Data Science In Biomarker Discovery

Andrea Grioni – PyBiella 2023

Agenda:

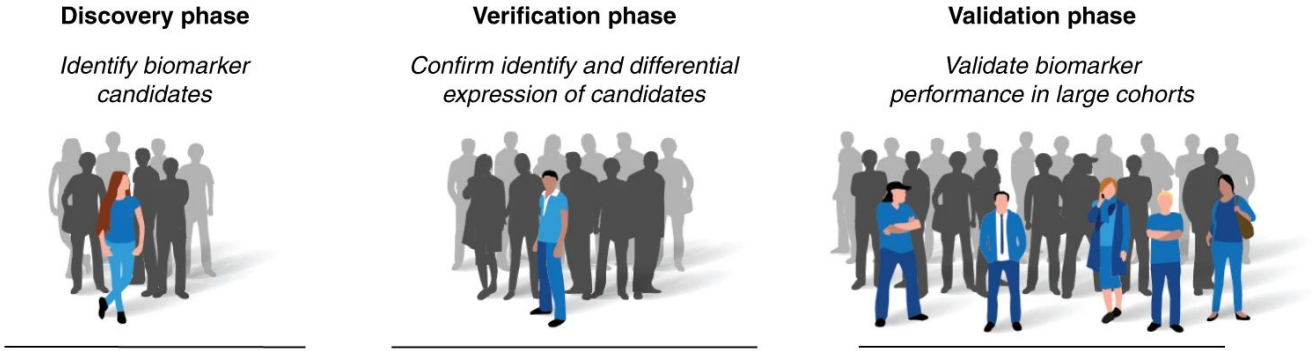
- Intro to Clinical Trial phases
- Intro to Biomarker discovery
- User Case in Python and R

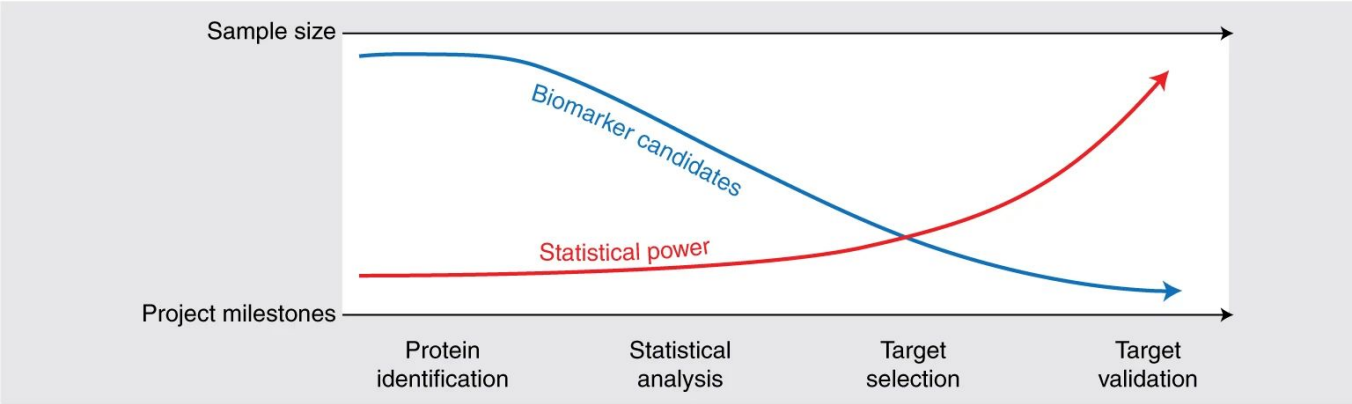
Source: https://www.mdanderson.org/patients-family/diagnosis-treatment/clinical-trials/phases-of-clinical-trials.html

Biomarker Discovery

A biomarker is a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention. In other words, a biomarker is a measurable substance or characteristic that indicates a specific biological condition, such as a disease, or a response to a treatment. Biomarkers can be found in different forms, such as molecules, genes, imaging patterns, or physiological changes, and are used in a variety of fields, including medicine, pharmacology, and environmental science. The discovery and validation of biomarkers is important because they can help in early disease detection, disease progression monitoring, and treatment response evaluation.

Biomarker Discovery





How Does Data Science Support Biomarker Discovery?

Data Science supports biomarker discovery in several ways:

1. Data Collection: Data Science techniques help collect large amounts of data, such as genomics data, medical records, and imaging data, which are used to identify potential biomarkers.
2. Data Pre-processing: Data Science helps clean and pre-process data, so that it can be analyzed effectively and accurately.
3. Data Analysis: Data Science techniques such as machine learning, statistical analysis, and dimensionality reduction are used to analyze data and identify correlations that may lead to the discovery of biomarkers.
4. Validation: Data Science helps validate the results by applying statistical tests, validating the findings through further experimentation and analysis, and creating predictive models to determine the reliability of biomarkers.
5. Visualization: Data Science provides visual representations of data, which can help researchers quickly identify patterns and correlations that may lead to the discovery of biomarkers.

User Case 1 - in Python and R

The idea is to take an open-source dataset and perform a data science analysis oriented into statistics and bioinformatics to identify candidate biomarkers predictive of survival in breast cancer.

A cleaned version of the dataset is available on kaggle

The scientific publication is available online

Environment Setup:

1. Install:
    a. Mamba here
    b. VS Code here
2. Open VS Code - Extension:
    a. Install Copilot [optional]
    b. Install R support
    c. Install Python Support
3. From VS Code:
    a. Clone pybiella_2023 repo here
4. From terminal
    a. Create conda environment from config file:
    `conda env create -f environment.yml`