Paper 4648-2020

Statistical Programming in the Pharmaceutical Industry: Advancing and Accelerating Drug Development

Harper Forbes, Hoffmann-La Roche Limited

ABSTRACT

In pharmaceutical product development, statistical programming and the analysis of clinical trial data has always played a key role in helping to deliver medicines to patients. Over the last decade, the industry landscape and analysis needs have changed dramatically. Most organizations have shifted from a one-size-fits-all approach to personalized healthcare, where genetics and other biological information influence individual treatment. Clinical trial design and analysis has become more innovative—and complex—with the use of synthetic control arms, patient reported outcomes, real-world evidence, and biomarker data. Industry data standards under CDISC have created guidance for most data sources but novel endpoints might require additional considerations. Statistical programmers and analysts need to ensure that data is FAIR—findable, accessible, interoperable, and reusable; this is critical to the longevity of an organization who prioritizes their data as an asset. SAS continues to lead the industry in data analysis but its use has also evolved to where it now can be used with other programming languages, contributing to interactive visualizations and automation. This influences how and what we use to analyze data. This presentation provides an overview of statistical programming and analysis in the pharmaceutical industry, including how skills and responsibilities have adapted and advanced the impact of bringing medicines to patients sooner.

INTRODUCTION

With significant changes in drug development over the last decade − regulatory landscape, innovative trial design, genomic diagnostics, external data sources, and new technologies − the impact on the roles and responsibilities for a statistical programmer in the pharmaceutical industry has evolved. Each pharmaceutical company has a particular strategy to get medicines to patients sooner, but more often, we are observing novel endpoints and smaller trials, with traditional molecule development from phase I to phase II to phase III for drug approval being less common due to time and cost. Regulators have also welcomed the pharmaceutical industry's willingness to assess therapeutic benefit using new technologies and novel endpoints (Macdonald 2019). Emerging technologies, global diverse stakeholders, new data types and maintaining FAIR principles are just a few reasons for those in pharmaceutical data science roles to prioritize both technical and behavioral skills. Being a statistical programmer in a rapidly evolving landscape can be challenging but this paper will help provide context and guide those new and existing within pharmaceutical development.

WHERE HAVE ALL THE GOOD TIMES GONE...? (THE PAST)

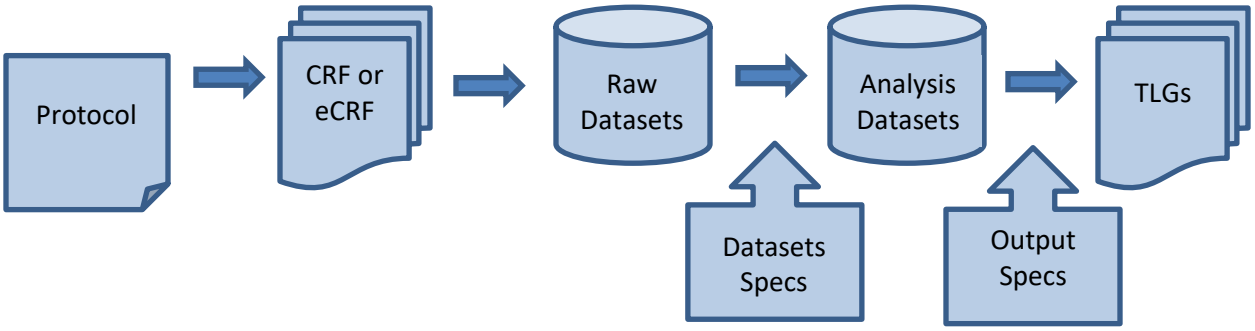
In the past, the role of the statistical programmer was much less complex than it is today. Most companies took the one-size-fits-all approach to treating patients, often targeting disease areas that are most prevalent in markets like the US and Europe. Ideas such as data FAIRness (Findable, Accessible, Interoperable, Reusable) were not necessarily promoted or enforced. Statistical programmers were typically more task-based, creating

static statistical outputs based on specifications. Statistical programming interactions were often limited to the project biometric groups and stakeholders were less diverse.

Some of the main challenges could include deciphering internal data standards and coding from scratch to create statistical summaries in the form of tables, listings and graphs (TLGs). The study team typically provided mock outputs with details such as output format, derivations, aesthetics and subgroups for the statistical programmer to reference. While Clinical Data Interchange Standards Consortium (CDISC) standards were available 20 years ago, pharmaceutical obligation to fully CDISC-compliant submissions was mandated in 2014 when the FDA announced sponsors whose studies start after December 17, 2016 must submit data in FDA-supported formats listed in the FDA Data Standards Catalog (U.S. Department of Health and Human Services Food and Drug Administration 2014).

One of the key responsibilities of a statistical programmer in pharmaceutical drug development is to apply statistical methodology to analyze clinical trial data using statistical software, such as SAS. To help ensure accurate statistical concepts and derivations, along with traceability and transparency, metadata in the form of dataset specifications and output specifications are as important as the code itself.

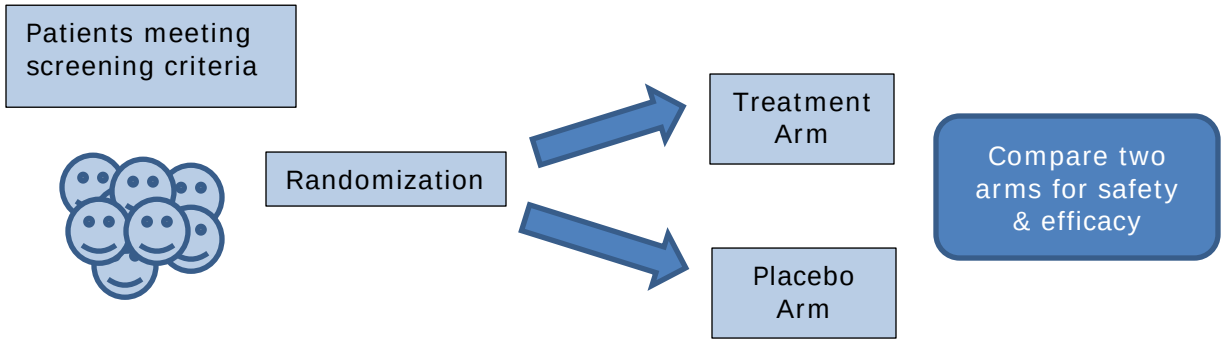
Figure 1 is a simple flow of clinical study information from protocol to statistical outputs.



Personalized medicine was born more than 20 years ago through the ambitious plan to sequence the first reference human genome. This was accomplished in 2003 at the Sanger Centre in Cambridge, for the first time, they had an essentially complete sequence and map of all the genes in the human body (Maxwell 2016). Prior to having the ability to readily determine genetic characteristics within an individual, statistical programming primarily focused on clinical safety and efficacy data collected on Case Report Forms (CRFs) or electronic Case Report Forms (eCRFs), which was data entered at a clinical site or hospital where the trial was being conducted or measured at laboratories.

With many organizations taking a one-size-fits-all approach to drug development, trial design was less complex and involved less treatments.

Figure 2 is a simple schematic of a randomized, double-blind, placebo-controlled trial



Historically statistical programmers were still very busy individual who had to help ensure data quality and coding integrity was maintained but data sources were less varied than today's personalized approach to medicine.

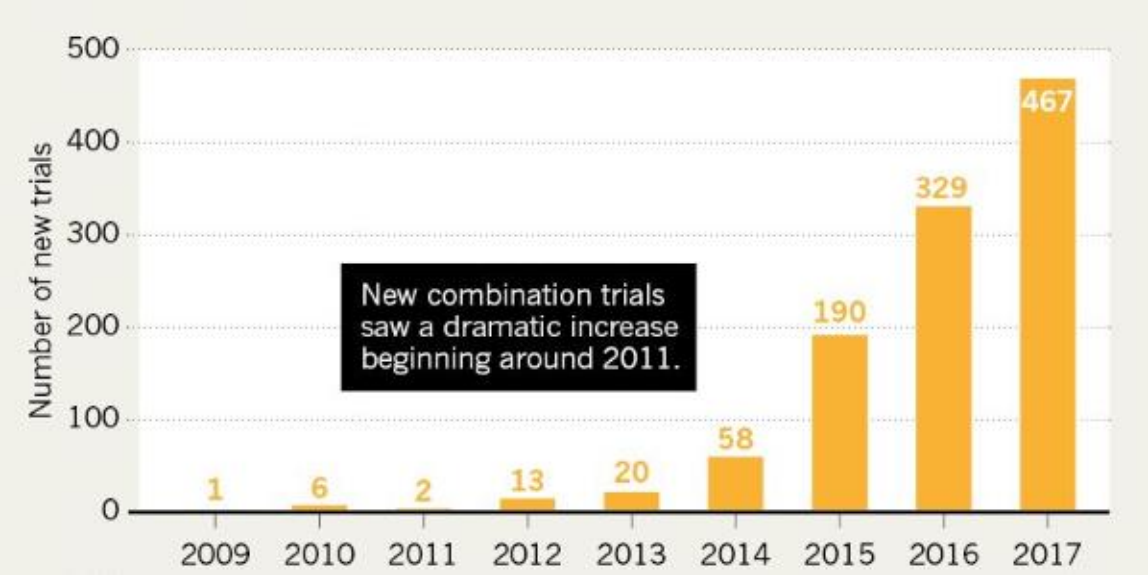
RIGHT NOW (THE LAST DECADE)

The following sections discuss several of the primary changes over the last decade, which have a substantial impact on statistical programmers.

COMBINATION THERAPY

Combination therapy – where patients receive more than one drug treatment - are on the rise. Combination therapies now account for more than 25% of clinical trials in oncology, and trials supported by the National Institutes of Health (NIH) are significantly more likely to use drug combinations than those supported by industry (Wu et al. 2014).

Figure 3 illustrates the rise in combination therapy over the last decade (Schmidt 2017)



The increase in combination treatment regimens administered in a clinical trial affects a statistical programmer in several areas. Each treatment may have associated safety signals to consider and creates an overall increase in number of statistical summaries, such as treatment related adverse events outputs. If teams are resourced by a primary molecule (Trt A) and have regulatory deliverables that may be used in combination with other treatments (Trt B) it is important Trt A team reach out to Trt B programming teams for any programming concerns (Adverse Event Baskets, MedDRA versioning, etc.). Effective and collaborative communication is key to ensuring the statistical programming team has all the information required to do their work efficiently.

COMPUTING ENVIRONMENT, PROGRAMMING LANGUAGES AND AUTOMATION

The biggest change affecting statistical programmers in pharmaceutical drug development is around the computing environment, automation and programming languages. These technological advances have revolutionized the ways of working - and the work itself - of a statistical programmer over the last decade.

It is now common for statistical programmers to work in and across many systems and platforms. Statistical programmers may be run their code in a PC or UNIX-based environment, depending on if the code is in development, validation or production, to help accelerate turnaround and leverage the speed capabilities of one system vs compliance on another. Coding can be held in a combination of stringent version controlled systems or collaborative open-sourced environments. The statistical programmer must be able to work efficiently in a complex computing landscape

Similarly, for programming languages where a decade ago almost all programming was in SAS, we now see much more agile statistical programmers - using SAS, R, Python™ and other languages - sometimes together. Each programming languages having unique capabilities and leveraged based on fit-for-purpose mindset.

Automation is becoming more prevalent, not only with the advancements in technology like powerful computing environments and diverse programming languages, but also increased adoption of standards. Having protocols, CRFs (CDASH), datasets (SDTM and ADaM) and CSR content standardized enables automation to remove any simple or redundant work and increases traceability from methodology to collection to analysis. If an organization has a capable statistical programming group this will be an opportunity to expand their impact by applying their talents outside the usual regulatory outputs.

Comprising components of the three aforementioned technological changes one significant impact statistical programmers have helped accelerate drug development is by production of dynamic visualizations and data displays. If an organization has a suitable framework and skilled statistical programmers these applications can be provided to key stakeholders enabling scientists insights into data sooner, greatly increasing the ability to make informed decisions quickly and potentially bring medicine to patients sooner.

F.A.I.R DATA

The concept of FAIR (Findable, Accessible, Interoperable, Reusable) data principles has achieved worldwide recognition by various organizations including FORCE11, National Institutes of Health and the European Commission as a useful framework for thinking about sharing data in a way that will enable maximum use and reuse (Australian National Data Service 2018). This concept is readily applicable to our own clinical trial data and should be a key mindset of statistical programmers who store, analyze and interpret data. Having

findable data is one of the primary expectations of project information − it provides metadata to help describe and locate the data. The accessibility of the data must be outlined with primary contact names to help make it shareable to an increasing set of stakeholders − other programmers, biostatisticians, data managers, scientists, outcome research, data curators, etc. Interoperability of the data is important so users are aware of the project related agreed formats, metadata and project related details (i.e. Statistical Analysis Plan). Reusable data may refer to the standards implementation − if the clinical data in raw and analysis form are CDISC complaint or still maintaining a legacy format. With the increasing and understandable priority on these principles, it is necessary for statistical programmers to have a framework and mindset in place for this information.

INNOVATIVE TRIAL DESIGN

Accelerating drug development in a competitive regulatory landscape using a traditional approach can be expensive and time consuming. The length of time it takes for a drug to be tested and approved can vary greatly but if following a traditional path it might take 10 to 15 years or more to complete all 3 phases of clinical trials before the licensing stage (Cancer Research UK 2015). Furthermore, traditional randomized phase 3 trials may not be feasible, or ethical in rare disease settings.

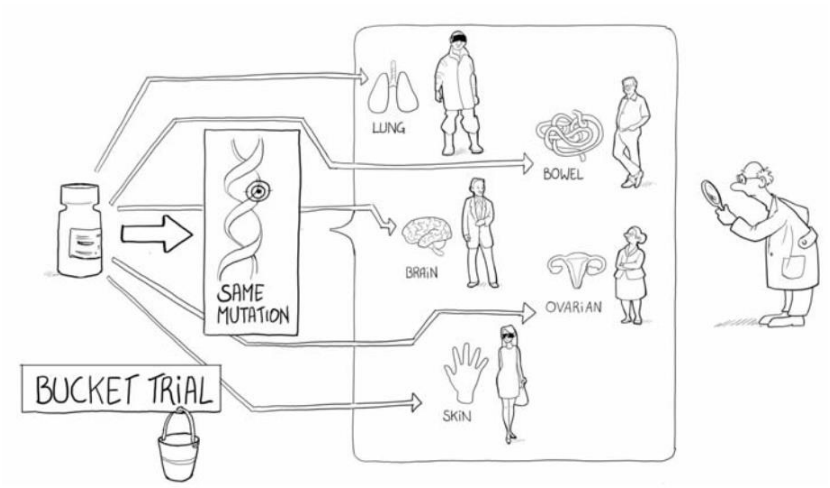
Regulatory authorities have encouraged new approaches to allow medicine to patients in need. To modernize drug development, improve efficiency, and promote innovation, the U.S. Food and Drug Administration (FDA) has initiated efforts focused on advancing complex innovative trial designs (CID), which may provide potential benefit across a range of therapeutic areas. Designs under the CID umbrella include, but are not limited to, complex adaptive, Bayesian, and other novel clinical trial designs, which often require simulations to determine the statistical properties of the trial (U.S. Food and Drug Administration 2019b).

In an umbrella study, the effect of different drugs are tested on mutations in a single type of cancer or disease area. Patients are enrolled based on a genetic mutation in their tumor and treated with a number of medicines known to target this specific mutation. Although complicated, umbrella studies enable researchers to test a number of different treatments in patients with a similar disease. This allows researcher to identify patient subgroups who would most benefit from those medicines tested. For a statistical programmer the database and collection (visit schedules, endpoints) can differ among each tumor type making it difficult to propagate statistical programming due to the number of treatments administered while also needing to consider the various genetic mutations across cohorts.

In a bucket or basket study, a single treatment regimen is administered across different cancer or disease areas on patients with a similar genetic mutation. The design allows researchers to analyze each cancer type individually, as well as assess the effect of the drug or drug combinations aggregated across all cohorts. While less complicated than an umbrella trial programming considerations may need to be adaptive from one cohort to another as primary endpoints may differ across tumor types or disease locations.

In adaptive trial designs researchers are allowed to modify the trial, including adding cohorts and new treatment, based on current scientific information. This allows the effective treatment to be determined swiftly and ineffective or unsafe treatment to be terminated quickly. The design allows flexibility for researchers but statistical programmers must ensure programs are robust to adapt to these potential changes in the study.

Figure 4 illustrates several innovative trial designs



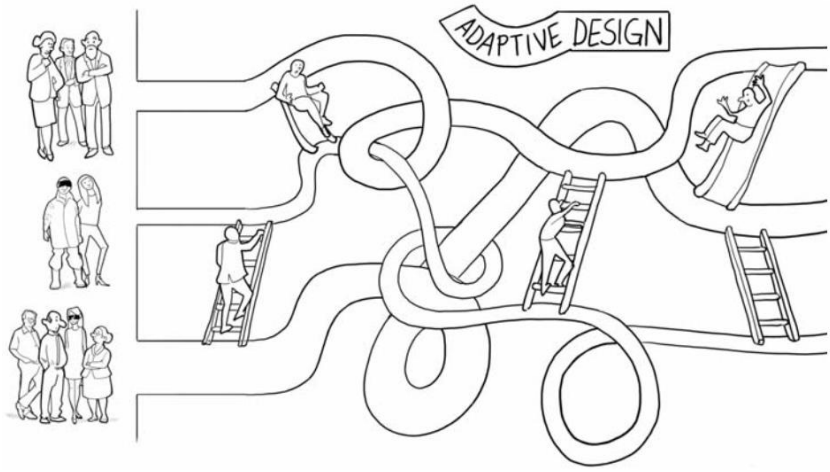


Illustration source:
https://www.roche.com/research_and_development/who_we_are_how_we_work/clinical_trials/innovative-clinical-trial-design.htm

NEW DATA SOURCES

One of the largest changes over the last 10-20 years has been the inclusion of new data types into the analysis of clinical trials for drug development. As noted earlier, the ability to characterize genomic features allows organizations to personalize their treatments to meet the needs of specific patients, especially within disease areas like cancer, where certain cancers may target specific pathways. Biomarker data can be very complex – simply mapping biomarker into SDTMv data can be a difficult due to the necessity for in-depth understanding of the biological assay. Generally, the source biomarker data will be delivered in disparate file formats with inconsistent structure across assays and data sets. This is further compounded by the lack of standards across labs. All of this combined makes it difficult to standardize downstream programming pipelines (Precision for Medicine 2018).

Patient-reported outcomes (PROs) can be included in clinical trials as primary or secondary endpoints and are increasingly recognized by regulators, clinicians, and patients as valuable tools to collect patient-centered data (Mercieca-Bebber 2018). Although their importance is justified, the data can sometimes be difficult to analyze due to collection issues (i.e. sometimes they may be collected using paper or electronically, or both). Also, unless defined as a key consideration in the trial the data quality may end up deprioritized until late in the study lifecycle, usually ending up with statistical programming trying to analyze messy data as appropriately as possible.

Inclusion of real-world data (RWD) and real-world evidence (RWE) into regulatory submissions for new disease treatments is becoming increasingly more common by creating additional insights into therapeutic data. Particularly, in rare disease settings, where randomized trials are not feasible, and observational RWD may be used alternative to a traditional control arm. Real-world data consists of data related to health outcome or care from various sources including mobile devices, smart watches, hospital records and other sources of health information. Real-world evidence is the clinical evidence regarding the usage and potential benefits or risks of a medical product derived from analysis of RWD. RWE can be generated by different study designs or analyses, including but not limited to, randomized trials, including large simple trials, pragmatic trials, and observational studies (prospective and/or retrospective) (U.S. Food and Drug Administration 2019a). The challenge for statistical programmers is that real-world data is not held to the same rigor as a registered clinical trial. Real-world data is more likely to have data quality issues such as missing data, outliers or unreliable metadata or specifications. Depending on how the data was collected it may be very difficult to map to SDTM, making it problematic to merge with other patient data as well as have endpoints or data variables unfamiliar to the clinical trial statistical programmer.

The use of wearable biosensors and smartphone applications capturing large amounts of health-related data has been rapidly accelerating and its use in clinical trials to create patient insights is underway. As of February 2020, clinicaltrials.gov shows that approximately 460 wearables studies are underway, and according to Kaiser Associates and Intel, 70 percent of clinical trials will incorporate sensors by 2025. Pharmaceutical and medical device companies have an opportunity to take advantage of these large data sets and developments to optimize both their clinical trials and their products to improve treatment efficacy (Jansen and Thornton 2020). These large datasets or "big data" will need to be handled by the statistical programmer presenting the challenge of another new data source to understand but also compounded by the size of the data; these analytical considerations may dictate the programming languages and computing environment of the statistical programmer.

Understanding new data sources such as biomarker, patient-reported outcomes, real world data and device data, while liaising with those subject matter experts directly is critical for statistical programmers today. Without this understanding, statistical programmers cannot

help create the appropriate statistical datasets and summaries to generate important medical insights.

NOVEL ENDPOINTS

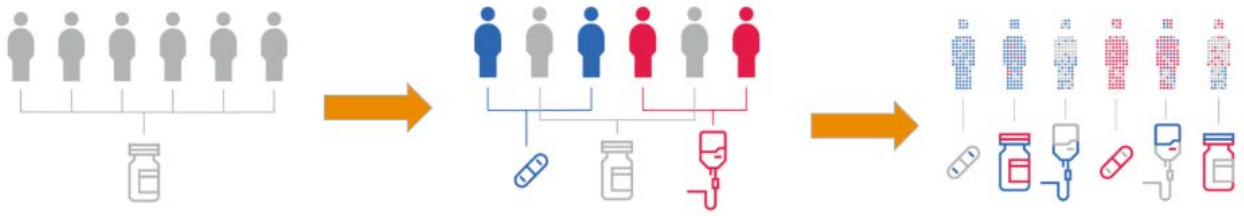
As the drug development landscape continues to evolve, and treatments continue to advance, the use of novel surrogate endpoints is necessary in order to gain faster insights on treatment benefit, allowing pharmaceutical companies to deliver new drugs to patients sooner. In the oncology landscape, overall survival (OS) is still considered the gold standard for assessing treatment benefit. However, the use of conventional surrogate endpoints, such as progression-free survival (PFS), has long been commonplace. More recently, as treatments become more effective, and patients attain greater benefit, even PFS can be an infeasible endpoint, as it requires an inordinate amount of time to reach.

The use of novel endpoints such as metastasis-free survival and minimal residual disease are being used more frequently as primary endpoints in clinical trials (Bennett 2018). It is critical for statistical programmers to have a good understanding of these endpoints and the study objectives in order to work within their organizational data governance to form appropriate CDISC-compliant standards and meet the analysis needs of the study.

TOP OF THE WORLD (THE FUTURE)

Clinical trial statistical programming and analysis landscape has evolved and successful individuals must develop both technical and behavioral capabilities. After all, data scientists have been named the "sexiest job of the 21st century" where roles are highly desirable and competitive (Davenport and Patil 2012). The authors expect the future to be even more of a continuation of the challenges listed today, including more new data types, increasing innovative and complex trial designs and continued global resourcing. Technical programming and analytical skills will still be required, but having the ability to communicate with other colleagues, influence stakeholders, project manage and feel comfortable making decisions will remain critical. Machine learning, natural language processing and artificial intelligence will need statistical programmers to implement successfully using clinical trial data.

Figure 5 illustrates the past, present and future of drug development and personalized health care



CONCLUSION

The statistical programmer is a key member of any pharmaceutical organization that recognizes the importance of data. A combination of their expertise in statistical programming, data analysis, or exploration are valuable assets they possess to help their team members generate insights from the clinical trial data. They are expected to be leaders and proactively seek solutions to summarize complex data. They are required to disseminate tasks to their supporting team members, sometimes on a global team.

REFERENCES

Australian National Data Service. "The FAIR data principles". ANDS. 2018. Available at https://www.ands.org.au/working-with-data/fairdata

Bennett, Christina. "A Powerful Force: Novel Endpoints Speed Up Drug Development." Oncology Business Review. 2018. Available at https://obroncology.com/article/a-powerful-force-novel-endpoints-speed-up-drug-development/

Cancer Research UK. "How long a new drug takes to go through clinical trials." 2015. Available at http://www.cancerresearchuk.org/about-cancer/find-a-clinical-trial/how-clinical-trials-are-planned-and-organised/howlong-it-takes-for-a-new-drug-to-go-through-clinical-trials

Davenport, Thomas H. and Patil, D.J. "Data Scientist: The Sexiest Job of the 21st Century." Harvard Business Review. October 2012. Available at https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Jansen, Yvette and Thornton, Grant. "Wearables & Big Data In Clinical Trials — Where Do We Stand?" Clinical Leader. February 25, 2020. Available at: https://www.clinicalleader.com/doc/wearables-big-data-in-clinical-trials-where-do-we-stand-0001

Macdonald, Gareth. "Novel endpoints in the digital age." PMLiVe. 16th December 2019. Available at http://www.pmlive.com/pharma_intelligence/Novel_endpoints_in_the_digital_age_1318755

Maxwell, Dr. Michelle. "Personalized Medicine: From 'one size fits all' to bespoke treatment." thepharmaletter. 09-06-2016. Available at https://www.thepharmaletter.com/article/special-report-personalized-medicine

Mercieca-Bebber, Rebecca. "The importance of patient-reported outcomes in clinical trials and strategies for future optimization." Patient Related Outcome Measures. 2018 November 1. Available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6219423/

Precision for Medicine. "Biomarker Data Management in Clinical Trials: Addressing the Challenges of the New Regulatory Landscape." precisionformedicine.com. 2018. Available at https://www.precisionmedicinegrp.com/pfm/wp-content/uploads/sites/3/2018/04/TIB-Biomarker-Data-Management-White-Paper.pdf

Schmidt, Charles. "The benefits of immunotherapy combinations." Nature. December 2017. Available at https://www.nature.com/articles/d41586-017-08702-7

U.S. Department of Health and Human Services Food and Drug Administration. "Providing Regulatory Submissions in Electronic Format — Submissions Under Section 745A(a) of the Federal Food, Drug, and Cosmetic Act". December 2014. Available at https://www.fda.gov/media/88120/download

U.S. Food and Drug Administration (a). "Real-World Evidence". 05/09/2019. Available at https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence

U.S. Food and Drug Administration (b). "Complex Innovative Trial Designs". 07/25/2019. Available at https://www.fda.gov/media/129256/download

Wu M, Sirota M, Butte AJ, Chen B. "Characteristics of drug combination therapy in oncology by analyzing clinical trial data on ClinicalTrials.gov." Pacific Symposium on Biocomputing. (2014) 2015:68–79.

CONTACT INFORMATION

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

Harper Forbes
Hoffman-La Roche Limited Canada
harper.forbes@roche.com
https://www.linkedin.com/in/harper-forbes-b8717252/