**Background**

Synthetic data is artificially developed using a variety of methods. Comprehensive synthetic data would mimic real-world data as the development process would use qualitative and quantitative features of a disease. Synthetic data is different to augmented or randomised data.

Data augmentation uses a process where the data is modified based on the same image. For example, an image of a flower can be used to augment a particular colour or shade of colour. Data randomizers on the other hand shift features from the data pool generating a new image or dataset. For example, the characteristics of the flower may shift between the images. If the synthetic data was applied, a new flower would be developed using the characteristics of the original flower, but this would look different to the original flower. Computerised methods are not the only method that can be applied to generate synthetic data. Mathematical and/or statistical approaches could also be used to generate synthetic data albeit this is a time consuming method.

There are two main type of synthetic data; full and partial. Partial synthetic data includes a combination of the real and synthetic data observations or measurements. The full type of synthetic data refers to simulated data. The decision to have either a partial or full dataset depends on the purpose as well as scope of the intended use. For example, models trained on synthetic data for the purpose of healthcare should have a verifiable and transparent process with a documented *know-how* to ensure reproducibility with other datasets.

**Use of synthetic data**

There are several key reasons for using synthetic data in healthcare settings as shown below;

1. Synthetic data could assist with addressing the missing data issues.
   1. The rate of missing data is problematic when using epidemiology data, especially from longitudinal studies. Engaging and gathering prospective data collection is cumbersome and expensive. This is often inconvenient to patients as well.
   2. The rate of missingness in clinical trials are a barrier to gathering long-term effectiveness data which is a barrier to decide suitability, acceptability and tolerability for patients.
   3. The rate of missingness in electronic healthcare systems is problematic, especially during transition phases. Most healthcare systems now do not use paper based medical records. This has meant older data may not always be available or there could be transcription errors during the transfer process. Equally, evolving healthcare practice has meant the data collected has somewhat changed.
2. Assisting to model and understand rare diseases using synthetic data.
3. Using synthetic data to help understand chronic conditions can be a powerful tool to develop patient centred and clinically optimal interventions.
4. Differences in governance, data sharing and privacy laws globally makes it challenging for researchers and innovators to learn from clinical data to develop better interventions. Synthetic data can be a viable alternative.