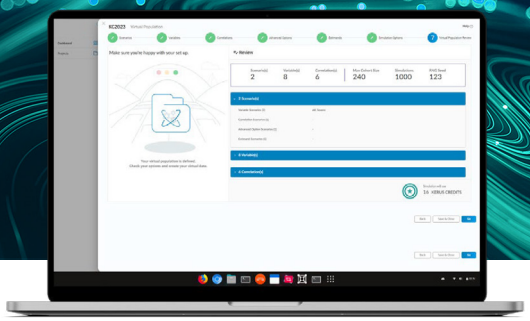
KerusCloud

exploristics.com/synthetic-dataset-generation

Synthetic Data

Generating realistic synthetic datasets for widespread applications.



Accessing subject-level data would make a significant contribution to driving innovation across many industries. The use of subject-level data has many applications such as:

- conducting research
- testing and validating products
- building and testing algorithms
- running simulations
- business intelligence

However, due to data privacy laws, reluctance and sensitivity around data sharing, it is often difficult to access data and many organizations have not been able to harness the potential power of this data. An elegant solution to data access issues is to develop synthetic datasets that are derivatives of the subject level data but do not contain any protected sensitive information. These datasets can be shared freely among investigators or those in industry, without raising privacy concerns. The availability of synthetic datasets would fill the void and allow organisations to accelerate the development of new innovative products, services and intellectual property.

Handling complex data scenarios

At the core of our approach is our unique synthetic data generation and simulation platform, KerusCloud. KerusCloud has broad utility for the generation and exploitation of complex datasets. Leveraging the combined power of analytics and cloud computing, it can handle diverse and complex data collected from a wide variety of data sources and use them to generate realistic virtual data. KerusCloud goes beyond the use of population level statistics as a basis for simulation, as it can model the intercorrelation between subject level data such as subgroups and strata, risk factors/covariates and multiple outcomes and data types.

With KerusCloud, special features can be added to the synthetic datasets such as missing data, truncation and censoring. This allows it to generate the most realistic synthetic version of the original data with broad utility for many users.

Creating the Synthetic Datasets

KerusCloud can be licensed for direct use in synthetic data generation. However, we also provide this as a technologyenabled service. Exploristics' **Data Science team** have key expertise in evidence synthesis and proprietary in-house tools that enable them to collate data from varied sources, process it and generate synthetic data sets from it. There are multiple ways to generate synthetic data: use a real, patient level dataset to derive the key characteristics for the synthetic data; use the KerusCloud interface to manually input the summary statistics that define the characteristics of the synthetic data; create an augmented dataset using both patient-level data and summary statistics.

The creation of synthetic data from patient-level data involves a five-stage process (Figure 1):

1. Identify relevant datasets, create a common data structure and prepare them for cleaning and standardisation.

2. Clean and standardise the data. This stage involves removing unnecessary fields, identifying and dealing with discordant values and implementing a missing data strategy. Once cleaned these data are ready to convert into representative summary statistics with this information captured within a data model.

Exploristics Ltd.
24 Linenhall Street Belfast,
Northern Ireland BT2 8BG

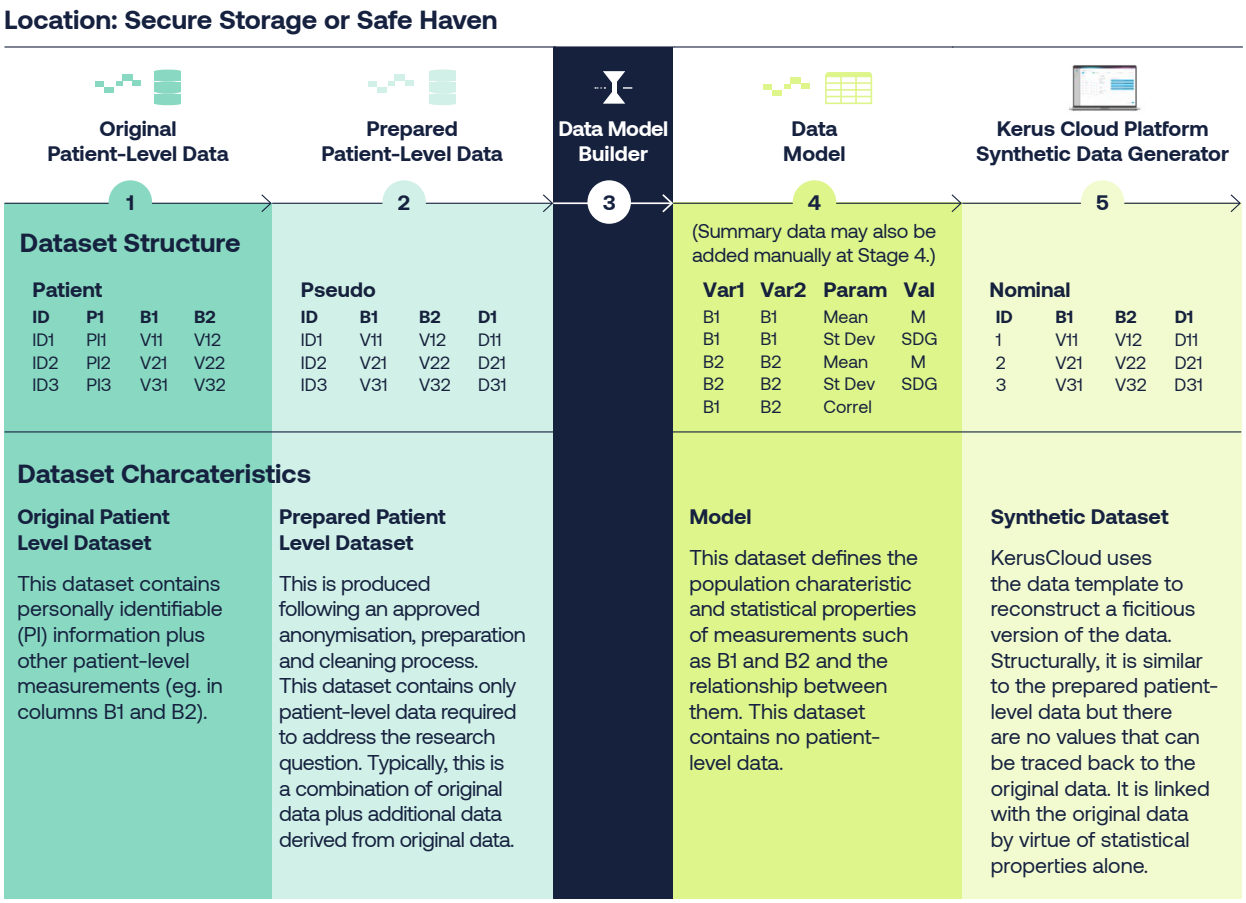
Find out more
T +44 (0) 28 9600 1996
info@exploristics.com

3. Create a data model using Exploristics' proprietary Data Model Builder (DMB). The DMB sits within a private, secure environment and imports the original patient level data; it extracts the key characteristics of the dataset in the form of summary statistics. The data model is structured so it is readable by KerusCloud and to facilitate automatic loading into KerusCloud.

4. Save the data models created from the Data Model Builder. An alternative approach at this stage is to manually input summary statistics obtained from other sources into the data models to augment the information contained in the data model or to create a new data model.

5. Import the data model into KerusCloud. The data model can be used to create realistic synthetic data using the software's synthetic data generation engine. These synthetic data accurately describe the structure and characteristics of the real source data providing a realistic alternative while not infringing privacy.

Figure 1: The five stages of the process to generate synthetic data



Once the original subject/patient level data has been identified (Stage 1) and prepared (Stage 2), they are translated via the Data Model Builder into data models (Stage 3) which are population level summary statistics that describe the features of the data. Note, the models do not contain any subject-level data. The data model is saved to a standard structure (Stage 4). At this stage further information can be manually added to the data model or a new data model can be created entirely from summary statistics. These models are automatically uploaded into KerusCloud (Stage 5) where they are used to randomly generate a synthetic version of the data.

Download the Synthetic Data

As KerusCloud uses the statistical properties of the original data, it can generate multiple unique variants of the source data. This provides the user with the opportunity to either select the variant that best matches the original data source or to use all the variants. All synthetic data generated for a project can be easily exported from KerusCloud by downloading it into a csv file. This ensures complete flexibility in its use going forward.

So, don't let data access issues stall your research. Harness the game-changing benefits of realistic synthetic data with Exploristics Data Science team and KerusCloud.

Exploristics Ltd.
24 Linenhall Street Belfast,
Northern Ireland BT2 8BG

Find out more
T +44 (0) 28 9600 1996
info@exploristics.com