VERISTA

Assay Data Management and Why It Matters

WHITE PAPER

INTRODUCTION

Data is a drug company's most valuable resource. A colleague of mine once said that we work in a data company that happens to make drugs. In this article, I discuss the data we generate during the discovery phase in the early stages of the drug development process. I then discuss the important role of assay data management and best practices in this area to ensure discovery data is most useful across a research organization.

First let's look at the drug development process. It's often shown as a pipeline:

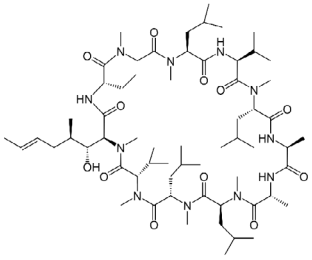
During the Discovery phase, research activities often involve experimenting on cell-based models or other simplified versions of biological processes. These experiments initially focus on basic biological research to understand how a disease progresses and may be influenced. Once disease researchers identify the mechanism that they want to target, the next step is to try to find a potential new drug that modulates that mechanism. More experiments are then performed to characterize these potential new drugs. All these results must be managed and organized to facilitate decision making in the research organization. A quality assay data management system can improve the efficiency of drug discovery in a number of different ways and potentially save hundreds of millions of dollars in wasted R&D costs.

DRUG DEFINITION

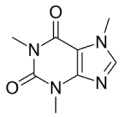
First, what is a drug? Drugs are what every pharmaceutical company is trying to make and sell. A drug is "any substance that causes a change in an organism's physiology or psychology when consumed".[1] Note "consumed" covers all routes of administration, including injection and topical application, as well as ingestion. For a pharmaceutical company, we generally expect drugs to have a primary beneficial effect, improving or curing the progression or impact of a disease or condition.

I will focus the discussion here on small molecule drug discovery. Discovery of large molecule therapeutics such as antibodies has many analogous steps and features, as well as some unique challenges.

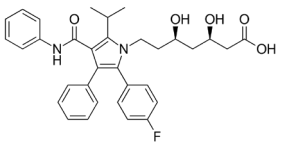
What is a small molecule? "A small molecule is a low molecular weight (< 900 daltons) organic compound that may regulate a biological process, with a size on the order of 1 nm. Many drugs are small molecules."[2] In short, small molecules are compounds that aspire to be drugs one day. Here are some examples:



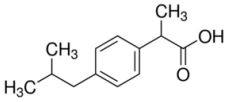
Cyclosporine



Caffeine



Atorvastatin (Lipitor)



Ibuprofen

1. Wikipedia. https://en.wikipedia.org/wiki/Drug

2. Wikipedia. https://en.wikipedia.org/wiki/Small_molecule

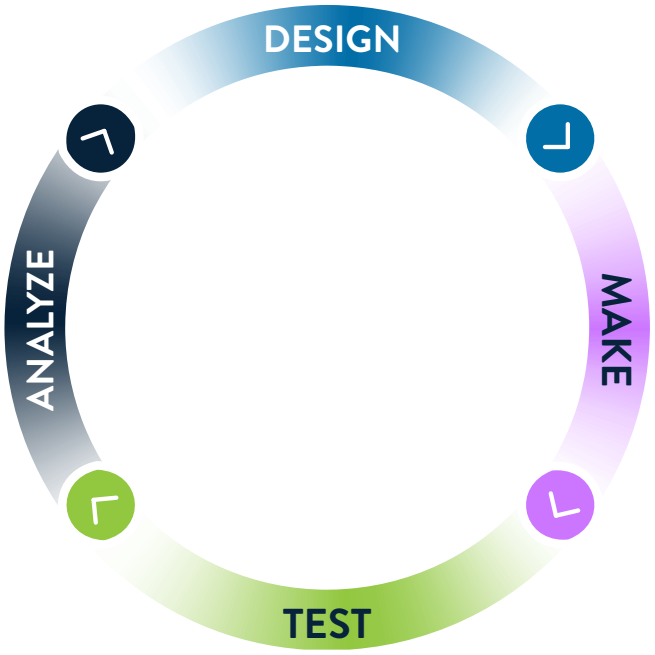
DRUG DISCOVERY

Once a disease mechanism of interest is identified, the search for a new drug that modulates that mechanism in the desired way begins. In addition to merely being effective in that mechanism, the new drug also needs to be specific to that mechanism to avoid unexpected adverse effects, be able to reach the part of the body where it is needed, and remain intact long enough to be effective. A variety of experiments are performed to assess all these properties, and new variants on the molecule are made to optimize its properties. A compound with the right combination of properties can then be advanced to preclinical animal testing, and then to clinical testing. Drug development activities cost much more than drug discovery, and it is highly desirable to only advance drugs to development that are thoroughly characterized and have the highest probability of success.

The process in which scientists repeatedly test compounds for activity, analyze the results, design new variants of the compound of interest, synthesize (make) the new compound and begin the testing process once more is frequently referred to as the DMTA cycle. This is a commonly used representation of the hit discovery and lead optimization portion of drug discovery.[3]

ASSAYS

Assays are experiments used to measure how drugs affect a certain biological system. Assay results are one of the most basic types of research data and are used to inform every decision in the drug discovery process. Assays are used to elucidate disease biology, to identify and characterize targets, to find the first hits in a program, and to optimize those hits until a lead candidate emerges. Scientists constantly invent new assays in response to project needs, technology advances, or simple curiosity.



Assay readouts take many forms such as cell imaging, colorimetric or fluorescent readouts, gene or protein expression measurements, or even small animal testing. The raw readouts are then normalized and converted to results that are useful for making decisions and comparing with other assays. Result types are very diverse and can vary widely in complexity from a single numerical result to many results, pictures, or qualitative observations. Accommodating this diversity is a major challenge for assay data management systems.

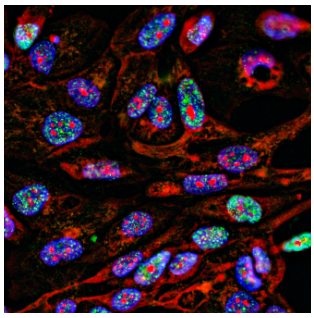
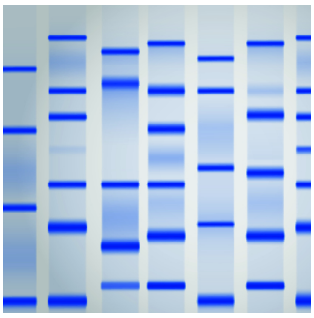
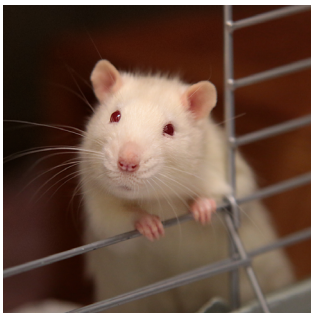
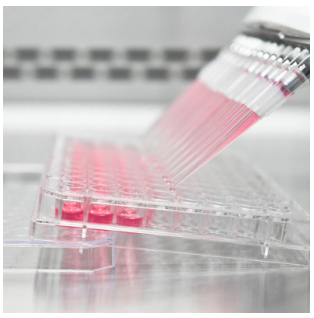
Assays used in drug discovery can be particular to a disease or drug target (e.g., enzymatic assays, cell models of disease), or generic to assess characteristics that we need to know about any new drug (e.g., chemical properties, metabolism, absorption, toxicity, etc.). Many assay results combine to profile a potential drug's ability to have the desired therapeutic effect without undesired effects.

Having a good assay data management system helps scientists gather many different assay results quickly to make decisions to advance or reject compounds. Having well-organized assay data can also mean not having to repeat assays and re-invent assays that were already performed.

3. Wiley Online Library. https://onlinelibrary.wiley.com/doi/10.1002/9783527677047.ch17

HIGH-THROUGHPUT SCREENING (HTS)

High-throughput screening is often the starting point to find compounds in a drug discovery program. In HTS, an assay is developed and miniaturized, then performed thousands or even millions of times. This is achieved by performing the assays in small wells containing tiny volumes. Wells are contained in plates of 96, 384, 1536, or 3456 wells. HTS can only be performed with the help of laboratory automation, and robotic systems are often used to move the plates.

Compounds that are active in the HTS are called “hits”. Hits are then tested again at multiple concentrations to confirm activity, and one or more other assays are usually performed at the same time to assess the hits for specificity. After the HTS is finished, other assays are used to assess the hits further. Compounds that are progressed from HTS and chosen for further optimization are called “leads”.

There is nothing scientifically different between an HTS assay and any other assay. The only difference is the capability to be performed at large scale in plates using automation. An HTS assay can be a variant of the same assay that was previously used to characterize the target or will later be used to assess lead molecules in the drug discovery program.

HTS generates large volumes of data that must be captured, processed, and analyzed to determine which compounds show activity of interest. Assay data management systems have traditionally been developed initially to support HTS, then later expanded to other parts of the research organization.

COMPUTATIONAL BIOLOGY AND CHEMISTRY

Testing compounds and generating assay results provides a benefit beyond just that one drug discovery program. If all the data is added to a well-organized assay data management system, computational biologists and chemists can then mine the data for new insights that add value to the entire research organization.

Aside from profiling specific drug candidates, identifying patterns of drug activity in different assays can help:

- √ Understand pathways
- √ Discover new targets
- √ Identify frequent hitters and recurring assay artifacts so scientists avoid wasting time following up on these compounds
- √ Create predictive models for activity to inform the creation of new molecules

Mining existing data is a major part of modern drug discovery, and the quality and completeness of the data accessible to computational biologists and chemists is a great determiner of how well they can mine insights from it. Hence, the data needs to be comprehensive, organized, and accessible.

FAIR PRINCIPLES

FAIR stands for Findable, Accessible, Interoperable, and Reusable.[4] FAIR principles were published in 2016 by a group of stakeholders from academia, industry, funding agencies, and scholarly publishers to ensure the reusability of data holdings.[5] Assay data system design should follow these principles.

4. GO FAIR. https://www.go-fair.org/fair-principles/

5. Nature.com. https://www.nature.com/articles/sdata201618

ONTOLOGY

All assay data needs to be annotated with controlled vocabulary to describe the assay, so that assay protocols and results can be accurately compared. While it is certainly helpful to have controlled vocabulary to help scientists compare assays themselves, the real value of controlled vocabulary lies in facilitating the use of algorithms for automated comparisons.

The formal representation of such controlled vocabulary that can be used in assay data management systems is called an ontology. The BioAssay Ontology (BAO) is one example of an ontology system used to describe assays that was developed by a consortium of academic HTS labs.[6] There are many others used in pharmaceutical research such as the Experimental Factor Ontology (EFO) or the NCIT for diseases.

In addition to describing and organizing internal assays, using industry-standard ontologies also allows for comparison of internal assay results with external data. A paper from Astra Zeneca[7] describes how the scientists were able to use assay annotations from the BioAssay Ontology to compare internal assay results with those published externally in PubChem to enrich their internal data.

ASSAY DATA MANAGEMENT

Putting it all together, an assay data management system needs to have the following characteristics to accommodate different needs throughout the research organization:

- √ **Flexible** enough to accommodate any assay that scientists can come up with, regardless of its level of complexity.
- √ **Scalable** enough to accommodate volumes of data generated by HTS.
- √ **Easy enough to use** so scientists can define their own assay protocols and result types and upload results without the involvement of the IT department.
- √ **User-friendly** enough to make scientists want to upload all their data consistently.
- √ **Compatible** with reporting tools so scientists can easily compare results from different assays to decide which compounds to progress down the pipeline.
- √ **Well-structured** so the data can be readily mined by computational biologists and chemists.
- √ **Consistent** with FAIR principles.
- √ **Uses ontologies** to ensure controlled vocabulary to compare assays both internally and externally.

6. BioAssay Ontology. http://bioassayontology.org/

7. Astra Zeneca. https://journals.sagepub.com/doi/10.1177/1087057114563493

ABOUT VERISTA

Verista is a leading business and technology consultancy firm that provides systems, compliance, validation and quality solutions to life science companies enabling them to improve health and improve lives. We help clients solve their most critical and complex challenges across the GxP lifecycle, from preclinical and clinical to commercialization, manufacturing and distribution – bringing together decades of knowledge, the most advanced engagement platforms and transformative technologies. This allows clients to benefit from the ease, efficiency and trust that results from working with one partner who excels across specialties. Verista's clients trust the company's 500+ experts to deliver consistent, safe, and high-quality results across the product development lifecycle.

Visit www.verista.com