**Project 1 - Align attribute names for stability testing**

**Problem statement**

Bristol Myers Squibb (BMS) has over 100 years of history and is one of the earliest adopters of electronic systems to record our data and go paperless. Our historical data and prior knowledge are of tremendous value, especially when it comes to leveraging the massive amount of existing data to feed into AI/ML models to gain more insights, enable predictive power, and eventually empower reliable data-driven decisions. However, over the decades, industrial standards/systems and regulatory requirements on electronic data have been evolving significantly along with the rapid development of technologies and digitalization. Although we have implemented standards to collect data in the electronic systems for many years, we observed various data issues (e.g., inconsistency, typos, etc.) across products, studies, etc. The historical data need to be cleaned and aligned to enable reliable and meaningful data-driven decisions by leveraging AI/ML models.

**Task**

1. Sign the data confidentiality consent form and return to BMS representative to gain dataset access.
2. Read regulator’s guideline (ICH Q6A and Q6B) on stability testing to have a brief understanding of the attributes.
3. Write Python script to efficiently clean and align the attribute names. (A list of standardized attribute names will be provided). Ensure **all** the attribute names are correctly mapped and there are no left out unless they are not in the standardized attributes list.
4. (Bonus) Leverage natural language processing (NLP) or large language models (LLMs) to clean the data more efficiently.