how to proceed, providing true decision support.

The best example of the development of predictive models is the research into biological markers (biomarkers) and the advent of personalized medicine.

Biomarkers are surrogate markers that can be objectively measured and evaluated as indicators of disease susceptibility and progression, safety concerns and therapeutic outcome [24]. Biomarkers can be anything from blood pressure to increasingly complex networks of individual traits [25, 26]. In the context of pharmaceutical R&D, biomarkers can help in the validation of disease targets and identification of suitable patient populations for the development programme, as well as providing early signs of safety issues and efficacy in order to facilitate 'go/no-go' decisions. The use of biomarkers in the development stage can also provide early indications of real-world effectiveness, which will be helpful for evaluation of the commercial viability of a drug early on.

Biomarkers are essential for personalized medicine. In recent years, it has become evident that developing new medicines cannot rely on the 'one size fits all' approach. Patient stratification is becoming a prerequisite not only in the real world but also in the design of successful development programmes.

In the field of prescriptive analytics, there are also projects underway looking at machine learning. For instance, the Memorial Sloan Kettering Cancer Center is working together with IBM to train the latest supercomputer, Watson, to support doctors in making better treatment decisions [27].

6 Big Data and Knowledge Management

In addition to having the capability to gain appropriate insight from Big Data, it is also vital to communicate these insights within the company. Companies must devise appropriate knowledge management strategies that enable the company to maximize the value of their Big Data initiatives. A survey by the Economist Intelligence Unit indicates that 41 % of pharmaceutical executives see knowledge management as one of the main drivers in productivity gains [28]. It is also clear that managerial ability and culture have a major impact on how Big Data initiatives fare [29].

Knowledge management can be divided into three areas: knowledge creation or research; knowledge utilization or new product development; and knowledge transfer or collaboration [30]. Depending on the primary aim of the Big Data initiative, different systems need to be put in place to support these initiatives appropriately. If the primary

objective is the discovery of new drugs, then companies need to look at implementing a personalization strategy that primarily aims to bring people together. Knowledge and information need to be shared in order to inform individuals about the latest advances. The goal is to create embedded knowledge. On the other hand, drug development needs to implement a codification strategy that allows many people to search for and retrieve codified knowledge from a repository without having to trace and interact with the source of knowledge. From an IT perspective, the personalization strategy requires implementation of highly bespoke systems, whereas the codification strategy requires systems that are optimized for data storage and retrieval.

7 Big Data Impact

While Big Data has been around for some time, and data sets in the pharmaceutical industry have always been complex, it is only now that all of the capabilities associated with Big Data analytics are slowly falling into place. The biggest leap to date has been seen in the Health Economics and Outcome Research arena, as the examples of Pfizer and Sanofi show [14, 15]. This can be attributed partly to the fact that a lot of the available data are more transactional in nature and therefore are easier to analyse; partly to the fact that marketing and sales departments have always been more 'customer focused'; and partly to the fact that health information systems and payer systems are now in place that allow for seamless gathering and integration of this information.

In the field of drug discovery, well established systems for the analysis of genomic data are now joined by systems evaluating the whole systems biology sphere [31].

In clinical development, Big Data is starting to make an impact, particularly in relation to patient stratification and recruitment. Evaluation of the available patient information can support the modelling of inclusion/exclusion criteria, as well as helping with the identification of suitable patients.

Moreover, the establishment of integrated systems providing centralized access to all available data is helping with the conduct of clinical trials—in particular, risk-based monitoring. The ability to compare and analyse information gathered from all clinical trial sites in a centralized setting allows companies to better evaluate safety issues, operational shortfalls and outright fraud by individual sites [32].

8 Conclusion

The pharmaceutical industry is only starting to implement Big Data initiatives, and a long road still lies ahead. Nevertheless, the industry has realized that it needs to focus on its main assets: its own data and the other available data. This will assist us to understand disease mechanisms better, define true unmet medical needs and deliver better medicines at affordable prices to an increasingly stretched healthcare system, ultimately helping those who need it most: the patients.

Acknowledgments Peter Tormay is affiliated with Capish Nordic AB, a solution provider of data integration and visualization software to the pharmaceutical industry. No funding was received for the preparation of this article.

References

1. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nat Rev Drug Discov. 2010;9:203–14. doi:10.1038/nrd3078.
2. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. Nat Biotechnol. 2014;32(1):40–51. doi:10.1038/nbt.2786.
3. Getz KA, Campo RA, Kaitin KI. Variability in protocol design complexity by phase and therapeutic area. Drug Inf J. 2011;45(4):413–20. doi:10.1177/009286151104500403.
4. Kaitin K. Deconstructing the drug development process: the new face of innovation. Clin Pharmacol Ther. 2010;87:356–61. doi:10.1038/clpt.2009.293.
5. Pammolli F, Magazzini L, Riccaboni M. The productivity crisis in pharmaceutical R&D. Nat Rev Drug Discov. 2011;10:428–38. doi:10.1038/nrd3405.
6. Hu M, Schultz K, Sheu J, Tschopp D. The innovation gap in pharmaceutical drug discovery & new models for R&D success. 2007. http://www.kellogg.northwestern.edu/biotech/faculty/articles/newrdmodel.pdf. Accessed 10 March 2015.
7. Munos B. Lessons from 60 years of pharmaceutical innovation. Nat Rev Drug Discov. 2009;8:959–68. doi:10.1038/nrd2961.
8. Laney D, Beyer MA. The Importance of 'Big Data': a definition. https://www.gartner.com/doc/2057415/importance-big-data-definition. Accessed 16 November 2014.
9. Durbin RM, Altshuler DL, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467:1061–73. doi:10.1038/nature09534.
10. Torjesen I. Genomes of 100 000 people will be sequenced to create an open access research resource. BMJ. 2013;347:f6690. doi:10.1136/bmj.f6690.
11. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, et al. Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. Nat Genet. 2007;39:1329–37. doi:10.1038/ng.2007.17.
12. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447:661–78. doi:10.1038/nature05911.
13. Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, Robson S, et al. Genome-wide association study of CNVs in

16,000 cases of eight common diseases and 3,000 shared controls. Nature. 2010;464:713–20. doi:10.1038/nature08979.
14. Phillips B, Roberts C, Rudolph AE, Morant S, Aziz F, O'Regan CP. Switching statins: the impact on patient outcomes. Br J Cardiol. 2007;14:280–5.
15. Sanofi. Sanofi reports results of new meta-analysis reinforcing Lantus® safety profile at the World Diabetes Congress [press release]. 2011. http://en.sanofi.com/Images/29269_20111207_Lantus_en.pdf. Accessed 2 March 2015.
16. McKinsey. The big data revolution in healthcare. 2013. http://www.mckinsey.com/~/media/mckinsey/dotcom/client_service/healthcare%20systems%20and%20services/pdfs/the_big_data_revolution_in_healthcare.ashx. Accessed 2 March 2015.
17. AstraZeneca. AstraZeneca and HealthCore announce real-world evidence data collaboration in the US [press release]. 2011. http://www.astrazeneca-us.com/media/press-releases/Article/20110202-astrazeneca-and-healthcore-announce-realworld-evidence. Accessed 2 March 2015.
18. Hede K. Project data sphere to make cancer clinical trial data publicly available. J Natl Cancer Inst. 2013;105:1159–60. doi:10.1093/jnci/djt232.
19. Strom BL, Buyse M, Hughes J, Knoppers BM. Data sharing, year 1—access to data from industry-sponsored clinical trials. N Engl J Med. 2014;371(22):2052–4. doi:10.1056/NEJMp1411794.
20. Johnson & Johnson. Johnson & Johnson announces clinical trial data sharing agreement with Yale School of Medicine [press release]. 2014. http://www.jnj.com/news/all/johnson-and-johnson-announces-clinical-trial-data-sharing-agreement-with-yale-school-of-medicine. Accessed 3 March 2015.
21. Altshuler JS, Balogh E, Barker AD, Eck SL, Friend SH, Ginsburg GS, et al. Opening up to precompetitive collaboration. Sci Transl Med. 2010;2(52):52cm26.
22. Berg A, Dahlbo C. The Capish information model—simplify access to your data. Pharmaceutical Users Software Exchange (PhUSE) Annual Conference; 12–15 October 2014; London. http://www.phusewiki.org/docs/Conference%202014%20PP%20Papers/PP35.pdf. Accessed 10 March 2015.
23. Cheung K-H. Prud'hommeaux E, Wang Y, Stephens S. Semantic Web for Health Care and Life Sciences: a review of the state of the art. Brief Bioinform. 2009;10:111–3. doi:10.1093/bib/bbp015.
24. Strimbu K, Tavel JA. What are biomarkers? Curr Opin HIV AIDS. 2010;5:463–6. doi:10.1097/COH.0b013e32833ed177.
25. Guo NL, Wan Y-W. Network-based identification of biomarkers coexpressed with multiple pathways. Cancer Inform. 2014;13:37–47. doi:10.4137/CIN.S14054.
26. Jin G, Zhou X, Wang H, Zhao H, Cui K, Zhang X-S, et al. The knowledge-integrated network biomarkers discovery for major adverse cardiac events. J Proteome Res. 2008;7:4013–21. doi:10.1021/pr8002886.
27. Bassett J. On cancer: Memorial Sloan Kettering trains IBM Watson to help doctors make better cancer treatment choices. 2014. http://www.mskcc.org/blog/msk-trains-ibm-watson-help-doctors-make-better-treatment-choices. Accessed 10 March 2015.
28. Economist Intelligence Unit. Foresight 2020: economic, industry and corporate trends. 2006. http://graphics.eiu.com/files/ad_pdfs/eiuForesight2020_WP.pdf. Accessed 10 March 2015.
29. Hopkins MS. Big Data, analytics and the path from insights to value. MIT Sloan Manag Rev. 2011;52:21–2.
30. Hansen MT, Nohria N, Tierney T. What's your strategy for managing knowledge? Harv Bus Rev. 1999;77:106–16.
31. Robinson SW, Fernandes M, Husi H. Current advances in systems and integrative biology. Comput Struct Biotechnol J. 2014;11:35–46. doi:10.1016/j.csbj.2014.08.007.
32. TransCelerate Biopharma Inc. Position paper: risk-based monitoring methodology. 2013. http://www.transceleratebiopharmainc.com/wp-content/uploads/2013/10/TransCelerate-RBM-Position-Paper-FINAL-30MAY2013.pdf. Accessed 2 March 2015.