BIG DATA Infrastructures for Pharmaceutical Research

Christian Seebode, Matthias Ort
ORTEC medical
ORTEC medical GmbH
Berlin, Germany
{seebode,ort}@ortec.org

Christian Regenbrecht, Martin Peuker
Charité – Universitätsmedizin Berlin
Charité – Universitätsmedizin Berlin
Berlin, Germany
{christian.regenbrecht,martin.peuker}@charite.de

Abstract— Big Data is an emerging paradigm covering production, collection, processing, analysis, access and presentation of huge data-sets: Big Data infrastructures represent an opportunity to approach this paradigm on an organizational level. We describe challenges and opportunities of big data infrastructures for the pharmaceutical industry. Pharmaceutical research and product development are huge investments and require intense exploration and analysis of data. Future trends show that pharmaceutical companies need to develop new methods of data and information processing to quicker and more precisely respond to changing markets and the need of patients and providers. The individual case and patients will become more influential in healthcare delivery and in consequence for the rationale behind these decisions. Our approach of a platform for semantic exploitation of BIG DATA supports a knowledge based infrastructure for deep analysis of clinical information from structured sources and clinical narratives. This infrastructure is able to pave the way towards the necessary big data-management possibilities. Example applications for cancer research will be given.

Keywords-bigdata;pharmaceutical research;outcome oriented medicine;semantic information processing;

I. INTRODUCTION

The pharmaceutical industry faces deep challenges [1]. R&D efforts that lead to successful products (“blockbusters”) are unlikely to repeat in the future. Many of the successful drugs will run out of patent protection in the next few years. Pharma has to re-invent itself and work on new therapeutic strategies and improve the existing ones. Comparative trials are much more difficult and costly than just proving an effect initially.

Mainly it is the economic perspective that drives new inventions. Clinical cancer research for example is required to be very specific and targeted on individual characteristics. Sometimes the differential information is just not available or known. It is a huge effort to identify exact genetic conditions or biomarkers that characterize the sub-class of tumors successfully.

Pharmaceutical companies are desperately looking for innovation and new business models. It is fairly probable that the pharmaceutical industry will not survive based on the same strategy to produce drugs that target a symptom or illness that is accompanied by a low probability for success. There is a tendency to personalize therapies or to follow outcome oriented strategies.

II. OUTCOME ORIENTED MEDICINE

We define outcome as the final result of all medical interventions, procedures and collaborations in medicine. Outcome has indeed always been the main driver behind medical science. However it has been measured in a very coarse scale. Medical science applies statistics and scientific procedures like randomized controlled trials to produce and describe outcome measures. The notion to understand medicine in terms of a population is not wrong but implies that experience with many patients tells a story about individual patients too.

This notion is changing and picks up possibilities, inventions and models from a broad scale of sciences. Looking at the individual in a more precise and straightforward way is an approach that is included in approaches like personalized and/or patient-centered medicine. The level of detail needed to perform individual or personalized medicine is much higher. The results are not measured in terms of large groups but for any individual patient. For example the result of a therapy may be to extend the life of a patient for some month on average. This may not be relevant for an individual patient. Identifying the goals and characteristics of individual patients requires much more analysis and collaboration between healthcare providers and patients.

On the other hand more data and information is needed to identify individual traits of patients in order to separate them from a population. Outcome oriented approaches consider genetic information they may be linked to the health record of the patients in order to stratify a population for a therapy [2].

III. FROM BIG DATA TO BIG INFORMATION IN THE FIELD OF CANCER RESEARCH

According to the data management criteria listed below not every medical discipline delivers the needed precision or detail to perform deep personalization. The other crucial dimension is diversity and origin of data. The human capability to integrate huge amount of diverse information is incomparable. It is difficult to do this integration with
computerized information systems. The example of cancer research shows that diagnostics and procedures are already available to produce a greater level of detail to power personalized medicine.

A. State of the art of oncological research data administration

The amount of data acquired during patient’s medical treatments is immense. For example, data of tissue analyses, medical imaging, and haemograms, etc. add up to terabytes of medical data depending on the specific diagnostic approach. On top of that the often very long medical history of cancer patients adds additional volume and complexity. Especially in large hospitals and university medical centers lots of additional data is generated by the scientific staff.

These data become more and more relevant as it is often generated from patient specific tumor tissue. The big challenge today is an “iron curtain” separating clinical from scientific data. Scientific data is usually administered within individual, separate working groups, eventually uploaded to project specific servers and private clouds.

In smaller projects, data management is often realized using wikis or similar content management systems - systems managing arbitrary contents of web pages. For large-scale projects, more professional solutions are needed. In cases where pharmaceutical companies participate in these projects their proprietary in-house solutions are commonly used.

B. Current topics and their relationship to data management

All of the following topics are influencing the requirements on management of big data. We mention a few of them to document that there is no generic way to understand big data management. Origin, structure and processes are influencing the way how big data management should be analyzed and understood.

- **High throughput of sequencing data.** Within the next 10 years sequencing cancer genomes will become clinical routine in Germany and large parts of Europe. Today these technologies are mainly applied within a scientific environment, meaning that the processes and data management procedures are largely established by the needs of scientist and not by medical practitioners and/or patients.
- **International collaborations.** Large, multinational research projects have been shown to be way more effective than smaller initiatives. Yet, their full potential has still to be unleashed by cross-national unified, data management and data protection strategies.
- **Data protection and security.** Data protection is a challenge in many ways. Indeed there are so many diverse aspects regarding data security that this justifies a separate discussion. To mention just a few that show the influence of data security aspect on big data management. These aspects strike process, quality and intention of big data usage and management:
  1. Incidental finding must not be reported to the treating physician or the patient directly.
  2. Different countries have dissenting laws of the degree of data protection.
  3. In Germany there is no opt-out for this.

Again there is no generic way to handle access and management of big data-sets for data security. The requirements need to be analyzed for particular applications in the big data domain.

C. Future goals of oncological research

Genomics is surely one of the most important influences on the outcome oriented and personalized clinical research. The tendency to understand and integrate clinical records with information from the patient’s own genome sequence together with the transfer of innovation back to medical services will lead to a stratified medicine. Stratification breaks down a population into distinct subpopulations where each population may receive different treatments for the same diagnosis. It is not very improbable that today’s notion of diagnosis based on the population as a whole may have to be revisited too.

IV. BIG DATA

All these influences from clinical medicine, pharmaceutical drug development and medical research require huge quantities of data and information from heterogeneous sources. **BIG DATA** is a term that describes data-sets that are too big to be managed with common tools and procedures used for smaller data-sets. Not only origin but also processing may include spatial distribution of data and of course different data formats. Based on the above description of practices the goal of management strategies should always be defined according to the use case supported. Big data management inherits more generic tasks from general data management like:

- Collection
- Storage
- Integration, Curation
- Access, Retrieval
- Scalability, Distribution of data.

However in order to deliver the value proposition of innovation and improvement for the pharmaceutical industry or cancer research the data-management alone is not enough. The value delivered by big-data sets comes from in depth comprehensive and value generating processing cycles with constant exchange and update of the informational value extracted from big data-sets [4]
• automatic data collection e.g. harvesting data from distributed repositories
• support for conversion to standard formats. E.g. special markup languages used in a particular domain
• upload to public repositories
• data organization and modeling like data warehouses and semantic web. E.g. support for tagging and annotation of data in RDF [5] or OWL [6]
• Access control. E.g. controlling and auditing access across organizations
• Possibility for extension of functionality to access and process more data-sources. E.g. text-mining to access information from unstructured texts
• Support for collaboration across organizations

Figure 1. BIG DATA Process

We document such a cycle in Figure 1. It shows the iteration on data collection in a structured way.
• collect data from various sources based on domain models
• extract information based on the application or research question according to the domain model
• associate or link the information with the already existing fact base
• optionally update domain models or ontologies if there is new information contained in the data that has not been recognized so far

• discover new information according to model change
• integrate and update information in fact base and prepare new iteration

Being able to collect and process big data-sets in a value-oriented process supports a continuous cycle of discovery of new information. This is mainly what drives technological development behind the big data paradigm. Among the many aspects of the big data revolution in healthcare the value-oriented processing and seamless integration of data and information generated by machine with those authored by human beings has the biggest impact on innovation in healthcare [7].

Human contribution is independent of its intention, e.g. text, speech, collaboration. Everything may be processed according to the contributing value it contains.

It is necessary to rethink organizational models and governance structures of all organizations participating in healthcare and scientific research in order to maximize the value. The separation of clinical and scientific data (‘iron curtain’) is just one example of a barrier to innovation. Another one is the traditional exclusion of patients from shared decisions or participation in healthcare delivery. Patient participation and empowerment together with the improvement of health literacy in healthcare delivery will boost outcomes [8].

V. A PLATFORM FOR SEMANTIC EXPLOITATION OF BIG DATA

Our approach for a platform that integrates and processes big data-sets is based on the semantic exploitation of clinical data collected from various resources like hospital information systems, clinical narratives, scientific literature and collaboration platforms.

This platform is developed within a joint project with Charité and Vivantes, - Europe’s leading healthcare providers - and ORTEC medical GmbH. Main driver for development is the support for cohort identification in clinical trials for recruitment and feasibility studies but is not limited to specific use cases. It supports the flexible and just-in-time exploitation of clinical data from various sources, including clinical texts, to leverage unlimited clinical or patient-centered use cases. Knowledge is represented by ontologies and engineered by the domain experts themselves in a straightforward manner in a semantic workbench. Semantically enriched facts are processed by domain-specific applications that support specific use cases. Semantic exploitation of clinical data from structured HIS records and clinical texts provides an up-to-date evidence base for medical research and healthcare delivery.
Physicians and medical experts are able to engineer ontologies and adapt the knowledge needed by an application or use case. For example, patient recruitment scenarios require knowledge that represents study criteria for case selection. The knowledge engineering process itself is always done by validating data and ontologies in an iterative fashion supporting a round-trip engineering. Ontologies can always be validated and tested against the real data flowing into the system from structured sources or clinical texts. E.g. experts identify ontological concepts in texts by just scraping and annotating them.

Cross-cut concerns like collaboration on individual facts, semantic patients records or ontologies support the required flexible administration and collaboration across organizations.

The data integration and annotation possibilities of this platform address lots of the requirements for the management of big data.

Platform and applications are currently piloted in various departments of Charité and Vivantes [9].

VI. CONSEQUENCES FOR PHARMACEUTICAL RESEARCH

Given that scientific progress in oncology is a good candidate for improvement of medical procedures on a more detailed level the results provide a powerful force to drive change regarding the challenges for pharma. The tendency to support individual aspects in clinical research and drug development is reforming the way healthcare is delivered and the pharmaceutical industry may focus on their core competencies to drive that change. The technical possibilities are available however the intelligent application and iterative deployment in changing disciplines is an ongoing challenge.

Achieving conditional licenses based on trials performed on small cohorts [10][11] is only the beginning of a tendency to close the information gap between scientific reasoning and the methods that help to approve and market the products of these efforts. We believe the possibilities to digest and enrich information is a crucial step in rightsizing clinical research and healthcare delivery. The technical requirements for that is the possibility to manage big datasets with the aspects listed above. However innovation can only be successful if it picks up the necessary culture and promotes innovative people and mind-sets. For the pharmaceutical industry this might imply

- Rethink the provision for the pharmaceutical industry away from drug delivery alone [12]
- Promote collaboration and value-oriented research between all stake-holders in healthcare
- Cooperate to establish new regulatory models and procedures that follow the proposition of value-oriented and personalized medicine
- Educate and perceive patients as partners in healthcare and research

There are already some examples that adopt models of shared innovation [13][14]. One important aspects of these models is also the collaboration with patients. Patient-Centered-Infrastructures [16] are technological perspective for patient-oriented and outcome-oriented services in research and healthcare delivery. Information as a first level resource in healthcare reaches the significance of therapies as it directly influences outcomes. The tendency to break down the risky model of information silos which are based on the false and dangerous understanding of values and markets is under way. To appreciate patients as a partner and a valuable source of information at least in chronic conditions influences outcomes [8] and is a way to change the economic situation of the healthcare system too[15].

VII. CONCLUSION

Establish big data strategies for the pharmaceutical industry may unlock a multitude of possibilities. The goal of outcome-oriented research and medicine directly applies these possibilities to improve outcomes and the
economical impact of research and healthcare in general. However the transition from a drug research oriented market approach and clinical trials culture that addresses big markets to a fine grained and personalized development of market offers is not easy. The required technology is available. We propose a platform for semantic exploitation of big data which may contribute to a large-scale integration of data from various sources in the healthcare domain. Finding the right way to apply these technological possibilities in order to drive innovation is still to be done. Patients should perceived as a partner and collaborate in research, innovation and healthcare delivery. Breaking down the informational barriers of information is a necessary step to learn about the possibilities for applying big data strategies for innovation.

ACKNOWLEDGMENT
This work was supported in part by Technologiestiftung Berlin (TSB) and by Europäischen Fond für regionale Entwicklung (EFRE).

REFERENCES