

Paper AS04**Data Sciences Project (Educating for the Future Working Group)**

Sascha Ahrweiler, PHUSE, Wuppertal, Germany

Aldir Medeiros Filho, London, United Kingdom

ABSTRACT

According to Google Trends, the word “Data Science” is currently at a peak interest for worldwide searches. Companies like Uber and Amazon have built entire business models using data science methodology. The healthcare sector has also seen new players especially in consumer devices, which increased awareness of healthy lifestyles by applying advanced analytics. The pharmaceutical industry is adapting to these new technologies. FDA approved or cleared devices are already used mainly for exploratory purposes in clinical trials and create huge volumes of data. We can create valuable insights when we connect these new data sources to clinical data and apply Data Science methods like Machine Learning, Deep Learning or Artificial Intelligence. This paper discusses the Data Sciences Project within Educating for the Future Working Group, with a view to educate the broader PHUSE community in data science techniques in healthcare so they can be prepared in dealing with new challenges.

INTRODUCTION

This paper introduces the scope of work of Data Sciences Project within the PHUSE Educating for the Future (EftF) Working Group, why we decided to first focus our attention on the clinical drug program development domain as well and tentatively delineate and differentiate the potential role of a Clinical Data Scientist within this domain.

The PHUSE Working Group, “Educating for the Future”(EftF) was initiated prior to the PHUSE CS Symposium (CSS) event in 2018. The goal of this EftF Working Group is to keep up with the evolving industry and to educate the PHUSE community at large on relevant topics. Initially these topics are; Design Thinking, Data Engineering and Data Sciences (including Machine Learning and AI).

We propose a big picture historical review of how Clinical Data Sciences emerges from the “Simplicity” paradigm towards the “Multiplicity” paradigm in clinical research methodology. We also briefly explore how Data Sciences and the Clinical Data Scientist can help evolve the clinical drug program development from the statistical era to the fully digital era of medicine.

Our objective is to help to educate the broader PHUSE community about the highly significant specificities of data science for the bio-pharma industry, so they can understand, adapt and evolve their skills to responsibly embrace and promote the new sustainable digital era in clinical drug program development.

Therefore, the educational frameworks are designed to inform the PHUSE community on the importance of topics where the Working Groups feel the PHUSE community has gaps, the details of the topics themselves and how they can be used to drive innovation in the industry.

The Data Sciences group is a spin-off from the earlier founded “EftF: Data Engineering” project. It has taken on the mission to dive deeper into Data Sciences. With the formation of the “Data Sciences” project itself in late 2018, the team has taken on the mission to explore what data science means for the biopharma and healthcare industry and how changes in the digital landscape affect them.

The project takes a holistic approach which tries to address the main challenges we are facing in our data science education efforts, which are:

- The huge diversity of functional and educational background in our industry, specifically amongst the PHUSE community;

- The huge disparity of academic and commercial educational content curricula packages. Most of them are mainly brand specific (including open source) computer and software orientated, with no focus on the use of such tools based on scientific methodology, medical and statistical thinking education.
- Another issue with such academic and commercial training packages, may be due to their novelty, is that their training is rather focused on one-off, speed-delivery short-term “projects”. None of them offers a holistic approach required by the complexities of long-term sequential experimental projects that we face in the biopharma R&D.

The PHUSE Data Science project goal is to offer curated resources and eventually also “learning pathways” towards data sciences in the biopharma industry for different educational and functional backgrounds. The Data Science project aims to develop a website as a one stop shop for talented and self-learners, who want to dive deeper in to biopharma data science matters, and to keep up to date with new educational topics. All curated material will be published on the Working Groups webpage, which can be assessed at <http://education.phuse.eu/data-sciences>

BIOPHARMA INDUSTRY AND DATA SCIENCES

Up to the last decade we divided the biopharma industry into three different Domains:

Two R&D experimental driven domains requiring long-term evidence data collection:

- Pre-clinical – using in vitro and in vivo data
- Clinical drug program development (CLINICAL) – using in human data

And one covering “rapid” analysis of real-time data,

- Post-Marketing using Real-World Data (RWD) and Real-World Evidence (RWE) Data

Commercial BIOPHARMA		
R&D – EXPERIMENTAL “EXPLORATORY” or “CONFIRMATORY” Long term evidence collection Prospective, gold standard RCT		“OBSERVATIONAL” “Rapid” Analysis of Real-Time Data mix observational, prospective, retrospective
Pre-Clinical	Clinical Development	Post-Registration, Pharmacovigilance/Post Market Surveillance
	Phase I to Phase III Clinical trials	“Phase IV” clinical trials, Real-World Data Real World Evidence
	Medical, Regulatory Clinical Operations, BIOMETRICS Pharmacovigilance, ...	Pharmacovigilance, Medical Affairs ^[1] , Marketing, Business Development

Such domains are the result of a long evolutionary process in drug development research throughout the last few centuries. The statistical era of medicine is one of the latest conquests that firmly shaped the three domains above over the last 100 years.

One of the key reasons for the separation into these domains, particularly for NDAs, are the intrinsic nature and complexity of the data at intra and inter-domain levels, the specificity of their regulatory, safety, medical and ethical frameworks and the high diversity of educational background of departmental teams working in each domain.

The context of data acquisition in biopharma R&D is primarily:

- Pre-clinical and clinical data for a New Drug Application (NDA[2]): patient data, with the aim to generate scientific evidence of the safety and efficacy of new biopharma compounds, or
- Real-World Data and Real-World Evidence (RWD & RWE): extend the indication and/or label update for already approved and commercially marketed medicines. It also includes the monitoring of safety signals and pharmacovigilance.

There are fundamental differences in data collection, management, usability and generalizability between those two categories. For the remainder of the paper, we will focus on the highlighted Clinical Drug Development program, which is described in more detail in the following section

Clinical Drug Program Development (CLINICAL)

A very long and rather slow stepwise set of experiments with humans organized in phases. Each phase is composed of several clinical trials. Phases 1 and 2 are primary safety studies and secondary, “exploratory” or “proof of concept”. Finally, the Phase III are confirmatory clinical trials that focus on efficacy.

- **Phase 1- First in Humans** – highly selected healthy volunteers are enrolled, and the drug is tested for safety and pharmacokinetics. The focus in phase I is looking at what the drug does to the human body and what the body does with the drug while subjects are within a very controlled environment. Phase I trials usually include a small number of subjects (typically up to a few dozen).
- **Phase 2 – First-in-Patients**: If a new compound is found to have an acceptable benefit-risk balance at the conclusion of phase I clinical trials, it can then be tested in patients to further explore its safety profile, optimize dose finding and depending on the study design, explore some hints about its efficacy (“proof of concept”).

There is an important cue here, particularly for statisticians and data scientists new to biopharma R&D. For NDAs, first-in-patient clinical trials are exploratory in nature and most often can only recruit, for safety and ethical reasons, patients who did not respond to local standard of care treatments (that often vary significantly between regions [3]). This has important implications for any kind of statistical & data sciences inferences made with this kind of data, particularly when the clinical trial datasets are used as “historical data”. This important matter will be further discussed later on this paper.

- **Phase 3** – They are usually called confirmatory studies, as the study design and its primary study objectives and endpoints rather focus on “therapeutic efficacy assessment” for the intended therapeutic purpose. These phase III clinical trials, as in the previous phases, includes controls groups, who receive a placebo or approved “standard of care” treatment. Usually the statistical design for efficacy phase III trials requires hundreds, sometimes thousands, of patients per study arm. Once again, that may vary significantly according to the therapeutic & labelling indication or regulatory strategy chosen to be followed by the Company clinical drug program development.

Usually patient numbers increase over the course of the different phases. While Phase 1 studies include a few subject (6 to 48), Phase 2 studies include (in the range of 100s) and finally leading to Phase 3 studies, which could include a couple of hundreds of patients.

It is important for Clinical Data Scientists to know, that pre-clinical and clinical drug program development is often tailored and customized according to the regions (NA, EU, Asia) in which the Company intends to market the drug. It is often the case that regulatory authorities from key markets in each region (FDA, EMEA, PPDA and now China and India) have their own specific requirements for pre-clinical and/or clinical data study designs (i.e. inclusion and exclusion criteria, primary endpoints, study procedures based on local standard of care, etc.). What may be relevant for one may not be to another, so increasingly the number of clinical trials performed [4]. Therefore, such factors must be taken in account about the Generalizability (definition provided later on in this paper) of the clinical trial datasets when used as historical data.

CLINICAL DATA SCIENTIST (CDS)

The nature of biopharma R&D (pre-clinical and clinical) and healthcare data (RWE & RWD) are very different compared to the usual kind of data dealt by Data Scientists working in other industries (Amazon, Uber, Airbnb, etc.). Clinical data is special due to scalability, selective patient populations, diversity between various clinical trials. These aspects differentiate with other areas, where data science approaches are effectively utilized.

Due to the important safety element, the biopharma R&D and healthcare data is not a simple commodity. Its misuse has potential to result in significant safety issues for the patient population, sometimes, at the global level.

Closely collaborating with Clinical Data Engineers on the integration and harmonization of these complex web of data sources sets, from the medical, scientific and ethical point of view, require a reasonable understanding of the interactive dynamics of the triad Genome & Exposome & Phenome.

One-offs and/or intermittent data sciences projects, usually common amongst data-scientists in other industries, may come with significantly unwanted consequences due to the highly web nature of the human organism and ecosystem diversity during the acute, remission or chronic phases of the disease. To help the Clinical Data Scientist understand the risk-assessment of complex web of human organism and clinical drug development, here two examples in phase I, the FIAU [5] in the USA in the 90s and more recently the BIA-10-2474 in Europe [6]. These failures are not a failure of data sciences necessarily but highlight the importance of careful planning for clinical trials.

Another challenge is that most often biopharma data is not publicly available due to patent and patient data privacy regulations. Therefore, data science in the biopharma industry requires a modified approach, which the PHUSE Data Sciences project tries to address by creating awareness of this special data domain.

To make the best use of PHUSE's volunteers Data Science Groups time, and our diverse background, we decided a domain stepwise approach. And the first domain we are focusing our efforts on will be the role of Data Sciences and Data Scientists in the Clinical Program Development (Clinical).

After having long explored and debated what Data Sciences means we decided to explore its potential role in the clinical domain, we started to assess how Data Scientist would fit in the actual classical clinical drug program development ecosystem.

The complexity of the clinical data collection, optimization and analysis in a highly regulated environment requires specialized people, who focus on each of these areas.

Based on the usual biopharma constellation of departments involved in the collection and management of clinical trials data we can very schematically for the purpose of this paper, divide them in the triad "operations", pharmacovigilance (PV) and "biometrics".

It is important to remind the Clinical Data Scientist that two separate key databases will be built and run in parallel during the clinical trial data collection: The Clinical Database (managed by Data Management) and Safety Database (managed by the PV department).

BIOMETRICS departments are involved at several levels, before, during and after each clinical trial:

- The **Statistician** ensures optimal study design and plans the statistical analysis of the clinical trial. He or she estimates the required patient sample size needed in a trial to detect and correctly interprets statistically significant differences in safety and efficacy of treatments.
- The **Clinical Data Manager** ensures the collection of high-quality and reliable data. He or she ensures that data is adequately collected, cleaned, and securely stored for further processing.
- The **Statistical Programmer** ensures the optimization of collected data to allow statistical analysis and later submissions to regulatory agencies. He or she transforms collected clinical patient data into CDISC SDTM and ADaM format and prepares the statistical displays.

When defining Clinical Data Science, it is important to understand how Clinical Data Sciences may integrate the above schematized constellations of departments working in throughout the clinical development process. While the focus of the classical biometrics is eventually the regulatory approval in various regional regulatory agencies, the CDS should have awareness of different regional requirements affecting data but could work boundless for his own data insights.

Clinical Data Sciences, for the time being, cannot and will not replace, any of the departments mentioned above. That's because all those departments fulfill very specific regulatory requirements for NDAs submission by all the major regulatory agencies worldwide monitoring clinical drug development by the biopharma industry and providing final assessment for approval or not to the market and labelling.

Such regulatory and scientific methodological requirements led over the years to a web of interdependent internal Study Operational Procedures between all those departments. It helps justify the business model upon which the biopharma industry operates. It is important to remind that such SOP's web is "replicated" and further extended across all supporting vendors services (i.e. Clinical Research Operations (CROs), Central Labs, etc.)

Technologies are significantly expanding the possibilities for clinical data collection and diversifying the toolsets required to manage and analyze it. As that was not complicated enough some of tools or toolsets have a survival span that may last less than the duration of clinical trial and even less the overall span of a clinical program development lifecycle.

We are therefore proposing to use Data Sciences as an overarching cross functional supporting role acting at the planning level to ensure datasets consistency and sustainability within the clinical program development domain throughout its lifecycle. It is important that the Clinical Data Scientist closely collaborates with the classical biometric roles.

We understand and tentatively define the biopharma Clinical Development Data Science (in short from now on CDS) as follows:

CDS is inherently an integrative discipline, ensures well planned traceable data collection, harmonized optimization, integration, analysis and display of different data sources throughout the clinical drug program development. It reduces uncertainty and creates knowledge and their collective use to achieve progressive results in the treatment processes.

In this new digital era, the CDS will help planning by finding a sustainable balance between “Simplicity paradigm” that are the conquests of the statistical era and the “Multiplicity” paradigm” based on innovative proposals from the Data Sciences and Clinical Research Informatics in this new digital era.

IMPORTANCE OF BIOSTATISTICS FOR THE CLINICAL DATA SCIENTIST

A thorough understanding of biostatistical methods is of critical importance for a CDS to ensure traceable and repeatable data insight. With regards to the importance of biostatistics, clinical research historian Harry M. Marks claimed “By all apparent indicators, the second half of the twentieth century represents the “statistical era” of clinical medicine.” [7]. Further to that quote, Marks refers to statistics as a methodology and tools for experimental controlled clinical trials to assess effectiveness and safety for drug evaluation in clinical settings, which emerged after the 1950s as the paradigm of scientific experimentation.

Biostatisticians contributed to the evolution of clinical trials for example in the following:

- Randomized Clinical Trials (RCT) includes i.e. sample size calculations and choosing the proper study design (parallel-group, crossover, adaptive) at the beginning of a trial,
- Descriptive analysis of trial sample characteristics, modelling of dose-response levels, over hypothesis testing (superiority, non-inferiority, equivalence),
- Inference about the estimate of the ATE (average treatment effect) between treatment groups,
- Survival methods like cox-proportional hazards models and Kaplan-Meier estimators,
- Multiple testing procedures to reduce publication bias,
- Sophisticated adjustments of the false-discovery rate in adaptive designs up using meta-analysis of multiple RCTs results to establish a broader picture.

Biostatisticians contributed significantly to address challenges such as clinical bias, missing and un-traceable data, compliance, oversupply of non-reproducible verifiable data and others. Nowadays, with an increased availability of data, researchers face for the reproducibility, replicability and generalizability of their studies whatever within the simplicity or multiplicity paradigm.

Therefore, today clinical drug researchers have the privilege to dispose of enough historical data, to avoid costly past mistakes and potential misuses of new technologies. Or even maybe use new Data Science and Engineering technologies to succeed where “classical statistics”, involuntarily contribute to the failure of far-reaching ambitious studies.

Since all these statistical methods have an impact on the data collection and analysis, the CDS needs to understand the statistical methods and concepts to effectively generate additional data insight.

During the first two decades of the XXI century, the fast-paced technological evolution, opened new opportunities that are challenging [8] the methodological constraints of the “classical statistics” in clinical medicine as well as for clinical drug development programs.

Very schematically we are now, in the start of the third decade of the XXI century, in a process of Assimilation and/or Accommodation phases¹ (depending of the therapeutic domain) between “SIMPLICITY” and “MULTIPLICITY” paradigms in clinical medical research for drug development.

¹ “**Assimilation** in which new experiences are reinterpreted to fit into, or assimilate with, old ideas. It occurs when humans are faced with new or unfamiliar information and refer to previously learned information in order to make sense of it.

In contrast, **accommodation** is the process of taking new information in one's environment and altering pre-existing schemas in order to fit in the new information. This happens when the existing schema (knowledge) does not work and needs to be changed to deal with a new object or situation.

The “statistical era” of clinical medicine, was based on methodologies and tools crafted towards the statistical paradigm of simplicity, in other words, randomize and simplify (ie. hypothesis and then collect pre-specified curated data).

Here the medical statistician plays a key role to ensure that the pre-specified data collected is adequate to answer the pre-specified question by using only one clinical primary endpoint.

Today, driven by technology advancements and an ever-increasing availability of data, researchers are tempted to follow a data mining approach (ie. raw data first then hypothesis), which contradicts the work of medical statistics. The CDS should avoid following this data mining philosophy and follow good hypotheses generation practices as medical statistics did in the past.

THE SIMPLICITY PARADIGM – “SIMPLE” CLINICAL DRUG DEVELOPMENT PROGRAM

In the past, clinical drug development programs were setup in a very simplified manner. The CDS needs to understand the way how clinical studies setup in the past were. The common point across all the clinical trials of the “statistical era” that shaped the “simple” “traditional” drug development, is the statisticians’ injunctions - randomize and simplify.

“Simplify” from the statistician’s point of view, broadly speaking means controlling the experimental settings to narrow the analysis to a single parameter to answer a single straightforward question.

One aim of a clinical trial is to make a calculated judgement about the likely clinical effectiveness results that would be seen if the treatments tested were to be used for all suitable patients. The eligibility criteria for a patient entering a clinical trial should ensure that they are candidates suitable for the treatments being tested and compared. However, due to the controlled experimental setup of a clinical trial the patients selection criteria might be too restrictive to recruit sufficient sample sizes. Hence, in these trials it is generally helpful to relax the inclusion and exclusion criteria as much as possible within the target population, while maintaining enough homogeneity to permit precise estimation of treatment effects. However, patient populations in a clinical trial are very special and selective compared to the general patient populations in the real world to allow evidence-based decisions when comparing two drugs.

In a randomized trial, the set of all randomized patients is known as the ‘intention to treat population’, or the ITT population. This clinical trial study population is intended to represent suitable patients and to be reflective of what might be seen if the treatment was used in clinical practice. Therefore, the ITT population should normally be the basis for inferences about the effectiveness of the treatments.

Significance tests (e.g., chi-square and t-tests) are used to determine the chances of finding a treatment difference as large as the effect observed by chance alone; that is, how strong is the evidence for a genuine superiority of one intervention over another. In hypothesis testing, the null hypothesis and one’s confidence in either its validation or refute are the issue: The basic overall principle is that the researcher’s theory is considered false until demonstrated beyond reasonable doubt to be true... This is expressed as an assumption that the null hypothesis, the contradiction of the researcher’s theory, is true... What is considered a “reasonable” doubt is called the significance level. By convention in scientific research, a “reasonable” level of remaining doubt is one below either 5% or 1%. A statistical test defines a rule that, when applied to the data, determines whether the null hypothesis can be rejected. Both the significance level and the power of the test are derived by calculating with what probability a positive verdict would be obtained (the null hypothesis rejected) if the same trial were run repeatedly (Kraemer and Thiemann, 1987, pp. 22–23). A clinical trial is often formulated as a hypothesis as to whether an experimental therapy is effective. However, confidence intervals may provide a better indication of the level of uncertainty. In the clinical trial setting, the hypothesis test is natural, because the goal is to determine whether an experimental therapy should be used. In clinical trials, confidence intervals are used in the same manner as hypothesis tests. Thus, if the interval includes the null hypothesis, one concludes that the experimental therapy has not proved to be more effective than the control.

For example, patients with high blood pressure would be randomly assigned into two groups, a placebo group and a treatment group. The placebo group would receive conventional treatment while the treatment group would receive a new drug that is expected to lower blood pressure. After treatment for a couple of months, the two-sample t-test is used to compare the average blood pressure of the two groups. Note that each patient is measured once and belongs to one group.

Accommodation is imperative because it is how people will continue to interpret new concepts, schemas, frameworks, and more. Piaget believed that the human brain has been programmed through [evolution](https://en.wikipedia.org/wiki/Piaget%27s_theory_of_cognitive_development) to bring equilibrium, which is what he believed ultimately influences structures by the internal and external processes through assimilation and accommodation.” (taken from - https://en.wikipedia.org/wiki/Piaget%27s_theory_of_cognitive_development on 07 Jan 2020)

The Independent Samples t-test compares the means of two independent groups in order to determine whether there is statistical evidence that the associated population means are significantly different. In the Independent samples t-test, the difference between the observed means in two independent samples is calculated. A significance value (P-value) and 95% Confidence Interval (CI) of the difference is reported. The P-value is the probability of obtaining the observed difference between the samples if the null hypothesis were true.

Hypotheses:

Null: The means of the two groups are not significantly different.

Alternate: The means of the two groups are significantly different.

The Independent Samples T-test is used to test the following:

- Statistical differences between the means of two groups
- Statistical differences between the means of two interventions
- Statistical differences between the means of two change scores

If we are following the “simplicity” paradigm, this primary endpoint concept with the above-mentioned statistical methods would fit the purpose

We may say that the Controlled Randomized Clinical Trial, with a single primary clinical objective and endpoint, is still the flagship of the “statistical era” of simple traditional clinical drug trial development for drug evaluation.

The Clinical Data Scientist may be reminded that the techniques used in this simple traditional clinical program development usually better accommodate small, short-term trials of acute episodes of a disease.

Warning done! Nowadays Clinical Data Scientist and Engineers may use the Multiplicity paradigm to resuscitate and succeed where the methodological statistical constraints limited the scope of the overarching medical studies in the past.

THE MULTIPLICITY PARADIGM – “COMPLEX” CLINICAL DRUG DEVELOPMENT PROGRAM

There are several factors contributing to a complex clinical drug development program.

CLINICAL RESEARCH INFORMATICS

One of them is the significant expansion and combination of toolsets available for collecting, managing, cleaning and analyzing data - EHR, EDC [9], Patient-Report-Outcome (PRO), consumer-grade or approved FDA medical device wearables, CTMS [10], etc.

The CDS must have deep understanding of the clinical research informatics tools and where and how they can be combined and applied. This is a fundamental part of his/her skills set and overarching role to ensure sustainable clinical drug program development.

STATISTICAL THINKING

Another factor of high interest for the Clinical Data Scientist faced with the potential oversupply of data, is the issue of Multiplicity from the statistical analysis point of view.

For example, the progress in wearables’ polyvalence (i.e. smartphone, smartwatches, etc.) creates a tsunami of data increasing the number of possible comparisons.

Definition of Multiplicity provided by Biostatisticians, Alex Dmitrienko and Ralph B. D’Agostino in a review article for the NEJM in May 2018:

“Multiplicity, or the use of many comparisons in a clinical trial, increases the likelihood that a chance association could be deemed causal. This problem commonly arises in clinical trials that have several clinical objectives based on the evaluation of multiple end points or multiple dose-control comparisons, evaluation of several patients’ populations, and other factors. Multiplicity considerations play a central role in the assessment of efficacy evidence in the presence of competing clinical objectives. The more comparisons that are made, the more likely it is that a comparison that appears to be significant will be falsely so.”

SUSTAINABILITY, REPRODUCIBILITY, REPLICABILITY, GENERALIZABILITY DEFINITIONS

The overarching opportunity for the Clinical Data Scientist, is to help their colleagues on the benefits-risk assessment between the multiple toolsets for data collection and the statistical thinking of clinical trials design to ensure the sustainability of the data collected throughout the clinical drug program development.

Here, sustainability means, the reproducibility, replicability and generalizability of data.

Recognizing such terms were a source of great confusion, the USA National Academies of Sciences, Engineering and Medicine, recently delivered a consensus report [11] where they clearly defined these concepts for the modern era where computing tools (hardware and software) are totally pervasive in all domains of research, including clinical drug program development.

As they are fundamental concepts for the modern Clinical Data Scientist to help address the issues of Multiplicity - toolsets and for statistical analysis - in clinical trials in this era of "big data", they are displayed here after:

"Reproducibility is obtaining consistent results using the same input data; computational steps, methods, and code; and conditions of analysis. This definition is synonymous with "computational reproducibility," and the terms are used interchangeably in this report. Reproducibility involves the original data and code.

Replicability is obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data. Two studies may be considered to have replicated if they obtain consistent results given the level of uncertainty inherent in the system under study. Replicability involves new data collection to test for consistency with previous results of a similar study.

Generalizability, another term frequently used in science, refers to the extent that results of a study apply in other contexts or populations that differ from the original one.

Here after a few examples of complex projects & programs where the Clinical Data Scientist may benchmark Multiplicity paradigms from the past and present to extract important learnings.

Past	Present
In chapters 6 and 7 of his "The progress of the experiment" book, Marks take up in greater details on these matters.	Governmental – "All of us" NIH program [12]
In chapter 6 he explores the history of the Diet-Heart study, an experiment that, biomedical researchers ultimately concluded, was too costly and too logistically complex to undertake.	Private sector – Consortium model - "ONE BRAVE IDEA" mega program [13]
In chapter 7, he examines the long and furious debate over randomized controlled trials of drugs to prevent the complication of diabetes: The University Group Diabetes Program (UGPD) study.	Company lead model: GSK The Salford Lung Study: a pragmatic, randomized phase III real-world effectiveness trial in COPD" [14, 15, 16]

FURTHER EXAMPLES FOR COMPLEX DRUG DEVELOPMENT PROGRAMS USING REAL WORLD DATA OR DIGITAL BIOMARKERS

So far, we have discussed drug development programs, where data was acquired from the patient following traditional approaches and recently also complemented by new digital data sources measuring actually the same data, but in a more state-of-the-art method. With continuously increasing new data sources, the digital transformation leads to completely new opportunities. In the future, drug development programs might include digital biomarkers or real-world data sources, which might completely change the face of clinical trials. This will come with a lot of multiplicity challenges as described above.

REAL WORLD DATA SOURCES

Real-world data (RWD) and real-world evidence (RWE) are playing an increasing role in health care decisions. The FDA is already using RWD and RWE to monitor post-market safety and adverse event data. The FDA 21st Century Cures Act passed in 2016, enables the health care community to use RWD and RWE to show potential benefits or risks derived from sources other than traditional clinical trials.

A nice example how to utilize RWD has been presented by Nicole Thorne at the PHUSE US Connect in Orlando (Reference: Data Standards Used for Electronic Submission of Real-World Evidence in New Drug Application (NDA)). She used a real world matched clinical and genomic dataset of patients with the disease under study and explored the predictive value of a specific tumor biomarker alteration to the current standard of care therapy. Using the RWE data, she showed the low response for subjects with the specific targeted tumor biomarker compared to their own collected datasets.

Approaches like in this example in huge RWD sources requires advanced Clinical Data Science techniques.

DIGITAL BIOMARKER

Digital biomarkers can support continuous measurements outside of the clinical environment, which creates new opportunities for patient care and medical research. However, there are still huge steps required to ensure the appropriateness and quality of digital biomarkers as well as their safety and effectiveness. ([reference](#)).

Digital biomarkers are defined by Wang et al as “consumer-generated physiological and behavioral measures collected through connected digital tools that can be used to explain, influence and/or predict health-related outcomes. This excluded patient-reported measures (eg. survey data), genetic information, and data collected through traditional medical devices and equipment.”. The growth in consumer generated healthcare data can lead to new additional insights from a population, which previously have been excluded from the patient data collection process. This potentially can produce a lot of new holistic insights into the development of medical conditions and eventually can completely change the drug development process.

Just as traditional biomarkers, digital biomarkers require validation through hypotheses testing and repeatable results to demonstrate specificity, sensitivity, positive and negative predictive values, before they can be used as endpoints in clinical trials. The huge amount of healthcare data and the corresponding analytical challenges required for this will be a future working area for a Clinical Data Scientist.

CONCLUSION AND OUTLOOK

The EftF Data Sciences project parsed the biopharma industry in three different domains. The first domain we will focus our attention is the Clinical Drug Program Development domain and the evolution based on increasing available data sources.

The role of Data Sciences in R&D clinical drug research, due to its safety profile, is fundamentally different from other industries that popularized the Data Science. (i.e. Airbnb, Uber, etc).

Clinical Data Sciences will not swipe away statistics. Rather the contrary, sound statistical thinking must be in the core of any clinical drug program development design. It will be the sanity check of clinical data sciences and help ensure to create evidence-based decision making.

Clinical Data Sciences will not replace the actual constellation of departments involved in the clinical drug program development. It will rather closely work with them to ensure a sustainable and consistent datasets collection, management and analysis throughout the clinical drug program lifecycle.

Clinical Data Scientists must understand the significant differences in scope and context between clinical drug development for NDAs and RWE&RWD. The CDS must understand the significant differences in scope applied to Pharmacovigilance matters on both contexts.

The CDS must understand that historical trials data from NDAs studies, due to methodology of clinical trials design and ethical constraints, are by definition, built on outlier patients. Therefore, any attempt to assemble RCT historical trials datasets¹ from NDAs studies in data lakes for inferential statistics and machine learning training should carefully assess the medical and scientific value of their proceedings and results.

In the start of the third decade of the XXI century, Clinical Data Sciences is in a process, to borrow Piagets' terminology, of Assimilation and/or Accommodation phases (depending of the therapeutic domain) between “SIMPLICITY” and “MULTIPLICITY” paradigms in clinical medical research for drug development.

Clinical Data Science as the intersection of biopharma and tech, must be aware of the trust issue that affects both industries. CDS' overarching role to help build sustainable clinical drug program development at global level will need to pivot its models around our two key customers, the healthcare community and the patients.

Therefore, empathy and communication skills to link and glue a diverse range of stakeholder's background will be essential in the new Clinical Data Sciences era of clinical drug program development. That will help differentiate the biopharma industry from the quackery and reduce the “garbage in – garbage out” as medical statistics helped do it in the last century.

ACKNOWLEDGEMENTS

This paper has been prepared with a lot of valuable contributions from the entire Educating for the Future: Data Sciences project with outstanding support from our PHUSE project manager Wendy Dobson. Many thanks in alphabetical order to our team members: Alexander Ullmann, Amar Mahidadia, Arteid Memaj, Chidam Kumar, Girish

Regmi , Hong Qi, Iraj Mohebalian, Jaishree Alladi , Karnika Dalal , Katina Manley , Mario Widel , Meenakshi , Murshed Siddick , Nicole Thompson , Nicole Thorne , Prasanna Murugesan , Sameer Bamnote , Shelley Fordred , Sonali Das , Sumesh Kalappurakal , Walter Cedeno , Mekhala Acharya , Tad Lewandowski and Surabhi Dutta who all contributed to the preparation of this paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Sascha Ahrweiler

Aprather Weg 18a

42113 Wuppertal, Germany

Work Phone: +49-202 365323

Email: Sascha.Ahrweiler@phuse.eu

Web: <http://education.phuse.eu/data-sciences>

REFERENCES

- [1] Medical Affairs (i.e. <https://www.medicalaffairs.org>)
- [2] https://en.wikipedia.org/wiki/New_Drug_Application
- [3] <https://ascpt.onlinelibrary.wiley.com/doi/full/10.1111/cts.12631>
- [4] <https://ascpt.onlinelibrary.wiley.com/doi/full/10.1111/cts.12631>
- [5] The Cure That Killed | DiscoverMagazine.com. Discover Magazine. n.d. Accessed 22 Jan 2020., <http://discovermagazine.com/1994/mar/thecurethatkille345>.
- [6] <https://doi.org/10.1136/bmj.i2727>
- [7] “The progress of Experiment – Science and Therapeutic Reform in the United States, 1900-1990” – Harry M. Marks – (1995) – Part II introduction – page 127
- [8] Anderson, C.: The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired Magazine <https://www.wired.com/2008/06/pb-theory/> (2008)
- [9] <https://blog.xclinical.com/a-brief-look-at-electronic-data-capture-in-clinical-trials> (consulted on 17 Jan 2020 – dated 27 May 2019)
- [10] https://en.wikipedia.org/wiki/Clinical_trial_management_system (consulted on 17 Jan 2020 – last page update 21 Oct 2019)
- [11] N Engl J Med 2018; 378:2115-2122. DOI: 10.1056/NEJMr1709701
- [12] <https://www.nap.edu/download/25303>
- [13] <https://www.joinallofus.org/en/program-overview>
- [14] <https://www.onebraveidea.org>
- [15] <https://www.nihr.ac.uk/documents/case-study-delivering-real-world-research-the-salford-lung-study/11555>
- [16] <https://clinicaltrials.gov/ct2/show/NCT01551758>
- [17] <https://doi.org/10.1038/s41533-019-0123-0>
- [18] <https://doi.org/10.1002/cpt.1608>

Brand and product names are trademarks of their respective companies.
