For reprint orders, please contact reprints@future-science.com

Drug discovery and development in the era of Big Data

First draft submitted: 8 August 2016; **Accepted for publication:** 12 August 2016;
Published online: 26 September 2016

Keywords: Big Data • data science • drug discovery • informatics • interdisciplinary R&D

Jurgen Bajorath, University of Bonn

Q To what extent do you believe that the application of Big Data can help stem high attrition rates & stagnant pipelines in pharmaceutical industry?

This is uncertain. The underlying assumption would be that increasing amounts of increasingly complex discovery data will help lower attrition rates. Over the past decade, essentially exponential data growth has already occurred, but attrition rates have not declined and pipelines not been filled. The key will be to learn how to extract knowledge in a better way from data and exploit this knowledge for discovery. One should also consider that biology and bioinformatics have been dealing with Big Data phenomena for more than a decade. By contrast, Big Data is still news for chemistry.

Q What are the challenges of implementing end-to-end integration of all data generated during the pharmaceutical R&D process?

Domain expert knowledge is required in addition to technical know-how when building computational infrastructures. Scientists from all disciplines within the drug discovery path are required to participate. Ultimately, it is an interdisciplinary endeavor that is equally challenging from a scientific and technical perspective.

Q What are the difficulties of enriching in-house data resources with external datasets?

Given the proprietary culture of drug discovery, many investigators in the pharmaceuti-

cal industry are – in my experience – not used to working with public domain data. Often the value is not understood (after all, the data are for everyone and not proprietary ...) and there might be only little internal rewards for investigating public domain data. This mind set needs to change and large-scale data mining approaches should be more seriously considered in discovery environments.

Q To what extent do different data definitions & data models impact the analysis of pooled data from disparate sources?

To a great extent. Many of those who currently speak about or promote Big Data are probably only partly aware of all criteria that affect data definitions and reliability. Handling Big Data requires serious science, as illustrated by the emerging field of 'data science'. Furthermore, it is essential to clearly distinguish experimental data from modeled data. This should be self-explanatory, but the line is often not clearly drawn.

Q Is there a need to regulate biomedical data collection, particularly with regard to Big Data consortia, to ensure that data are comparable & can be effectively integrated?

Concerted community efforts will be helpful. However, in my view, focusing on data collection alone will not be sufficient. The field is in need of well-defined data standards and transparent curation methods. Implementation is another issue. A few pharma companies are beginning to exchange screening collections and open up

Jürgen Bajorath*,¹

John Overington²

Jeremy L Jenkins³

Pat Walters⁴

Author affiliations are listed at the end of this article.

**FUTURE
SCIENCE** part of

fsg

to larger scale collaborations with academic institutions but these initiatives typically do not integrate data. Generating proprietary in-house data using alternative compound sources probably is the main motivation.

Q Does the application of Big Data call for a more collaborative & interdisciplinary approach to drug R&D?

Without any doubt. As stated above, the notion and utilization of data on a large scale is still in its infancy in chemistry-based drug discovery and collaborative efforts do not (yet?) address Big Data issues. But it is clear that thinking about Big Data calls for further increasing interdisciplinary actions. Otherwise, the whole business does not make much sense. If the Big Data concept should impact drug discovery, data from different disciplines and stages of the discovery process could not possibly be generated in isolation or analyzed outside of an interdisciplinary context.

Q Do you feel that, in general, drug discovery specialists are enthusiastic to incorporate Big Data analysis into their design strategies?

It will definitely require more work to go off the beaten path and pay more attention to large volumes of internal and external data and their complexity and heterogeneity. Further increasing workload is probably not *per se* accepted enthusiastically by practicing drug discovery scientists, especially if the benefits are not 'immediate'. Data science specialists will most likely be required.

Q Do you feel that drug discovery researchers generally lack the appropriate training to work in a highly data-centric environment?

Yes, in my view, this is certainly the case. For example, chemists are not trained to work with large amounts of internal and/or external data in their projects. Integrating more extensive informatics education into chemistry curricula would be a good start, but will ultimately be insufficient. Domain experts will be required who should best be chemists by training and then experience graduate education in data science. In practice, they need to function in project teams. First specialized educational programs moving into this direction are beginning to appear such as the EU funded Horizon 2020 'BIGCHEM' project (www.bigchem.eu) that aims to integrate academic and industrial education in chemical informatics for drug discovery. However, these are currently small-scale efforts that hopefully should have a pilot function. Data science needs to enter drug discovery environments on a grand scale.

John Overington, Stratified Medical

Q To what extent do you believe that the application of Big Data can help stem high attrition rates & stagnant pipelines in pharmaceutical industry?

It crucially depends on what type of Big Data, but for me part of this aspect is in reliable access to the raw and highly diverse drug attrition data. Part of the issue is that full text mining is still out of reach due to both methodological, but also data access rights and the transient nature of many documents on the web (as company websites or press releases are changed or disappear for example). The move toward Open Access papers and preprint servers for the published literature is very promising, as is reporting of clinical trials in places such as ClinicalTrials.gov. Historically, attrition data have been collated by commercial vendors and licensed on a restrictive basis, this also holds back deeper analysis of attrition data. Key is distinguishing target-based versus compound-based attrition, but techniques such as Mendelian Randomization approaches to link genetic change to disease risk could go a long way to making this 'bad compound' or 'bad mechanism' argument go away.

A second major issue is related to the so called 'reproducibility crisis' – this again points to larger quantities of data, so that individual erroneous data points can be identified as outliers, but regardless, data analysis methods will need to be robust to noisy and incomplete data.

However, my feeling is that machine learning approaches, once unleashed on suitable data will be able to make substantial progress on attrition.

Q In what way do you think the 'Big Data era' has impacted pharmaceutical drug discovery & development so far?

Probably the first wave of really large-scale data has been in the genomics space, where it is clear that direct and testable insights into both common and rare diseases can be developed. The launching of many national scale projects for large-scale sequencing, and even in direct to consumer genotyping data, will make a change to the way we investigate disease.

Q What are the difficulties of enriching in-house data resources with external datasets?

It is often trivial things that get in the way of easy integration, the moving of an FTP site, or change in format or identifiers. It is certainly easier to integrate molecular objects (compounds, genes, etc.) as opposed to more phenotypic data (such as symptoms, diseases and physical measurements, among others) where significant effort often has to be expended in

data mapping. A second issue is where the same primary data is loaded and transformed into secondary sources, this leads to the same data being loaded multiple times, potentially leading to an overestimation of its reliability. We are generally cautious of using 'secondary' data sources, and prefer to integrate 'primary' sources – although the secondary resources often attempt integration or consolidation of many underlying resources, the inevitable changes that happens during this progress make tracking data complicated.

Q To what extent do different data definitions & data models impact the analysis of pooled data from disparate sources?

There is still substantial need to work on better dictionaries, definitions and ontologies for phenotypic data, and this is where current differences really bite when it comes to real-world use of the data. A further factor is that inevitably the more variables that exist in performing an experiment, the more potential variance there then is in measured outcomes – in general these can be estimated by analysis of variance of data – for example the variance in biochemical *in vitro* assays is lower than for cell-based activity data which again is lower than *in vivo* data. Potentially a hierarchical confidence model can be applied to account for this.

Q Is there a need to regulate biomedical data collection, particularly with regard to Big Data consortia, to ensure that data is comparable & can be effectively integrated?

This is a really complicated area, and one that will take significant time to play out – there will be inevitable national differences in required regulation, and the overhead of regulatory style conditions applied to early stage discovery data could slow down and make significantly more expensive early research.

Q Do you feel that, in general, drug discovery specialists are enthusiastic to incorporate Big Data analysis into their design strategies?

It is still early days with thinking through the implications of large-scale data analytics into much of planning of early-stage drug discovery. Especially as it starts to include novel issues such as cloud-based versus local storage of data which can have large cost implications. However, the industry certainly is looking for more effective and successful ways of doing things.

Q Do you feel that a more data-centric approach to drug discovery reduces scope for innovation?

Not at all, it's possible to experiment and play with data in the same way that it is to experiment with the

physical world. There is plenty of scope to find faster ways to solve a data analysis problem, and many classical approaches do not scale to very large, or sparse datasets. So there is still plenty of room to innovate.

Q What does the Big Data era mean in terms of the quest toward personalized medicine?

It is probably not long before many people have access to some form of genotype data – and the consequential prospects for more personalized and safe use of current and development of future medicines are incredibly exciting. It probably is the case that the technology will run ahead of the actual delivery and guidance, but undoubtedly there will be a major transformation in the way that the population is treated and remains healthy.

Jeremy L Jenkins, Novartis Institutes for BioMedical Research

Q In what way do you think the 'Big Data era' has impacted pharmaceutical drug discovery & development so far?

We are seeing the rise of interdisciplinary computational scientists that want to leverage and combine large-scale data. Data scientists in drug discovery can begin to ask 'meta' questions of our aggregated 'omic-scale datasets, cutting across chemistry and biology to a systems-level understanding. For example, compound profiling across hundreds of diverse cancer cell lines where we have quantified the genomic, transcriptomic and proteomic states of the cells has resulted in an unparalleled ability to discover drug mechanisms that stratify within cancer pathways or lineages. In this way the sequencing revolution has impacted cheminformatics as well as bioinformatics. As another example, large-scale compound–target bioactivity knowledge bases are being used to derive high-value compound tool sets with defined mechanisms-of-action to interrogate biology in complex, lower-throughput phenotypic assays. Finally, the advances in computational power and availability of Big Data sets has helped to fuel the evolution of machine learning methods, which are impacting lead discovery, as well as target and biomarker identification. Computational power is also driving the routine use of previously high-end virtual screening and docking approaches, such as docking to hundreds of protein x-ray crystal structures.

Biocuration as a discipline has emerged as a consequence of this 'Big Data era'. Curators and data scientists must work hand-in-hand, for example to transform unstructured data to machine-friendly form for data mining or to map identifiers between ontologies.

Q What are the challenges of implementing end-to-end integration of all data generated during the pharmaceutical R&D process?

There are many hurdles, both technical and cultural:

- To truly integrate across experiments and platforms, all data must be captured along with standardized metadata terms using a bioassay ontology that describes the key experimental components and methods (e.g., assay type, target, format, readouts) with a solid definition of result types (e.g., IC_{50} , percent_inhibition). In the pharmaceutical industry, the challenge is how to implement a simple yet effective and flexible assay registration system that enables such metadata capture for scientists, with appropriate support from biocurators;
- Especially in chemistry, pharmaceutical legacy systems are often oriented toward data browsing rather than data mining. Bioinformatics, as a younger discipline, suffers less from the problem of data and metadata formatted to public standards given its origins in academia. The legacy systems problem is most severe in development, where document formats persist and database integration has not been a focus or even discouraged historically;
- Incomplete knowledge of the myriad of platforms generating data and their data types limits the ability of data engineers to solve the integration problem without heavy input from data analysts and scientists generating the data. Further, institutional long-term commitment to data integration must be persistent in the face of constant organizational changes;
- Data can exist on many levels, from raw to normalized to aggregated, or one step further to summary conclusions after application of domain-specific statistical methods. While most organizations have been successful in creating pipelines for the first few levels of data, there are fewer mechanisms to integrate the postanalysis data level, for example, the statistically significant 'hits' in a screen (real or virtual). And yet, it is this level of data that is truly amenable to integration across heterogeneous platforms. However, data inference methods to draw and store 'conclusion-level' knowledge from aggregated data are slow to creep into the pharmaceutical industry, due to the slow uptake of metadata standards. While advances have been made with approaches like semantic web, there is a continuous burden on data scientists to implement rules or heuristics that automate knowledge across the R&D process.

Q What are the difficulties of enriching in-house data resources with external datasets?

Internalizing external data from public or commercial data sources immediately creates an obligation to format and store internal data in a way that's amenable to such integration. Pharmaceutical data spans backward for several decades, and therefore experimental metadata was not collected or stored according to modern standards of reference. Drug discovery researchers recorded experiments with hand-written descriptions in notebooks, not in databases with controlled terminology. Compound bioactivity data predates even the robotic automation that enabled screening of large compound libraries, whereas in contrast, the National Center for Biotechnology Information reference database Entrez Gene was not released online until 2003. It is unlikely that until the 21st century, the pharmaceutical industry would have been able to integrate public and private compound datasets at all. Despite this challenge, data integration becomes nearly impossible without dutifully migrating proprietary systems toward public references created by national or international efforts (e.g., National Center for Biotechnology Information, EBI and IUPAC). In that way standard identifiers for genes, proteins and compound structure can be used as conceptual handles to integrate data across internal and external datasets.

In many ways, commercial vendors have also helped expedite standardization. For example, controlled vocabularies around compound classes, mechanism-of-action and target identifiers were introduced by commercial compound bioactivity datasets such as the MDL Drug Data Report, WOMBAT (Sunset Molecular), GOSTAR (GVK Bio) and StARlite (Inpharmatica/Galapagos), which gave rise to ChEMBL. Most importantly, these were databases that could be queried by computational scientists rather than closed systems with licensed interfaces, such as SciFinder (CAS).

These efforts exposed arcane data collection practices in the pharmaceutical industry and inspired efforts to standardize data and metadata capture through assay registration systems.

Worth noting is that despite efforts to standardize chemical structures, there remains serious challenges to establish ground truth for compound structures, particularly for structures with many possible stereoisomers, as erroneous or inaccurate chemical representations have been found to propagate across data sources.

Q To what extent do different data definitions & data models impact the analysis of pooled data from disparate sources?

Pooled data may or may not be readily analyzed if data is collected according to different standards or models.

Even when standards and models are agreed upon, data aggregated across different laboratories, locations or dates can present a challenge depending on the level of data representation in the data hierarchy. One possible solution is to integrate summary assertions derived from data rather than attempting to integrate normalized data values from different laboratories. For example, instead of attempting to aggregate IC₅₀ values for Compound A against Target X across all literature and internal experiments, you might instead infer Compound A – inhibitor – Protein X with a particular confidence value inferred from rules about the strength of evidence from all data points for Compound A. In that way data published according to different data models, such as alternative algorithms for calculating an IC₅₀, may still feed into an overall knowledge stream.

Q Do you feel that drug discovery researchers generally lack the appropriate training to work in a highly data-centric environment?

The volume and complexity of data in drug discovery is outpacing the training of our scientists. At the present, Big Data requires computational experts trained to process, analyze and interpret large-scale experiments, who work collaboratively with experimental scientists. It is not tenable for many of our individual scientists to have deep expertise in both computational science as well as disciplines like chemistry or biology. While continuous learning should be every scientist's goal, the most efficient teams are those that combine 'wet' and 'dry' lab experts that leverage one another's capabilities to advance drug discovery projects.

Q Are there certain therapeutic areas where the application of Big Data in drug R&D is particularly beneficial?

We have seen that oncology has benefitted extensively from Big Data in R&D. The long history of kinase family inhibitor design has helped to advance methodologies in structure-based drug design, docking and QSAR. Due to the widespread immortalization of cancer cell lines, we have witnessed 'omic scale profiling of compounds across cancer cells and lineages, which has provided data to evolve methods for curve fitting, clustering, growth rate inhibition metrics and more. Together with cancer cell line sequencing and profiling, oncology researchers have been able to shine a light on compound mechanisms and the cellular features that create vulnerabilities in cancer lineages or driving pathways.

Q What does the Big Data era mean in terms of the quest toward personalized medicine?

Big Data is driving a growing base of knowledge that supports personalized medicine in a couple of

ways. First, whole genome sequencing is yielding clinical variants associated with pathologies, which provides insights into human disease biology. This is fundamentally different from our traditional *in vitro* (cell-based or biochemical) approaches in research. Human phenotypes resulting from genetic variations in genes or regulatory DNA regions point us toward potential drug targets or pathways that could be treated with personalized therapies. Second, big pharmacology data (e.g., compound-transcriptome data, compound bioassay/screening data) is helping to map chemical matter to the modulation of human proteins, which can be leveraged to develop personalized medicines.

However, despite these advances that push us closer to personalization through a genetic understanding of each individual's biology, we are not yet making equivalent leaps forward in our ability to make medicines. There is a limit to how much we can repurpose existing drugs for new indications. While we can sequence a human for US\$1000 in under 1 day, the timeline for developing new drugs remains on the order of a decade. Until we can rapidly engineer drugs tailored to individual genotypes, or solve limitations to the *in vivo* delivery of other therapeutic modalities, there will remain a gap between 'personalized' and 'medicine'.

Q Over the next 5–10 years, how do you envisage the field of Big Data progressing?

We will increasingly tailor structure-based design efforts toward individual genotypes. In drug discovery research, Big Data will progress toward a holistic mapping of chemical feature space to target space. Today we can create chemistry-based models for only a subset of the pockets in the proteome amenable to ligand binding – due to lack of chemical knowledge or protein structure knowledge, new breakthrough technologies may emerge, for example to evolve chemical matter under selective pressure to bind proteins or to modulate phenotypes – either virtually or experimentally. This will generate massive datasets with statistical models that allow us to efficiently drive medicinal chemistry for novel targets. Such efforts will work hand-in-hand with variant genomics in acknowledgment of the genetic differences from individual to individual.

Big Data, if properly linked to 'big metadata', will also benefit from the advances in machine learning and computing power through cloud computing, emerging computer chip and storage architectures, and alternative computing structures.

Increasingly, as R&D becomes data-centric at an unprecedented scale, successful organizations will employ data analysts on par with data producers.

Pat Walters, Relay Therapeutics

Q To what extent do you believe that the application of Big Data can help stem high attrition rates & stagnant pipelines in pharmaceutical industry?

The availability of data from public sources such as ChEMBL and Pubchem can provide a broader medicinal chemistry perspective. In the past, a medicinal chemist was limited to the data available in his/her organization, coupled with information gleaned from reading the literature. In the current environment, individuals working in drug discovery can mine these databases and utilize experimental data to answer practical questions. The availability of biological activity data also provides a benefit to those developing predictive models and enables others to reproduce and extend these models.

Q In what way do you think the 'Big Data era' has impacted pharmaceutical drug discovery & development so far?

To date, the impact has been somewhat limited. One of the most significant impacts of the availability of 'Big Data' has been the matched molecular pairs concept that was pioneered by Leach *et al.* This technique enables scientists to examine the impact of a particular change in chemical structure (e.g., phenyl to pyridine) on a physical property or biological activity. In the past, this knowledge was typically based on experience gained from personal observation of a limited number of drug discovery programs. With newer computational methods and the broader availability of experimental data, decisions can be derived from work on hundreds of drug discovery programs.

Q What are the challenges of implementing end-to-end integration of all data generated during the pharmaceutical R&D process?

One of the primary challenges is integrating biological assays. Assay conditions can change over the course of a drug discovery program and it can be very difficult to determine where assays are comparable. Another challenge is reflecting where values from the same assay can be compared. Any assay has an associated error, but databases typically only capture the experimental values (e.g., IC_{50} , Ki) that were derived from the assay. It can be difficult for a data analysis procedure to determine where two results are significantly different.

Q To what extent do different data definitions & data models impact the analysis of pooled data from disparate sources?

When dealing with data from biological assays, results are typically reported in a consistent fashion. Data

with more complex readouts from pharmacokinetic or toxicology studies can be more difficult to capture in a consistent form. Data from phenotypic assays can also be challenging to capture in a consistent fashion that can be easily compared.

Q Does the application of Big Data call for a more collaborative & interdisciplinary approach to drug R&D?

Yes. A broad skill set is necessary to make effective use of Big Data in drug discovery. As proposed by Drew Conway, effective data science requires three different skill sets:

- Domain knowledge – in order to impact drug discovery, one must have the requisite knowledge of chemistry and biology. The challenges encountered in drug discovery will not simply be solved by clever algorithms, one must understand the problems at hand and know where to steer the analysis;
- Math and statistics – it is surprisingly easy to over-train predictive models or to be fooled by trends which are, in fact, not significant. Those taking advantage of Big Data must have extensive knowledge of statistics to avoid potential pitfalls;
- Hacking skills – Big Data is often messy, and making effective use of this data can require a high degree of computational facility.

Q Do you feel that drug discovery researchers generally lack the appropriate training to work in a highly data-centric environment?

In many cases, yes. Scientists who are trained in non-computational disciplines may not have the exposure to the computational tools that are necessary to take advantage of data resources. Integrating disparate data can also require some programming skills. While most experimentalists are not programmers, it is relatively easy for them to acquire some fundamental programming skills that are invaluable in working with Big Data. There are now a wide variety of free programming courses available through online learning platforms such as Coursera. In addition, a number of open source tools are available in the Python and R programming languages to facilitate data analysis.

Q Are you aware of any success stories in the application of Big Data in drug discovery & development?

In my experience, gains from the analysis and integration of public and proprietary data have been incremental. The matched molecular pairs example

mentioned earlier has provided a number of insights that have helped to propel drug discovery programs. Public databases have also become valuable sources of probe molecules. Biologists who wish to interrogate a particular target or pathway can now search databases such as ChEMBL to identify molecules with pre-cedented biological activity. By cross referencing activity databases with commercially available compounds, one can rapidly identify probe molecules that can be purchased.

Q Over the next 5–10 years, how do you envisage the field of Big Data progressing?

Hopefully, successful analysis of the currently available data resources will lead to increased data availability. Positive results may also provide an incentive for granting agencies to provide funds for meta-analyses of public data. Providers of commercial databases will realize the value of data integration and provide application programming interfaces that will streamline data integration. Online resources such as the IPython Notebook will make it easier for others

to reproduce, learn from and extend computational analyses. Scientific journals will provide better methods for publishing computational workflows. New and more powerful software tools will make data analysis available to a broader portion of the scientific community. Statistics, visualization and data analysis will become key components of scientific education.

Financial & competing interests disclosure

J Bajorath participates in the EU-funded Horizon 2020 'BIG-CHEM' Project (www.bigchem.eu). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Disclaimer

The opinions expressed in this article are those of the interviewees and do not necessarily reflect the views of Future Science Ltd.

Author affiliations

¹Department of Life Science Informatics, University of Bonn, B-IT, Dahlmannstrasse 2, D-53113 Bonn, Germany; bajorath@bit.uni-bonn.de

²Stratified Medical, 40 Churchway, London, NW1 1LW, UK; john.overington@stratifiedmedical.com

³Developmental & Molecular Pathways, Novartis Institutes for Biomedical Research, Cambridge, MA 02139, USA; jeremy.jenkins@novartis.com

⁴Relay Therapeutics, 215 First St, Cambridge, MA 02142, USA; pwalters@relaytx.com